IMPERIAL COLLEGE LONDON
Department of Brain Sciences

PH.D. THESIS

IMPROVING MODELLING IN COGNITIVE SCIENCE

The application of state-of-the-art nested sampling methods of Bayesian inference to advance the theory of simple response time

Darije Čustović

supervised by

Professors Adam Hampshire and Claudia Clopath

Submitted in part fulfilment of the requirements
for the degree of Doctor of Philosophy at
Imperial College London,
22 February 2022

I hereby declare that this thesis was composed by me, and me alone; that all the work reported herein originated from and was undertaken by me, and me alone; and that information derived from the works of others has been acknowledged and referenced in the body of this work.

– Darije Čustović (22 February 2022)

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes, and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

To all future students of the topics herein:

I hope you find something useful.

All things I thought I knew; but now confess

The more I know, I know, I know the less.

— John Owens, Works, Book VI

General abstract

This thesis seeks to advance modelling in cognitive science in three ways. The first is by introducing the field to MultiNest, an exemplar of the overlooked *nested sampling* approach to performing Bayesian inference borrowed from advanced physics; it will be demonstrated that, in applications involving cognitive models and data, MultiNest offers a number of advantages over the prominently proposed thermodynamic integration and steppingstone methods when calculating the crucial Bayes factor.

Second, the thesis puts forward new and more expansive models of simple reaction time task (SRT) performance. Despite the importance of empirically well-established foreperiod effects in simple responding, extant mathematical models of SRT performance confine themselves to describing the marginal response time distribution only; the new models presented here, developed by extending evidence accumulator models of decision making to possess foreperiod-modulated time-dependent decision thresholds, simultaneously account for both foreperiod effects and response time distribution.

Finally, it is argued that modelling work in cognitive science, *pace* current tendencies towards simplifying assumptions, should focus on providing full accounts of complex cognitive phenomena. Bayesian model comparison – undertaken with MultiNest and SRT data from numerous subjects – shows that the new SRT models, despite their increased complexity, receive far greater support from the data across subjects relative to their predecessors. This result is interpreted to mean that the additional complexity in the new models echoes complex patterns inherent in the data that must be preserved in any satisfactory theory explaining the data: the data, it will be said, possess *irreducible complexity*. If this is true when considering a task as simple as the SRT, it is likely true across a host of other cognitive domains; thus, it is concluded, cognitive modelling would do well to embrace the goal of developing models which give full accounts of the irreducible complexity of cognitive data.

Contents

General abstract **vii**

1 General introduction **1**

1.1 Context, scope, and aims of this work 1

1.2 The major arguments and through-lines of this work 2

1.2:1 Nested sampling for Bayesian inference in cognitive science 2

1.2:2 Improved modelling of simple reaction time task performance 4

1.2:3 Attitudes towards the invention and interpretation of cognitive models . 5

1.3 Plan of work 6

PART I. FOUNDATIONS

2 Prolegomenon to MultiNest: Induction, statistics, and Bayesian inference **15**

2.1 Induction and the confirmation of hypotheses 16

2.1:1 The nature and function of inductive inferences 16

2.1:2 Evaluating inductive arguments 18

2.2 The Bayesian approach to induction 19

2.2:1 Confidence, plausibility, and Bayesian probability 19

2.2:2 Bayes' theorem in Bayesian probability theory 22

2.2:3 The elements and mechanics of Bayes' theorem for induction 25

2.3 Statistical inference 28

2.3:1 The scope of statistical inference 29

2.3:2 Mathematical formalism 30

2.3:3 The parameter space of a model 32

2.3:4 The problems of statistical inference 33

2.4 The general theory of Bayesian inference 34

2.4:1 Parameter estimation and model selection in Bayesian terms 35

2.4:2 The interpretation of Bayes' theorem for statistical inference 37

3 Nested sampling and MultiNest **43**

3.1 The practical challenges of Bayesian model analysis 43

3.2 Nested sampling . . . 45
3.2:1 Transforming the evidence integral: Λ . . . 45
3.2:2 Interpreting Λ in parameter space . . . 49
3.2:3 The idea of nested sampling: Direct integration over the prior mass . . . 50
3.2:4 The algorithm for nested sampling . . . 52
3.2:5 Final evidence estimate, errors, and posterior inferences . . . 53
3.3 The MultiNest algorithm . . . 54
3.3:1 The limitations of vanilla nested sampling . . . 54
3.3:2 Intermediate developments: Ellipsoidal and recursive clustered nested sampling . . . 55
3.3:3 MultiNest's refinements to nested sampling . . . 57
3.3:4 Evidence of MultiNest's efficacy in application . . . 58
3.3:5 Other nested sampling techniques and the choice to use MultiNest . . . 59

4 The simple reaction time task 61
4.1 Why study simple reaction time? . . . 61
4.2 Description of the task . . . 65
4.2:1 Experimental design . . . 65
4.2:2 Task measures . . . 67
4.3 The distribution of response times . . . 68
4.4 Mathematical models of response time and the SRT . . . 70
4.4:1 Shifted Wald model . . . 70
4.4:2 Ex-Gaussian model . . . 76
4.4:3 Ex-Wald model . . . 77
4.4:4 Ratcliff one-choice drift-diffusion model . . . 78
4.5 Foreperiod effects . . . 78
4.5:1 Overview . . . 78
4.5:2 Empirical findings . . . 79
4.5:3 Theoretical discussion . . . 81
4.6 Two problems of synthesis in the theory of simple responding . . . 83

PART II. INVESTIGATIONS

5 Demonstrating the basic viability of MultiNest 89
5.1 Introduction · Efficacy and competitiveness . . . 90
5.2 First test · A simple Gaussian likelihood . . . 93
5.2:1 The justification for and aims of the test . . . 93
5.2:2 General methods . . . 95

5.2:3 Assessment A: Accuracy and precision of Bayesian evidence estimate . . 99
5.2:4 Assessment B: Accuracy of Bayesian evidence uncertainty estimate . . . 110
5.2:5 Assessment C: Accuracy of posterior estimate 113
5.2:6 General discussion of the first test 122
5.3 Second test · The linear ballistic accumulator 125
5.3:1 The justification for and aims of the test 125
5.3:2 Methods 126
5.3:3 Results and discussion 129
5.4 Discussion 134

6 MultiNest and models of response time: A simulation study 139
6.1 Introduction 139
6.2 General theory and methods 141
6.2:1 Choice of models 141
6.2:2 Data generation procedure 142
6.2:3 Bayesian model selection and parameter estimation; MultiNest 146
6.2:4 Prior specifications 148
6.2:5 Log-likelihood calculations 150
6.2:6 Overview of subsequent analyses 152
6.3 Intra-subject model selection capabilities 152
6.3:1 Aims and methods 152
6.3:2 Results 155
6.3:3 Interim discussion 163
6.4 Trans-subject model selection capabilities 164
6.4:1 Aims and methods 164
6.4:2 Results and discussion 164
6.5 Comparison to information criteria model selection 165
6.5:1 Aims and methods 165
6.5:2 Results 166

7 An empirical study of the simple reaction time task 171
7.1 Introduction · A verification study 171
7.2 General methods 172
7.2:1 Data collection 172
7.2:2 Data pre-processing: Classifying responses and extracting response times 174
7.3 Response time distributions and hazard functions 177
7.3:1 Aims and methods 177
7.3:2 Results and discussion 178

7.4 Foreperiod distribution effects . . . 183
7.4:1 Aims and methods . . . 183
7.4:2 Results and discussion . . . 185
7.5 Asymmetric sequential foreperiod effects . . . 187
7.5:1 Aims and methods . . . 187
7.5:2 Results and discussion . . . 192
7.6 Discussion . . . 192

8 New models of simple responding 195
8.1 Shortcomings in the theory of simple responding . . . 195
8.1:1 On the theory and significance of one-choice reaction tasks . . . 195
8.1:2 Empirical phenomena in simple responding . . . 197
8.1:3 Models of response time and their shortcomings . . . 198
8.1:4 The specific aims of this chapter . . . 199
8.2 General approach and preliminaries . . . 200
8.3 Foreperiod parameter-dependence models of simple responding . . . 200
8.4 Collapsing threshold models of simple responding . . . 203
8.4:1 The basic collapsing threshold model . . . 204
8.4:2 The collapsing threshold model with sequentially varying span and decay rate . . . 204

9 A Bayesian assessment of models of simple responding 207
9.1 Introduction . . . 208
9.1:1 General background . . . 208
9.1:2 Model selection in cognitive science . . . 208
9.1:3 Simple responding, foreperiod effects, and models of response time distribution . . . 210
9.1:4 Aims and plan of this chapter . . . 212
9.2 Bayesian analysis of models . . . 213
9.2:1 Fit data and sampling parameters . . . 213
9.2:2 Models, likelihood functions, and priors . . . 213
9.2:3 Overview of analysis . . . 216
9.2:4 General notes . . . 216
9.3 Results . . . 217
9.3:1 Model comparison at ensemble or group level . . . 217
9.3:2 Model comparison at subject level . . . 217
9.3:3 Posterior estimates . . . 219
9.3:4 Comparison of model postdictions to empirical data . . . 219

9.4 Discussion . . . 225
9.4:1 Summary of work . . . 225
9.4:2 On the use of collapsing threshold models of simple responding . . . 229
9.4:3 On the IID assumption in RT modelling . . . 230
9.4:4 On the use of MULTINEST in cognitive science . . . 232
9.4:5 Epilogue . . . 233

PART III. CONCLUSIONS

10 General discussion **239**
10.1 Summary of work . . . 239
10.2 On the irreducible complexity of cognitive data . . . 241
10.3 Attitudes and approaches towards modelling cognition . . . 243
10.4 Future directions . . . 244
10.4:1 Modelling and model comparison in other cognitive tasks . . . 244
10.4:2 Simultaneous modelling of and inference from multiple tasks . . . 244
10.4:3 Modelling in cognitive neuroscience with multimodal data . . . 245
10.4:4 Relating novel behavioural models of the SRT to brain measures . . . 245
10.4:5 Improving the implementation of current models . . . 245
10.4:6 The utilisation of other advanced nested sampling algorithms . . . 246
10.4:7 Nested sampling for other purposes . . . 246
10.4:8 Translation to clinical application . . . 246
10.5 Closing statement . . . 247

Acknowledgements **249**

Appendices **251**
A The origins and obstacles of modelling . . . 251
B A visual proof of Bayes' theorem . . . 255
C Inference from multiple datasets . . . 256
D Prior sensitivity and absolute superiority . . . 259
E Markov Chain Monte Carlo . . . 260
F The linear ballistic accumulator model of choice response . . . 261
G Kernel density estimation . . . 263
H Permissions for reproduced works . . . 264

Bibliography **267**

Improving modelling in cognitive science

CHAPTER 1

General introduction

1. CONTEXT, SCOPE, AND AIMS OF THIS WORK

The purpose of cognitive psychology is to understand the nature of cognition: the components, structure, and operation of mental processes, broadly construed. Put differently, the discipline aims to understand how the mind acquires, generates, processes, transforms, evaluates, and acts upon mental content – thoughts, sensations, goals, memories, knowledge, and so on. To pursue the relevant questions empirically, cognitive scientists create tasks which aim to place a few simple, tightly controlled demands on the processes of cognition; administer the tasks to willing subjects; and observe the resultant behaviour (since cognitive processes are observable only in their effects). In this way, rich and varied data have been collected on many aspects of behaviour under a huge variety of cognitive demands.

Of course, collecting data is not, by itself, enough. To understand cognition, the cognitive scientist must attempt to identify patterns inside the behavioural data, and thence to draw conclusions about the mental processes involved. Complementing experimental work, then, is theoretical work, of which an important part takes the form of *cognitive models*. Cognitive models provide formal, explanatory theories of the cognitive processes underlying complex human behaviour (Annis and Palmeri, 2018; Farrell and Lewandowsky, 2018). The first essential feature of cognitive models, their formality, entails that they are expressed mathematically. Usually, this means – given the unpredictable variation inherent to behaviour – that they are stated as statistical models. The second essential feature, their explanatory nature, means that they are intended to do more than merely describe or summarise data; they are meant to offer rational, conceptually clear accounts of *how* the phenomena at hand come about. To this end, the purely mathematical components of the model are bestowed with psychological interpretations; or conversely, theoretical psychological constructs are given mathematical form and precision.

This thesis is grounded in the proposition that efforts to build cognitive models are laudable – and necessary, if we are ever to have a proper scientific understanding of cognition. Modelling work, however, presents a number of distinctive challenges, including how to adjudicate between competing models of the same phenomenon; how to choose which aspects of the data should be modelled and which should be set aside; and how to properly interpret the mathematical

formalisms of the models in psychological terms. (A general discussion on the challenges of modelling can be found in appendix A.) Not one of these challenges has a definitive solution in cognitive modelling; and providing new solutions to them is the central aim of this thesis. Put simply, this thesis asks: Can we improve our models of cognitive processes? And can we improve the way we conduct modelling work in cognitive science?

In light of this, the principal undertakings of this work are to provide a detailed examination of the methods employed, substance of, and attitudes adopted in cognitive modelling; and, on the basis of this examination, to recommend changes and improvements to be made in these various aspects, in order that the science might better achieve its ends. These undertakings are, naturally, a little too great to be definitively accomplished within the scope of a single work. Within this ambit, then, this thesis addresses three themes of current scientific interest: the use of Bayesian inference methods in cognitive science; modelling of simple reaction time task performance (and response time distribution modelling more generally), with a particular focus on the time-course of temporal preparation; and the attitudes taken towards the invention and interpretation of cognitive models.

In particular, this thesis seeks to advance modelling in cognitive science in three ways, which form the through-lines connecting the various parts of this thesis. The first way is to provide better tools, by introducing the field to MultiNest, a means of performing Bayesian inference borrowed from advanced physics. The second is to directly provide better models, by offering new and more expansive models of the simple reaction time task. The final way is to map out a new direction for cognitive modelling work, by arguing that modelling work in cognitive science, *pace* current concerns regarding simplicity in models, should focus on providing full accounts of the inseparable complex phenomena endogenous to cognitive data; and by further arguing that until such models are built, attempts to apply psychological interpretations to the mathematical models should be undertaken with caution. Let a little more be said about the motivating forces behind, arguments for, and relationships between each of these threads.

2. THE MAJOR ARGUMENTS AND THROUGH-LINES OF THIS WORK

2:1 Nested sampling for Bayesian inference in cognitive science

Recent decades have seen a proliferation of mathematical models of cognitive processes, with most mathematical accounts of cognitive data taking the form of statistical models (Evans and Annis, 2019). This proliferation of models, and their burgeoning complexity, has led to an upsurge in the use of Bayesian inference for the analysis of these models (Andrews and Baguley, 2013; Lee and Wagenmakers, 2013).

The mathematical theory of Bayesian inference (BI) provides a principled basis for selecting

that statistical hypothesis from amongst a set of candidates which provides the best explanation of the data – or, equivalently, of finding the relative weight of evidence in favour of each hypothesis as determined by the data (Jaynes, 2003; Jeffreys, 1961). The devices of the *posterior distribution* and the *Bayes factor* (the comparison of individual models' *Bayesian evidence* scores) rigorously quantify the weight of evidence for problems of *parameter estimation*, where the statistical hypotheses being considered take the form of different parameters for a given likelihood function, and *model selection* (or model comparison), where the hypotheses represent discrete theories of the situation at hand (D'Agostini, 2003). However, calculating these quantities to make an inference about the best hypothesis is, in practice, quite difficult; this is especially true of the Bayes factor.

The standard approach for making posterior inferences (i.e., of calculating posterior distributions) is to use so-called *Markov Chain Monte Carlo* (MCMC) techniques. Numerous methods for obtaining Bayes factors have been built on top of these MCMC techniques; some of these have been prominently promoted for use in cognitive science, most notably *thermodynamic integration* (TI) and *steppingstone* (SS) approaches (Annis and Palmeri, 2018). My contention here is that an overlooked approach, *nested sampling*, provides a better solution. In particular, the MultiNest algorithm, an advanced form of nested sampling developed for use in cosmology, astrophysics, and particle physics, will be advanced as an alternative to the above-mentioned approaches (Feroz et al., 2009).

After some introductory chapters which lay out the deep rationale for the adoption of the Bayesian approach (Chapter 2) and which provide an exposition of nested sampling in general and MultiNest in particular (Chapter 3), a series of three studies will attempt to justify this contention. The first (Chapter 5) will establish the inherent efficacy of MultiNest for problems of inference and its competitiveness with the TI and SS approaches mentioned above. The second (Chapter 6) will demonstrate MultiNest's suitability for use with response time models and data. The final study (Chapter 9) will vindicate its use in practice, through application to current issues in the theory of response time modelling and simple responding (more of which below). Further details about the contents of these chapters is given below, in the section outlining the plan of work.

In applying BI techniques to cognitive data, a few small extensions to the theory of BI itself are made, including a note on how Bayes factors can be interpreted given the limits of prior sensitivity (appendix D); and an extension of the notion of the Bayes factors (appendix C) to cover not just a model applied to a dataset, but "ensembles" of models applied to data "corpora" – that is, a set of models matched to independent-but-similar datasets collected from multiple units (as is common in cognitive science with, e.g., behavioural data from a single task and multiple subjects).

2:2 Improved modelling of simple reaction time task performance

The second major theme of this thesis is the modelling of simple reaction time task (SRT) performance. The SRT requires, on each trial, a speeded response to a known stimulus presented at a known location after an unknown time, the *foreperiod* (FP). With regards to the general concerns of this thesis, the choice to study the SRT has four motivations. First, the SRT yields data in the form of response times (RTs), and the study of response time has yielded perhaps the richest and deepest vein of quantitative, theory-based modelling in the whole of cognitive psychology. Moreover, response time is "psychology's ubiquitous dependent variable", and thereby has broad significance to many aspects of the field (Luce, 1986). Second, the SRT is the one of the simplest examples of a reaction time task, and is, in a sense, the kernel of all others, since the SRT demands nothing more than a response to a signal, and all other response time tasks require at least a response to a signal. Given these two considerations, any conclusions drawn about modelling from attempts to model the SRT therefore have some hope to be generalised to cognitive modelling more broadly. Third, the SRT is a basic measure of attention and prediction, and thus of basic interest to cognitive science as means of studying those aspects of mental operation. Finally, there are important lacunae in our theoretical grasp of the processes underlying simple responding.

These lacunae can be approached from two directions. Long-established in empirical SRT research is the presence of a family of so-called foreperiod effects (Los et al., 2014). One branch of this family comprises *foreperiod distribution effects*, whereby mean RT varies with FP according to the task's overall FP distribution; for example, uniformly distributed FPs tend to produce faster responses on those trials with short FPs. The other branch comprises *asymmetric sequential foreperiod effects*, whereby the FP distribution effect is itself modulated by the FP on preceding trials; a long FP on the preceding trial tends to exaggerate the within-trial FP effect, whereas a short one blunts it. These phenomena implicate complex mechanisms of attention, learning, and prediction in the performance of even the simplest of response time tasks. However, theoretical work to understand the mechanisms underlying foreperiod effects have thus far been non-mathematical pursuits, and have concomitantly ignored another important aspect of SRT phenomenology: the random variation of RT and the resultant RT distribution.

By contrast, mathematical modelling work around the SRT has been confined almost exclusively to the modelling of the marginal RT distribution (see, for example, Ratcliff and Van Dongen (2011)). The currently most popular models of response time are *evidence accumulator models*, which take the decision to respond to be the result of the noisy accumulation of so-called "sensory evidence" to a decision threshold (Evans & Wagenmakers, 2020). Importantly, EAM models (and other models of RT) have hitherto usually assumed that the trials of a task which yields a RT distribution can be modelled as independent and identically distributed, and that across-trial variability can be attributed to random noise (Steingroever et al., 2020; Whelan, 2008); foreperiod effects, of course, demonstrate clear violations of these assumptions. (Chapter 4 contains a fuller

review of the current state of both empirical and theoretical work in the SRT, and RT modelling more generally.)

In light of these lacunae, this thesis offers some first attempts at modelling the marginal RT distribution while taking account of the systemic effects of FP on mean RT. This will be done by extending the EAM framework to to include foreperiod-modulated time-dependent decision thresholds (Chapter 8). These novel models are of considerable greater complexity than extant ones (at a basic level, because they require individual RT distributions on each trial), raising the issue of of whether they their additional complexity is warranted by whatever improvement in fit they might achieve – put simply, do they provide better accounts of the data? To answer this question, Bayesian model comparison of both the new models and many extant ones is performed using MULTINEST and SRT data from numerous subjects; this analysis shows that the new models developed here receive far greater support from the data relative to their predecessors (Chapter 9).

2:3 Attitudes towards the invention and interpretation of cognitive models

On a basic level, the result of this Bayesian analysis acts to recommend the adoption of the newly proposed models of the SRT; the culminating thread goes beyond this, scrutinising the implications of the result in light of the fuller understanding of the Bayes factor outlined in the early chapters (Chapter 10). In essence, the result is interpreted to mean that the additional complexities in the new models echo complex patterns inherent in the data that must be preserved in any satisfactory theory explaining the data: the data, it will be said, possess *irreducible complexity*. If this is true when considering a task as simple as the SRT, it is likely true across a host of other cognitive domains; thus, it is concluded, cognitive modelling would do well to embrace the goal of developing models which give full accounts of the irreducible complexity of cognitive data. In practical terms, this suggests that, despite most discussions of model complexity emphasising the overfitting problem, current cognitive models might be underfitted.

Furthermore, insofar as one uses models to provide accurate accounts of the mechanisms of cognition, this understanding of the SRT modelling result cautions against the ready psychological interpretation of mathematical model parameters; this speaks against current ideas of using cognitive models as "measuring tools": fitting cognitive models to data and leveraging the psychological interpretations imputed on their mathematical parameters to infer the values of individual's latent cognitive parameters (Dutilh et al., 2019; Evans, 2019; Evans & Wagenmakers, 2020; Forstmann et al., 2011),

3. PLAN OF WORK

The three threads laid out above are woven together throughout this work, especially in the final few chapters. For clarity and reference, this section lays out the overall plan of work, with allusions to these three through-lines, around a series of chapter summaries; the chapter summaries presented here are adapted and abbreviated versions of the abstracts and synopses of the chapters proper.

This thesis is organised into three main parts, "Foundations", "Investigations", and "Conclusions". Across three chapter, the first part develops the basic ideas of Bayesian inference, so that the full import of later results and arguments might be understood; gives an exposition of nested sampling and MultiNest; and reviews the literature (both experimental and modelling) surrounding the SRT.

In particular, Chapter 2 is as a guide for the Bayesian initiate, explaining how Bayesian statistics repurposes probability theory to provide a completely general and coherent framework for performing inference with statistical data and models. The view presented, leaning heavily on the ideas of Jeffreys (1961) and Jaynes (2003), sees BI as providing rules for performing *inductive inference*, that is, the kind of reasoning that allows us to assess the support lent by data to competing hypotheses – the evidence for or against them. Key to this exposition will be the notion that statements of confidence in a proposition can be expressed as probabilities, which allows the formal tools of probability theory to be brought to bear on induction. Armed with this notion, it will be shown how the Bayesian framework accommodates both of the most common classes of statistical problem, *parameter estimation* and *model selection*. Particular emphasis is placed on the interpretation of Bayes factors.

Chapter 3 moves on to the practical developments of the ideas in Chapter 2, providing a comprehensive exegesis of the MultiNest method from the basic ideas of nested sampling up to the particulars of the algorithm itself. A deep understanding of the basic ideas, as well as theoretical reasons for and practical examples of the tool's effectiveness, are stressed throughout.

Suspending the Bayesian thread, Chapter 4 turns its attention to the SRT. The chapter first outlines the main features of the SRT's RT distribution in both its PDF and hazard function representations, as well as numerous existing mathematical models of RT distribution; this is followed by a review of the literature on foreperiod effects. The chapter closes by considering two theoretical syntheses that are required to advance our knowledge of the processes of simple responding, setting the stage for the issues to be be tackled in later chapters.

The second part, "Investigations", covers the original research reported in this thesis. This will be presented across four chapters.

Chapter 5 aims to establish MultiNest as a contender within the sphere of BI methods used in cognitive science. This will be done by demonstrating that MultiNest satisfies two important and necessary (if not sufficient) criteria for the positive assessment of any practical tool. The first

is *efficacy*: that is, that it is is capable of achieving the desired results, namely, producing accurate and precise estimates of models' Bayesian evidences and posterior parameter distributions. Second is *competitiveness*: that it is efficacious in a way superior to rival methods according to relevant desiderata. The above-mentioned thermodynamic integration and steppingstone methods will be taken as MultiNest's benchmark competitors; and the assessment of competitiveness is done along a number of practical dimensions, such as computational cost, ease-of-use, and accuracy of estimates.

The relevant demonstration are made via two sub-studies. In the first, MultiNest is applied to a simple Gaussian model where the posterior and Bayesian evidence are analytically tractable. This approach sets the technique against the only absolute benchmark available, and therefore strictly evaluates MultiNest's fundamental capacities. It is shown that MultiNest is capable of producing extremely accurate posterior and Bayesian evidence estimates, as well as useful error estimates.

A second test assesses MultiNest's performance in a cognitive modelling context, attacking a previously studied toy problem in model selection involving the linear ballistic accumulator model; here, it is found that MultiNest's estimates are in general agreement with earlier ones. These findings supplement and support the already-convincing literature on MultiNest's efficacy covered in Chapter 3.

Further, in both studies, comparisons to the benchmark methods can be made; these comparisons bear out that MultiNest offers numerous advantages and suffers from no serious drawbacks by comparison. In sum, it is argued that it is an accurate, efficient, general purpose, all-in-one and easy-to-use method for model inference.

Chapter 6 aims to ground MultiNest's reliability for applications involving RT models and data – as will be made in Chapter 9 – by testing it on simulated data for which the underlying generative model and parameters are exactly known. Models are chosen that, in light of current understanding, create plausible response time distributions to simulate those obtained from a simple reaction time task. Each model is used to generate simulated data for numerous virtual subjects, with realistic inter-subject differences in parameters, and the various models repeatedly fitted and compared to the simulated data using MultiNest.

Through this process, it is shown show that the Bayesian evidence of each model can be estimated for every dataset in a feasible timescale, and to high accuracy and precision. It is further shown that the Bayes evidence estimates obtained from MultiNest allow one to determine which model actually produced the data with high certainty. Moreover, that certainty is greater than that obtained through other inference methods (information criterion methods) that trade precision for speed through simplifying assumptions. It is therefore concluded that MultiNest is *prima facie* suitable for application to response time models for the purposes of both model selection and parameter estimation, and further bolster the evidence of its advantages over other, more

broadly used methods.

The next three chapters work towards the application of MultiNest to new models of simple responding that account for the foreperiod effects not incorporated into existing models.

Within that context, the primary purpose of Chapter 7 is to detail the empirical data which will be used in the model inference and to show that it exhibits the features expected of such data; to this end, I report the results of a simple reaction time experiment performed by 18 subjects and show that the findings are consistent with those reported in the prior literature when considered in light of the SRT's major phenomena: response time distributions, hazard functions, and the various foreperiod effects.

An auxiliary role of this chapter, somewhat separate from the main threads of the thesis, is to underscore certain difficulties in the interpretation and analysis of SRT data which re not usually discussed, and to the offer methods for their resolution. Most notably, I will foreground a phenomenon that will be referred to as *corrected responding*, occurring when a subject responds (in error) before the appearance of a response signal and subsequently responds correctly upon signal onset, which may contain insights into subjects' self-monitoring of performance.

Chapter 8 develops new models of the SRT which incorporate foreperiod effects. These will be developed by elaborating existing evidence accumulator models. Two new classes of response time model are introduced, *foreperiod parameter-dependence* and *collapsing threshold* models, each of which introduces foreperiod-modulated time-dependent decision thresholds into the standard EAM framework.

Chapter 9 contains the confluence and conclusion of the two major lines of argument developed so far in this thesis, namely, the argument for the adoption of nested sampling techniques in cognitive science, and the argument for the need for new and better models of one-choice – especially simple response time (SRT) – task performance. In this chapter, I apply MultiNest to test models of simple responding against real simple reaction time data. The models include traditional models of the SRT as well as the new, more complex models introduced in the last chapter that are designed to account for foreperiod effects. This thorough model comparison shows that existing models do not provide the best accounts of the data; rather, the novel, more functionally complex models, based on more elaborate theories of cognitive processing, consistently obtain higher evidentiary support.

From these basic results obtained, three immediate conclusions are drawn. First, MultiNest is vindicated in practice as an valuable tool for modelling work in cognitive science. Second, the dominant evidence accumulator framework of response time and decision making requires revision and extension if it is to continue to be scientifically useful. Third, our understanding of the basic mechanisms of stimulus response is deepened through the insights gleaned from the new models; in particular, it seems that temporal preparation effects directly modulate the signal-detection and decision-making aspects of cognition.

The final part, "Conclusions", comprising the final chapter only, develops the ideas outlined in §1.2:3. Using the results of Chapter 9 as groundwork, the broader argument is made that cognitive data exhibits an irreducible complexity that cannot be ignored in modelling work, and that future advances in understanding cognition can be driven by adopting a combination of (a) advanced tools for inference of the sort used here and (b) new guiding principles for theoretical work based on the acknowledgement of the irreducible complexity of cognitive data.

Ideas for future work are offered in three main areas. First, some potential lines of research in the just-outlined broad vision of future cognitive modelling work are sketched out. These include modelling and model comparison in other cognitive tasks, such as the stop-signal task; the simultaneous modelling of and inference from multiple tasks; and extensions from the focus here on pure cognitive science and behavioural data to cognitive neuroscience and multimodal data. Second, improvements and extensions to the toolkit and practical methods , including improving the computational implementation of cognitive models and the utilisation of other advanced nested sampling algorithms. Finally, practical uses are suggested for SRT models of the kind developed in thesis, with special consideration of their potential clinical applications.

But before we get too far ahead of ourselves, let us turn to the first set of questions we need to answer: What is Bayesian inference, and why is it important?

—ω—

Part I

Foundations

Science does not aim at establishing immutable truths and eternal dogmas; its aim is to approach the truth by successive approximations, without claiming that at any stage final and complete accuracy has been achieved.

— Bertrand Russell, Skeptical Essays

CHAPTER 2

Prolegomenon to MultiNest: Induction, statistics, and Bayesian inference

Synopsis. Despite growing interest in and use of Bayesian methods in the cognitive sciences, and much statements of their formal aspects, there have been few reviews for cognitive scientists of the essential ideas and logical structure underpinning Bayesian inference. This chapter is intended as a guide for the Bayesian initiate: it aims to show, starting from elementary ideas, how the Bayesian framework takes the basic results of probability theory and makes from them a completely general and coherent set of principles for performing inference with statistical data and models.

The view given here, heavily influenced by the work of Jeffreys (1961) and Jaynes (2003), is of Bayesian statistics as providing rules for performing *inductive inference*, that is, the kind of reasoning that allows us to assess the support lent by data to competing hypotheses – the evidence for or against them. Key to this exposition will be the notion that statements of confidence in a proposition can be expressed as probabilities, which allows the formal tools of probability theory to be brought to bear on induction. Armed with this notion, it will be shown how the Bayesian framework accommodates both of the most common classes of statistical problem, *parameter estimation* and *model selection*.

This will provide the background necessary to understand the technique of choice for Bayesian inference in this thesis, the nested sampling algorithm known as MultiNest, which will be covered in the next chapter. It will also set the terms required to understand the full import of the arguments of later chapters. Particular emphasis will be placed on a concept that will be the linchpin in later chapters – that of a principled, mathematically rigorous, and direct measure of the extent to which the data support one model over another: the Bayes factor.

Author's note: This chapter should be read as a synoptic overview – and therefore as something of a personal view – of the basics of Bayesian statistics, as gleaned from a synthesis of numerous sources. Beginner readers interested in learning more about the subject are directed to the introductory texts by Sivia (2012) and Lambert (2018); more advanced readers to D'Agostini (2003); and everyone to the foundational texts by Jeffreys (1961) and Jaynes (2003). For more on inductive reasoning, the reader is directed to the "Deductive and Inductive Logic" and "Confirmation and Induction" entries in Internet Encyclopedia of Philosophy (Huber, 2022; IEP Staff, 2022), and the "Inductive Logic" and "Problem of Induction" entries in the Stanford Encyclopedia of Philosophy (Hawthorne, 2021; Henderson, 2020).

1. INDUCTION AND THE CONFIRMATION OF HYPOTHESES

1:1 The nature and function of inductive inferences

At the heart of the scientific method lies the process of *inductive inference*, also known as *inductive reasoning* or simply *induction* . Induction – like its counterpart, deduction – is a form of argument and method of reasoning, meaning that a number of statements (*premises*) are put forward and another statement (the *conclusion*) drawn from them, with the premises intended to provide support for the claim in the conclusion. In contrast to deductive inferences, which – by virtue of their strict motions from statement to statement according to the rules of logic – makes their conclusions certain, inductive inferences offer no guarantees about the veracity of their conclusions;[i] the premises of an inductive argument are instead meant to provide *partial*, though not necessarily insubstantial, support to their conclusions. To wit, deduction deals with what is certain, induction with what is plausible; where a deductive argument can be *valid* or *invalid*, an inductive one can be *strong* or *weak*, with the strength of an inductive argument being just the extent to which its premises provide support to its conclusion.

Induction helps us to move from the specific to the general, from experimental data to reasonable conjectures and hypotheses, and, at its best, from observed regularities in nature to the natural laws which underlie them. Despite its intrinsic uncertainty, induction is both necessary and useful: necessary, since it alone allows for the development of successful general theories, and useful, for despite falling short of certainty, strong scientific evidence can still – indeed, can only – be developed by the proper application of inductive methods. Since our scientific focus is on the problem of providing good explanations of a phenomenon, we will restrict our attention to a particular subset of inductive arguments relevant to that problem.

Within the scope just defined, inductive reasoning works by taking a set of observations about a phenomenon, positing a hypothesis that could successfully account for those observations, and drawing a conclusion from the conjunction of the two. The observations consist of specific instances of some general type. The hypotheses can range from simple generalisations – say, "all swans are white" following numerous observations of uniformly white swans – to complex and precise mathematical formulations – say, "there exists a force, gravity, which acts between any two bodies in proportion to the multiple of their masses and inversely the square of the distance between them" following the realisation that the planets have elliptical orbits around the Sun. Statements of the form "It has been observed that (…)" and "The following hypothesis could account for (certain aspects of) those observations" then form the premises of the inductive argument.

From this combination of premises – one detailing particular instances, one proferring a

[i] In both cases, taking the truth of the premises for granted.

general account, which we shall call the *empirical* and *theoretical* premises respectively – it seems, at first glance, that two kinds of inductive argument can be drawn. The first kind seeks to draw conclusions about the observations, be they the instances in the empirical premise or future (and otherwise as-yet unobserved) instances of the same kind. This sort of conclusion provides an answer to the question, "What does the theoretical premise have to say about the observations?" The second kind seeks to draw conclusions regarding the validity of the proposed hypothesis; it answers the question, "To what extent does the empirical premise provide support to the theoretical premise?" To illustrate the difference, compare the statement "The next swan we see shall be white" to "Newton's Law of Gravity is well-supported by the evidence." We will refer to these two types of conclusion, and the arguments which reach them, as *predictive* and *confirmatory* respectively.

However, closer examination shows that predictive inferences ("the sun will rise tomorrow", "the next swan will be white") really involve two moves, first an inductive argument of the confirmatory kind that warrants the use of some model, and then a deduction to arrive at the predictions that model makes. For example, from "the sun has risen every day" to "the sun will rise again tomorrow", there is an induction for the implied theory "the sun rises and sets every single day", from which the predictive conclusion is drawn. The data do not definitively implicate this model; it is in observational terms indistinct from "there is a one in a billion chance the sun will not rise on any given day".

It follows that the key practical issue of induction is to provide a means of evaluating the strength of inductive arguments, that is, a means by which to assess to what extent a given set of premises support their intended conclusion, which in turn entails a method for determining the degree of support certain observations give to a certain hypothesis. Moreover, apropos the central claims of this thesis, the issue of assessing and comparing models can be seen as essentially a particular and specialised form of the issue of assessing inductive arguments, wherein the empirical premise is formed of the quantitative data collected from experiment and the theoretical premise consists of the mathematical model of that data. With that end in mind, the remainder of this section will be given over to the question of assessing inductive arguments in general; this will lay the groundwork for the next section on Bayesian probability theory, which provides an answer to that question.

Before proceeding, we will quickly note that inductive arguments, as understood here, do not give instructions on how to construct theories of a phenomenon. The theoretical component has to be provided independently. The process by which general accounts of a phenomenon are obtained from particulars – that is, the process of theory or hypothesis formation – will be treated here as a black box. On a pragmatic level, we know that we are capable of forming such theories; the concern of induction is rather with evaluating theory in light of the knowledge granted by evidence, and evidence in light of the knowledge derived from theory.

1:2 Evaluating inductive arguments

We are seeking, as it were, the rules of an "inductive logic" to provide principles by which one can infer the degree of support the evidence provides to various claims. To be precise, we will call the extent which the empirical premise supports the posited hypothesis the *credibility* of the hypothesis. It is, recast in this language, absolutely essential to *any* inductive argument be able to determine the credibility of the hypothesis. This determination in turn requires a theoretical framework to inform it, which must specify

(I) what forms the the premises may take;
(II) some measure for the credibility of the hypothesis;
(III) and a means to calculate the credibility.

As we shall see in the next section, Bayesian probability theory specifies just such a framework.

Before we turn our attention to Bayesian probability theory, we will make a few remarks on the general desiderata we would expect a successful inductive logic to satisfy. We note that the credibility of a hypothesis with respect to the evidence can intuitively be viewed as arising from a combination of various factors; key amongst these factors are the answers to the following questions:

1. *Is the hypothesis well-defined enough that it makes definite predictions about which observations should and should not occur?* If the posited hypothesis could account for the actual observations, but could equally well account for numerous other potential observations, its ability to explain or describe its object phenomena is limited. Why should this set of observations, as opposed to some other set, be the one we in fact observe?

2. *Can the posited hypothesis properly account for all the known facts, and any others which, although excluded from the analysis, might be relevant?* A hypothesis which is unable to fully account for its topic cannot be more than a partial explanation or description; this nevertheless does not mean that the hypothesis is not useful, and it is reasonable to assign weight to the hypothesis in proportion to the degree to which it is capable of accounting for its topic. Furthermore, any credible hypothesis cannot be selective in the data it chooses to explain; once a general phenomenon is identified, the hypothesis ought to be able to account for the specific instances of that phenomenon and everything that falls within its scope. This question is also closely related to the question of predicting future data, which can conceptually be viewed as evidence currently "excluded" from analysis.

3. *Relatedly, does the hypothesis make novel and testable predictions (which turn out to be true)?* While not strictly essential for establishing its credibility, a hypothesis which makes successful (and precise) predictions about what should be observed in circumstances not yet seen can claim additional support from the evidence. Positive predictions are scientifically

meaningful threefold. First, they provide a comparison point to other hypotheses which would allow for successful differentiation between them. Second, they guide the search for additional evidence with which to determine the support given to the hypotheses. Finally, making a prediction definitely avoids the danger of *ad hoc* modifications to a hypothesis to make it conform to existing evidence.

4. *Does the hypothesis meet the demands of parsimony, to be as simple as possible, but no simpler?* Intuitively, claims which are more complex are *a priori* less likely to be correct, since they require more factors to obtain / sub-claims to be true than simpler claims, and it is less likely that more components are all true, as the complex claim necessitates, than that fewer components are true, as a simpler claim necessitates. The only warrant for introducing complexity, then, is that in doing so a proportional or greater increase in evidentiary support is gained.

A strong inference – *viz.* a well-formed hypothesis – will answer all those questions positively. A weak inference will fail under scrutiny by one or more of these measures. A successful inductive logic therefore should assign its credibility score in a manner that is congruous with these intuitions. The reader is encouraged to keep these factors in mind in what follows.

This concludes our survey of the fundamentals of inductive inference. The next two sections will outline two approaches to issue of how to successfully perform inductive inference, starting with Bayesian probability theory, and followed up by statistical inference.

2. THE BAYESIAN APPROACH TO INDUCTION

The Bayesian approach to induction identifies the measure of the credibility a hypothesis with a formalised mathematical version of the commonsense notion of plausibility (Jaynes, 2003; Jeffreys, 1961). This section therefore begins with a careful conceptual analysis of plausibility, which we shall also refer to as *confidence*, and its relationship to probability theory. The mathematical formalism obtained thereby will be brought into the subsequent exposition of the inductive capacity of Bayes' theorem, which, we will see, consists of its ability to assign a meaningful metric of credibility to hypotheses of a certain sort. This discussion of the eponymous theorem will later form the starting point of our treatment of Bayesian inference.

2:1 Confidence, plausibility, and Bayesian probability

In broad terms, "confidence" and "plausibility" are here used in their everyday sense, as the degree of belief or reasonable expectation placed in a particular statement.[ii] On an intuitive level, the

[ii] The concept denoted here by "confidence" or "plausibility" is referred to by many names in different contexts, including "credence", "degree of belief", and "evidentiary support"; for consistency's sake, we shall refer to it here exclusively as *confidence* or *plausibility*, and reserve those terms for use in the sense just mentioned.

concept can be summed up simply as the odds we place in something being the case or occurring, or as the answer to question, "How likely is it that statement X is true?" The notion of plausibility is axiomatic to Bayesian statistics and rooted in common experience; we shall now formalise it and give it technical substance. Crucially, we shall see that statements of plausibility can be treated as statements of probability.

On the basis of the broad definition above, let us grant that, in any formal account, we desire confidence (or plausibility, denoting the same) to have the following properties:

(a) Confidence is a quantity, expressible as a number, and assigned to a statement. To see this, it is sufficient to realise that we talk of being more or less confident in one thing than another, or of finding the same statement more or less plausible following new events and observations, which relations can be captured numerically. It is also clearly absurd to suppose that confidence inheres in facts themselves, as opposed to statements of fact: to give an extreme illustration, it is either true or false that extraterrestrial life exists, regardless of one's confidence in the truth of the statement.

(b) Confidence is always assigned against a set of background information and assumptions, even though this will often be done implicitly. Therefore, when we talk of the plausibility of some statement X, we mean the plausibility of statement X with respect to background information and assumptions I. This background should determine, or at least constrain, how plausible we find various statements. Concomitantly, if the assumptions used or the information available changes, the degree of confidence held in various statements should also change; moreover, an assessment of plausibility should change only under those conditions.

(c) Confidence must be distributed over a set of possibilities comprising mutually exclusive and exhaustive statements, since when we talk of our confidence in some statement, it is, at least implicitly, placed in contrast to something else – at the very least, the negation of the original statement. Part of the background assumptions going into an assessment of confidence may well be that the mutually exclusive possibilities being considered are indeed exhaustive.

(d) The total quantity of confidence must be delimited; that is, with regards to a particular set of (presumed) exhaustive and mutually exclusive statements, it is fixed and finite. It makes sense to talk of having complete, but not of infinite, confidence in a particular statement, and we must be certain (that is, have full confidence) that *something* is the case.

(e) Two mutually exclusive possibilities cannot share the same unit of confidence; in the extreme, we cannot be completely confident in both of two mutually exclusive things. Equivalently, we can say that confidence is additive: given two mutually exclusive state-

ments, X and Y, our confidence in the statement formed by the conjunction of the two should be the sum of the confidences of X and Y considered individually.

(f) Confidence must be expressed positively, in that it is either present to some degree or lacking entirely with respect to some particular statement, but it is in any case incoherent to talk about negative confidence.

Granting these properties to confidence, and letting $p(X)$ denote the confidence given to statement X (bearing in mind that more precisely, we mean $p(X \mid I)$, using the notation above) and $\mathfrak{X}$ denote a set of mutually exclusive and exhaustive statements, we can without loss of generality express the following constraints on the distribution of confidence across as a set of statements:

$$\begin{cases} \sum\limits_{X \in \mathfrak{X}} p(X) = 1 & \\ p(X) \geqslant 0 & \forall\ X \in \mathfrak{X} \\ p(X_1\ \&\ X_2) = p(X_1) + p(X_2) & \forall\ X_i \in \mathfrak{X} \end{cases} \tag{2.1}$$

Now, these constraints are just those that are placed on and define probability distributions, since probability[iii] too is a delimited, positively expressed quantity which must be partitioned over a fixed set of possibilities.[iv] Put plainly, every feature we desire of probability distributions, we also desire of distributions of confidence. This structural homomorphism[v] means that we are mathematically entitled to treat distributions of confidence, which encode the relative plausibility of a set of statements, as probability distributions; to wit, we may speak of the "probability" of a particular statement X just as well as we may speak of our confidence in it.[vi] [vii] As we shall

[iii] We use "probability" here in the more usual sense presented in most textbooks, namely, as the limiting proportion of outcomes of a random experiment which satisfy a particular condition; a random experiment is understood to be an infinitely repeatable process that on each trial results unpredictably in one, and only one, of a well-defined set of possible outcomes.

[iv] This is not a mathematically complete and rigorous statement of the conditions for a probability distribution per the mathematical standard – the so-called "Kolmogorov axioms" which require the advanced mathematics of measure theory to be stated rigorously – but it does capture the essence of the similarity between the confidences and probabilities. For a more detailed account, see Jeffreys (1961) and Jaynes (2003).

[v] A homomorphism, from Greek *homo* (same) + *morphē* (form), is a mapping between two objects which preserves the relationships between the parts of the mapped-from object in the mapped-to object – that is, objects in a homomorphic relation have the same internal structure.

[vi] It should be noted that this approach to viewing confidence as probabilities, via its desired properties, is the author's own (though the broader idea is not).

[vii] It is germane at this point to remark that this homomorphism can be taken both ways in Bayesian theory. Here, we have gone from considering degrees of certainty, or confidence, to probabilities in the mathematical sense. Roughly speaking, the Bayesian justification for the inverse – for treating probabilities as degrees of confidence – lies in this homomorphism combined with the observation that in everyday use, when we speak of the probability of an event, we do *not* speak in the frequentist sense of the limiting proportion of occurrences of that event in an infinite series of random trials; rather, we talk about the degree to which we expect or believe that certain outcomes will come to pass, i.e., confidence. This does not entail that "we are all Bayesians" in everyday life, or that our everyday treatment of these kinds of beliefs is Bayesian in nature; in fact, one could view the purpose of Bayesian theory as to identify the

now see, the correspondence between confidences and probabilities underpins the application of Bayes' theorem, a basic result regarding probabilities, to the question of induction.

Before we proceed, a brief aside on subjectivity and Bayesian probability: The property of inhering in statements means that confidence is an epistemological, not ontological, property; that is, it relates to knowledge of things rather than things themselves. Consequently, confidence must belong to the same entity that that knowledge belongs to, that is, to the same entity that holds the statement to which the confidence is attached. Insofar as this is the case, and insofar as Bayesian probabilities must be interpreted as belonging to some kind of user or agent, the Bayesian notion of plausibility is "subjective"; that notwithstanding, the Bayesian concept of plausibility should not be equivocated with one's actual feelings of confidence in a statement, any more than the notion of the truth of a statement in deductive logic should be read as a person's belief (or lack thereof) in that statement.

2:2 Bayes' theorem in Bayesian probability theory

We return now to the showing how Bayesian probability theory can help with inductive inference. Having ascertained *that* degrees of confidence can be treated as degrees of probability, there arises the question of precisely *how* confidence should be divided amongst a collection of statements – that is, what numerical value should be given to the confidence placed in (or plausibility of) each statement, given the relevant background information and assumptions. One of the key insights of Bayesian probability theory is that, with the view of confidence-as-probability, the crux of Bayes' theorem is that it provides an important part-answer to the above question: given the initial degree of confidence we hold in all of a set of statements, the theorem governs how the confidence we hold in those statements should change in the face of new information. In simple terms, Bayes' theorem can be used to tell us how we should adjust our beliefs in the face of new evidence. More precisely, it tells us precisely how much we should rationally change the degree of confidence we assign to various statements, given that certain events or outcomes have been observed, and presupposing a certain set of background assumptions.

Of course, the question of how plausible various statements are in the face of particular information is closely related to the main practical issue of induction, *viz.* determining the degree of support that the premises of an argument lend to its conclusion, and especially determining the credibility of a theoretical premise given a certain empirical premise. Bayesian thinking approaches induction by identifying the the ratio of the degrees of support for the conclusions of two inductive arguments with the ratio of the their Bayesian plausibilities. Now suppose we let the statements being compared be the conclusions of two confirmatory inductive arguments with different theoretical premises and the same empirical premise. Each conclusion has its associated

rational basis for the treatment of such beliefs and to show that such a treatment is mathematically coherent (Jaynes, 2003; Jeffreys, 1961).

credibility in light of the empirical premise. By identifying the inductive credibility ratio with Bayesian confidence ratio, Bayes' theorem then allows us to establish the relative credibility of a number of hypotheses with respect to certain evidence. The major aim of this section is thus to explain *how* Bayes' theorem manages to do this.

The mathematical statement of Bayes' theorem is

$$p(H \mid E) = \frac{p(E \mid H)\ p(H)}{p(E)} \tag{2.2}$$

where E and H are events from sample spaces $\mathcal{E}$ and $\mathcal{H}$ which have a joint probability distribution defined over them (for the unfamiliar, see appendix B for a basic outline of the logic of Bayes' theorem as a purely mathematical statement). Each component of the formula has an associated technical term: $p(H \mid E)$ is the *posterior*, $p(E \mid H)$ is the *likelihood*, $p(H)$ is the *prior*, and $p(E)$ is the *marginal likelihood*.

Stated as such, Bayes' theorem is a straightforward result of probability theory, and seems to contain no special content with regards to inductive inference. Its full significance is brought out only through the careful application of a particular interpretation of E, H, and the various terms in equation (2.2). We begin with a brief overview of this interpretation before giving a fuller explanation.

The first step in the inductive interpretation of Bayes' theorem is to let E denote a collection of observable events drawn from a well-defined set of similar such events $\mathcal{E}$; the E could be the outcome of some procedure, data from an experiment, and so on, with the $\mathcal{E}$ being the set of all conceivable outcomes from that procedure or experiment, etc. It is presumed that there is some true underlying but unknown generating process G which gives rise to the observations in E. In turn, H denotes a statement regarding a particular hypothesis, condition, or state drawn from a well-defined set of such hypotheses (etc.). The H is intended to describe, predict, explain, or otherwise account for the observed events E and others of the same kind generated by G. We will (suggestively) refer to the empirical component E as the "event" or "evidence", the theoretical part H as the "hypothesis", and the generating process G as the "phenomenon".

Continuing this interpretation, the prior encodes the current degree of confidence held in each hypothesis being the best available account of the phenomenon, i.e. of G; the likelihood gives the probability of particular evidence E being observed, assuming a particular hypothesis H holds; the posterior gives the new degree of confidence that should be held in each hypothesis being the best available account, if the aforementioned evidence is in fact observed; and the marginal likelihood plays the role of a normalising factor. If new evidence E shows an increase in the relative value of one hypothesis H_1 against another H_2 when their posterior ratio is compared to their prior ratio, then it is said that E supports H_1 over H_2, with extent of the change in the ratio being indicative of the strength of the support.

The functioning of Bayes' theorem as a tool for inductive inference is clear: the evidence E form the empirical premise of the inductive argument; the hypothesis H is the theoretical premise, the account of those events; and the posterior ratios provide the conclusion, namely, the extent to which the empirical evidence supports one hypothesis over another. In the terminology of the previous section, Bayesian statistics is concerned with providing a framework for making inductive arguments with confirmatory conclusions by calculating the relative credibility of various hypotheses.

Some mathematical remarks

A few brief comments should be made before we take a more detailed look at the nature and function of each component of Bayes' theorem in light of this interpretation. To begin, we note that there are three mathematical prerequisites for rendering Bayes' theorem sensible. The first is that there must exist a fixed set $\mathcal{H}$ of mutually exclusive hypotheses (or conditions, etc.) H which are being taken under consideration. Second, there must be a fixed set $\mathcal{E}$ of mutually exclusive basic events (or outcomes, etc.) E' that can be observed, from which E can be constituted. Finally, there must be a map $\mathcal{M}$ which, for each possible hypothesis H, returns a probability distribution over all the possible basic events E'. Stated in mathematical notation,

$$\exists\ \mathcal{M}\quad :\quad \mathcal{H} \longrightarrow \mathbb{P}_{\mathcal{E}} \quad \text{where} \quad \forall\ H \in \mathcal{H},\ \exists\ P_H \in \mathbb{P}_{\mathcal{E}} \tag{2.3}$$

and where $\mathbb{P}_{\mathcal{E}}$ is the set of probability distributions over $\mathcal{E}$. The reasons for and consequences of these requirements will become clear in what follows.

It is also worth remarking that notation of the form $p(X)$ in equation (2.2) should not be read as the value of a single mathematical function p with input argument X; rather, it should be read straightforwardly as the statement "the probability of X holding (or occurring)" and, according to context, also the value between zero and one assigned to that probability statement. There are instead four distinct mathematical functions involved, one each for the prior, likelihood, posterior, and marginal likelihood:

$$\begin{aligned}
&\textit{Prior} && \pi : \mathcal{H} \to (0,1), && \mathcal{H} \mapsto \pi(H) \equiv p(H) \\
&\textit{Likelihood} && \mathcal{L} : \mathcal{E} \times \mathcal{H} \to (0,1), && (E,H) \mapsto \mathcal{L}(E,H) \equiv p(E \,|\, H) \\
&\textit{Posterior} && \mathcal{P} : \mathcal{H} \times \mathcal{E} \to (0,1), && (H,E) \mapsto \mathcal{P}(H,E) \equiv p(H \,|\, E) \\
&\textit{Marginal likelihood} && \mathcal{Z} : \mathcal{E} \to (0,1), && E \mapsto \mathcal{Z}(E) \equiv p(E)
\end{aligned} \tag{2.4}$$

The exact nature of these functions will be covered below.

2:3 The elements and mechanics of Bayes' theorem for induction

The prior and posterior

Let us now turn our attention to examining the components of equation (2.2) in detail, starting with the prior and posterior. Read naïvely, the prior $p(H)$ is "the probability of hypothesis H." Compared to a more mundane statement, such as "the probability of flipping heads," this may seem like a strange statement; after all, unlike the results of coin flips, there is no cogent sense in which we can randomly sample hypotheses. This is where the concept of confidence-as-probability comes into play: a more productive way to look at the prior is as *the distribution of confidence the user of Bayes' theorem has in the relative ability of each hypothesis to account for the known facts, within the set of hypotheses under consideration.* We then have that $p(H)$ encodes the proportion of confidence assigned to an admissible hypothesis H before any (new) evidence E is received; $p(H)$ is often called the "prior belief" in H.

The posterior $p(H \mid E)$ is the counterpart to the prior *after* the evidence E has been observed; that is, it is a distribution over the hypotheses under consideration which describes the degree of confidence imputed to each hypothesis H after event E has been observed. Analogously to the prior, $p(H \mid E)$ is called the "posterior belief" in H. It is clear that the posterior is the essential component of Bayes' theorem as far as induction is concerned, for it encodes the confidence we ought to hold in, and therefore the credibility of, various hypotheses when given certain evidence.

It must always be kept in mind that, per the first mathematical prerequisite mentioned earlier, the process of assigning confidence is done across a predefined set of possible hypotheses $\mathcal{H}$; without this, it would be clearly impossible to talk consistently of the probability assigned to a particular hypothesis. Stated in the language of confidence assignment, this restriction says that part of the background assumptions for the assignment of confidence is that the hypotheses under consideration form a mutually exclusive and exhaustive set.

This point has an important conceptual consequence: Bayes' theorem, when used for induction, is fundamentally comparative in nature. More precisely, the principal measure of the plausibility of the hypotheses, the posterior, is a metric of the *relative* abilities of the proposed hypotheses to account for the available evidence, rather than a metric of the absolute ability of each hypothesis to account for the data. This is because, within a Bayesian induction, the fixed set of hypotheses is all there is under consideration, and, furthermore, we *must* apportion our full confidence over the available hypotheses, no matter what they are or how closely they fit the data. There is therefore nothing incoherent in comparing a number of bad hypotheses, and finding one which nevertheless has much higher posterior values than others. This may seem like a fundamental problem with the technique, but is in fact a natural feature of constructing explanations of phenomena; this point will be discussed in detail (later).

The comparative nature of Bayes' theorem means that we cannot interpret the posterior

probability of some particular hypothesis H given evidence E in isolation; the posterior value of a hypothesis by itself is not, in any sense, its "credibility score." We can, however, ask about how credible one hypothesis is compared to another; this is simply encoded in the ratio of their posterior values. In the Bayesian framework, we are made to place bets (as it were) on which hypothesis is *best* within a set, not which is *good* in some absolute sense; put differently, the statement $p(H)$ should not be read as the confidence we have in H being true, but in H offering the best account of the phenomenon out of all the members of $\mathcal{H}$.

Our final remark on the credibility of hypotheses within a Bayesian framework: Bayes' theorem, being the means for calculating changes in confidence, does not dictate the prior distribution of confidence that should be placed in each hypothesis. A full discussion on the topic of how to properly assign prior distributions lies beyond the scope of this work; for now, it will simply be remarked that different prior distributions result in different posterior distributions, but that, through the repeated application of Bayes' theorem, the posteriors of two different initial priors will (usually) converge as more evidence is presented. Nevertheless, it must always be borne in mind that the posterior distribution over a set of hypotheses is also partly determined by the prior distribution; hypotheses which seemed more likely to begin with will, all else being equal, still seem more likely after evidence has been presented. In plain language, the more convinced of a hypothesis you are to start with, the more evidence will be required to change your mind.

This reliance on the prior is a double-edged sword: it allows for the incorporation of prior information about the hypotheses where it exists, but can make choosing a prior difficult in its absence. The topic of properly choosing a prior distribution, and whether this is possible, is a subject of active debate which lies beyond the scope of this work. In practice, workable priors which give rise to useful posteriors are relatively straightforward to implement. In the context of the research presented in this thesis, we will therefore deal with the assignment of prior distributions on a case-by-case basis, rather than proceeding from some more abstract account.

The likelihood

The likelihood $p(E \mid H)$ is the probability of observing the event E if or when the hypothesis H holds. The ability to calculate the likelihood relies upon the remaining two mathematical prerequisites, which we recall are, first, that there is a well-defined set of events to be accounted for, and second, that there is a statistical model which assigns to each condition a probability distribution over those events, from which the likelihood value can be found or calculated.

The notation of the likelihood can be misleading; we therefore remark here that, while the value returned by the likelihood function $\mathcal{L}$ is always a probability, the function itself is not utilised as a probability distribution. Recall from equation (2.4) that the likelihood function $\mathcal{L}$ is defined over both events and hypotheses. When the hypothesis H is fixed, evaluating the

likelihood function $\mathcal{L}(E; H_{\text{fixed}})$ returns the probability distribution over the events as predicted by H_{fixed}. When the event E being accounted for is fixed, however, $\mathcal{L}(H; E_{\text{fixed}})$ is *not* a probability distribution; it is instead a function over the hypotheses, the value at each one being simply the probability of E_{fixed} occurring as given by the probability distribution associated with H.

In the context of Bayesian statistics, the calculation of the posterior belief calls for evaluation of the likelihood function for every hypothesis with a fixed event in mind; as the above considerations make clear, while the desired function $\mathcal{L}(\mathcal{H})$ will be derived *from* probability distributions, it will not itself be a probability distribution. The matter of calculating likelihood functions is fundamental to performing Bayesian statistics, and the particulars will be taken up at length in the section on Bayesian inference.

Note also the restrictions that the likelihood function places on the use of Bayes' theorem: every hypothesis *has to be* sufficiently well-defined that it produces a definite probability distribution over the possible events. Hypotheses which are not so defined cannot be the subject of inference by Bayes' theorem.

Bayesian updating

With this understanding of the prior, posterior, and likelihood, it is possible to fully grasp the mechanics of inductive inference under Bayes' theorem. Following equation (2.2), upon observing event E, our confidence in condition H should be adjusted to be proportional to the probability of observing of E under H and to the degree of confidence held in H before (that is, *prior to*) observing E; more succinctly, our posterior belief in H should be proportional to the likelihood of E under H multiplied by our prior belief in H. The relative credibility of one hypothesis to another can then be expressed as the ratio of the posterior value of the first to the second. Clearly, this process can be applied repeatedly as new evidence is provided, with the old posterior being used as the new prior.

Bayes' theorem results in a number of common-sense outcomes. First, if we take a so-called *uniform* prior, with equal confidence being assigned to all the hypotheses being admitted into consideration, then the condition with the highest posterior probability is simply the one with the highest likelihood, as one would expect: the condition we should have the highest confidence in is the one most likely to result in the outcome being observed, all else being unknown. When we choose a non-uniform prior, the prior simply acts as a weighting on the likelihood; as the commentary on iteratively applying Bayes' theorem demonstrates, this is nothing more than admitting into consideration any previous evidence, so that the most probable condition is that which is most likely to generate both the new observations under consideration and the previous observations encoded in the prior distribution.

The marginal likelihood

The marginal likelihood plays the role of a normalising factor. The prior and the posterior, as mentioned, are proper probability distributions. This entails that they satisfy the summation condition

$$\begin{cases} \sum\limits_{H \in \mathcal{H}} p(H) = 1 \\ \sum\limits_{H \in \mathcal{H}} p(H \mid E) = 1 \end{cases} \tag{2.5}$$

Since the posterior value of each hypothesis H is its prior value weighted by its associated likelihood, and since the likelihood for each posterior value can in principle take on any value between zero and one, the only way for the posterior summation condition to hold is if the likelihood for each hypothesis is one, namely, if each hypothesis predicts the event being accounted for with certainty. This is, needless to say, an extremely implausible (and trivial) scenario, so the posterior needs scaling to satisfy the summation condition. In order to preserve the essential information encoded in the *relative* values of the posterior beliefs across the hypotheses, the scaling needs to be uniform across the same. The scaling term which satisfies these conditions is the marginal likelihood, and it can straightforwardly be shown that it is

$$p(E) = \sum_{H \in \mathcal{H}} p(E \mid H)\ p(H) \tag{2.6}$$

That is, in the context of Bayesian statistics, the marginal likelihood is the *expectancy of the evidence:* the weighted average probability of obtaining E from all the hypotheses under consideration, where the weighting for each hypothesis is determined by its prior belief.

While the marginal likelihood is, formally speaking, a scaling term, it plays a significant role in Bayesian inference; this will be covered in §2.4.

3. STATISTICAL INFERENCE

In this section, we will present the elements of the theory of *statistical inference*, which – alongside Bayesian probability theory – is one of the sources that feeds into Bayesian inference. The chief concerns will be to give precise formulations of the subject matter of statistical inference and to present the tools and methods used to handle that subject matter. To these ends, this section contains detailed discussions of two of the central topics of this thesis: data and statistical models. These discussions will be used to give bounds and substance to two of the principal classes of problem which statistical inference (including Bayesian inference) addresses, *parameter estimation* and *model selection*.

3:1 The scope of statistical inference

The methods of *statistical inference* are concerned with drawing conclusions about phenomena of interest from observed data about those phenomena; in particular, statistical inference is concerned with doing so where the data exhibits unavoidable and indeterminate variation or is otherwise subject to considerable uncertainty (Johnson, 2011). We will refer to such data as *statistical data*, and remark for later that we expect statistical data to comprise a series of observations of the same kind, with an ordered list of quantitative measurements composing each observation. Clearly, the root problem that statistical inference addresses is the same as the fundamental problem of empirical science (cf. the General Introduction): namely, that while data are the *only* definite source of knowledge about the process underlying a phenomenon, they by nature provide only *partial* knowledge. From this stems the challenge of creating a reliable description of the underlying process using the data alone as the guarantor of reliability.

We have been pursuing a line of thinking that emphasises the role of inductive reasoning in solving such problems; we will therefore frame the fundamental aim of statistical inference as to provide a language and a logic for inductive inference in those cases where the empirical content that forms the basis of the inference is statistical data, and where we wish to make inferences about the underlying data-generating process. By a logic, we mean a framework for drawing conclusions and evaluating the statements of an inductive argument; by a language, we mean a kind of basic grammar from which statements can be constructed in the first place. As we have already seen, a logic of inductive inference can be founded on Bayesian principles; the elaboration of that logic to cover statistical cases is the subject of Bayesian inference. Therefore, this section is mostly concerned with establishing the language of statistical inference: a way to form meaningful statements upon which the laws of reasoning under uncertainty, as codified in Bayes' theorem, can operate.

Statistical propositions are made using *statistical models*. The constructs within a model provide the basic grammar with which inductive questions about statistical data and their underlying processes can be asked. Like all models, statistical models are tentative mathematical representations of the underlying process, formally embodying a conceptual scheme for describing, explaining, or otherwise accounting for some phenomenon and its associated data, through the specification of the relationships between various quantities, both observable and hypothetical. Statistical models therefore act as the bridge between empirical data and the underlying process.

The basic presupposition of all specifically *statistical* modelling is that the underlying data-generating process, at least for certain dependent variables, is stochastic (that is, subject to random variation), meaning that each datum of the dependent variables in the data-set should be treated as a random variable. The data-set as actually observed is thereby considered to be the outcome of a specific realisation of the underlying stochastic process, wherefore the aim of the modelling is *not* to reproduce the observed data-set exactly, but to account for its underlying features, both those

which are deterministic and characteristic of chance. Consequently, the basic requirement of a statistical model of a data-set is to provide a schema which specifies the probability distribution over the possible outcomes of the dependent variables for each observation; those probability distributions are the predictions of the model.

To say more about statistical inference, some terminology, concepts and definitions must be introduced. We start by stating a mathematical formalism for data in general, after which we will lay out the construction of a statistical model for a particular data-set.

3:2 Mathematical formalism

Data, measurements, and metrics

The intuition behind our formalism for a data-set is that data can be arranged in tabular form; one can imagine that the rows of the table correspond to observations of particular instances and that the columns correspond to readings taken of various different measures for each instance. Therefore, for our purposes, a data-set D can be conceived as a collection of p observations, where each observation consists of q measurements which can be recorded numerically, and where the mathematical form and conceptual interpretation of each measurement is the same across all the observations. We will index the observations by i, where $i \in I = \{n \ : \ n \in \mathbb{N},\ 1 \leqslant n \leqslant p\}$, and the measurements by j, where $j \in J = \{n \ : \ n \in \mathbb{N},\ 1 \leqslant n \leqslant q\}$, where I and J will be called the index sets of the observations and the measurements, respectively.

Every measurement records the specific value established by observation against a particular metric. Since we require that each measurement be capable of being numerically expressed, we can mathematise the notion of a metric simply as the set of values which measurements of that metric can potentially realise. Indexing them in the same way as their measurements, we will designate the metrics of a data-set μ_j, where it is understood that each μ_j is a well-defined subset of the integers or real numbers that depends on the nature and operational procedure underpinning the metric.

If the data contains measurements of more than one metric, the first step is to demarcate certain metrics (possibly all of them) as the dependent variables in need of explaining; these are to the ones be treated as random variables. The remaining metrics are taken to be independent variables; their significance will be addressed shortly. Formally, we will reference the chosen dependent variables through a subset of the index set of metrics, $J' \subset J$.

Statistical models

In the definition we will adopt here, statistical models are composed of six major elements:

1. A sample space Ω of possible outcomes based on the dependent variables in the data-set.
2. A set of probability distributions Ψ, where each distribution is defined over the sample

space Ω.[viii]

3. The parameterisation of the family of probability distributions, which can be broken down as follows:
   (a) A set of *free parameters* Θ. The specific parameter set $\theta \in \Theta$ decides the exact function of the model.
   (b) A set of *exogenous state parameters* Ξ determined by the independent variables in the data-set.
   (c) A set of *endogenous state parameters* Σ which are determined by the internal model structure.
4. A procedure Γ for generating predictions by assigning to each datum a probability distribution from Ψ in a manner dependent only on the specific values chosen from Θ, Ξ, and Σ.

A simulated data-set D is conceived of as a sampling or realisation of the set of random variables

$$D = \{X_i \quad : \quad i \in I\} \tag{2.7}$$

where each X_i is the random variable for the corresponding observation i, and where the predictions of the model for a given observation and a given parameterisation is the probability distribution

$$p(X_i = \omega) = \psi_i(\omega;\, \theta, \xi, \sigma) \quad \text{where} \quad \psi \in \Psi,\ \omega \in \Omega,\ \theta \in \Theta,\ \xi \in \Xi,\ \sigma \in \Sigma \tag{2.8}$$

Conceptually speaking, and assuming the truth of the model, the empirical data-set being modelled is one such sampling.

Constructing a statistical model of a data-set

For a given experiment, the set of metrics $\{\mu_j \quad : \quad j \in J\}$ forms the mathematical nucleus of statistical models of that experiment. Recall that our basic presupposition is that data should be treated as the realisation of random process, and that certain measures in the data are construed as random variables. Those random variables are the dependent variables of the model. Having so defined certain measures of interest, the sample space of a statistical model of an experiment is formed as

$$\begin{aligned} \Omega &= \prod_{j \in J'} \mu_j \\ &= \left\{(x_1,\, x_2,\, \cdots) \quad : \quad \forall n \in J',\ x_n \in \mu_n \right\} \end{aligned} \tag{2.9}$$

[viii] For the technically minded, following from footnote iv, probability distributions are technically defined over a set of "events" which form a so-called "σ-algebra" over the sample space. We will omit this technical detail in this work and assume that, in all cases, this procedure can be suitably carried out.

where the product sign indicates the Cartesian product of the sets μ_j. Since each μ_j is just the set of potentially realisable values for its associated metric, the sample space Ω corresponds to the set of all possible measured outcomes of the dependent variables, which is exactly what the statistical model needs to provide predictions for.

Now, recall that statistical modelling conceives each observation as the realisation of a random variable; more specifically, using the formalism just developed, we can now say that each observation can be treated as a combination of an Ω-valued random variable and the vector of values of any remaning independent variables. Since a statistical model must make predictions for the dependent variables, it must ultimately be able to specify, for each observation, a probability distribution over Ω. We denote the set of distributions to which the observation random variables map Ψ.

Broadly speaking, to give a model meaningful structure, the members of Ψ will form a family. By a family, we mean a set such that the distributions in the set all share the same functional form; the exact form chosen will depend heavily on the data being studied, but we can make a number of general observations. First, this family of functions may be parameterised by a set of free parameters Θ, particular values of which (uniquely) select functions out of the family. If free parameters are present, it is necessary to select parameter values $\theta \in \Theta$ to completely determine the distribution assigned to each observation. The free parameters for a model, once selected, are general; they cannot depend on the observation. A simple example of free parameters would be the mean and standard deviation of a normal distribution, drawn from the family of all possible normal distributions.

Second, we may wish to encode in the model certain deterministic dependencies between measures, specifically between the independent variables and the dependent variables. The probability distribution for a particular observation may therefore also depend on the values of the independent variables as given by some mathematical function with its own parameters; we formalise this dependency in general as a set of parameters in the model Ξ that we will call the *exogenous state parameters*. In a simple linear regression model, for instance, the slope and intercept would belong to this set.

Finally, we may wish to model dependencies between observations according to the values they obtain within a particular sampling of the model (as, for example, in a time series where the value at of a measure at $t + 1$ depends on its value at t). The variability of the probability distribution assigned to an observation attributable to such internal dependency, we encode in a set of parameters Σ that we call the *endogenous state parameters*.

3:3 The parameter space of a model

We introduce a concept now that will be used extensively later on, a geometric interpretation of the parameterisation of a model. The free parameters Θ of a model can be thought of as forming

an abstract entity known as the "parameter space" of the model. A model's free parameters are given as an ordered list of n numbers, that is, $\theta = [\theta_1, \theta_2, \cdots, \theta_n]$, where each θ_i is drawn from a certain range allowed by the model. We will take the θ_i to be continuous quantities, so that the allowed range of values of the i^{th} parameter is a subset of the reals, $\Theta_i \subset \mathbb{R}$. The restrictions on this range can be mathematical (for example, negative standard deviations are impossible), practical (we know from experience that the model parameters will fall within a certain range), or theoretical (we choose to restrict the range).

Now, it is well-known that a point in the three dimensions of ordinary physical space can be represented by a list of three ordered numbers, most commonly as x, y, and z coordinates read off three mutually orthogonal axes (we note, one for each dimension); conversely, once a space has been meted out by a set of axes, a list of three ordered numbers can be interpreted as giving the coordinates of one particular point in space. This one-to-one correspondence between points in space and lists of numbers evidently also holds in one dimension (the line) and two dimensions (the plane). Extending the notion, we may interpret *any* list of n numbers as specifying a point in an abstract "space" of n dimensions, with the values of those numbers beings those read off n mutually orthogonal axes in that space. The set of possible values of the free parameters Θ, therefore, can be mapped onto a "parameter space," with each Θ_i forming an axis, orthogonal to all others and delineating one dimension of the space. The range of values in Θ_i specifies the limits of the space along its dimension, as well as the value of θ_i to be read off along that dimension. Therefore, each point in the space represents a particular combination of parameter values $\theta = [\theta_1, \theta_2, \cdots, \theta_n] \in \Theta$ by its projection onto the axes, and conversely, each θ corresponds to exactly one point in the space.

This abstract analogue serves three purposes; by intermixing geometric and symbolic ways of thinking about mathematical objects, both are extended. First, the notion of the dimensionality of space is extended, allowing problems that would otherwise not seem geometric to be handled as such. Second, while it is, of course, impossible to visualise spaces with more than three dimensions, the construction of higher-dimensional spaces is based on the extension of our notions of low-dimensional ones; therefore, the intuitions we have working with more familiar geometries can often be parlayed into higher dimensions, allowing us to have a more concrete grasp on higher-dimensional behaviour. Finally, it is generally more conceptually enlightening, because more intuitive, to work with abstract entities in geometric representation rather than in purely symbolic form; we shall see this when discussing the mechanics of MultiNest in the next section.

3:4 The problems of statistical inference

In statistical inference, inferences about the data, and the process underlying the data's generation, are derived from a comparison of models to data. In the earlier language of inductive argument,

data provides the content of the empirical premise, and models the content of the hypothetical premise. Now, bound by those terms, the investigator's latitude for altering predictions is either to change a model's free parameters, or to select a new model entirely; hence, any conclusions drawn are statements rgearding the plausibility of either the choice of parameters within a model, or the selection of model itself. Hence, there are two key classes of problem within statistical inference: *parameter estimation* and *model selection*, which are the problems of inference within and between models, respectively. More specifically, the first class covers the the problem of determining, given a data-set and a single statistical model, which values of the model's free parameters give the closest fit to the observed data, and thence to make statements about the values of those parameters for the underlying phenomenon. The second class covers the problem of choosing the statistical model which performs best (whatsoever is meant by that) at accounting for the data from a set of candidate models. In the following section, we will apply the framework of Bayesian probability theory to the problems of statistical inference.

4. THE GENERAL THEORY OF BAYESIAN INFERENCE

We have now laid all the groundwork required to build the theory of Bayesian inference. The purpose of Bayesian inference is to evaluate the weight of evidence in favour of (or set against) a scientific theory; this is accomplished through the comparison of predictions made by competing theories. The methods of Bayesian inference draw from those of statistical inference in general: conclusions are drawn about statistical data through the use of statistical models, which provide the semantic content of statistical propositions. The basic idea is to use Bayes' theorem in its capacity (developed before) as an instrument of induction to address the problems of statistical inference. We recall that Bayes' theorem is

$$p(H \mid E) = \frac{p(E \mid H)\, p(H)}{p(E)}$$

where H is some hypothesis which assigns probabilities to events or evidence E. Given a fixed set of hypotheses $\mathcal{H}$, it is possible to calculate the relative plausibility of the various H in $\mathcal{H}$ through their posterior ratios, i.e., $\frac{p(H_1 \mid E)}{p(H_2 \mid E)}$. Recall also that the two major classes of problem in statistical inference are parameter estimation (determining which free parameters of a statistical model best account for given data) and model selection (determining which of a set of models best accounts for given data). The trick to Bayesian inference is to recast these problems in terms of Bayes' theorem. It will now be shown how this is done.

4:1 Parameter estimation and model selection in Bayesian terms

We start with the problem of parameter estimation. We have certain data, D, and we have a model M with free parameters Θ to describe that data. By construction, every set of parameter values $\theta \in \Theta$ furnishes us with a probability distribution over the dependent variables in D. It follows, per the discussion in §2.2, that Θ therefore satisfies the basic conditions necessary for a valid set of hypotheses $\mathcal{H}$ for use within Bayes' theorem; that is, we may treat every point in the parameter space as an individual hypothesis. Furthermore, since the model M at parameters θ makes predictions over the dependent variables for each observation in D, we can treat the observations in D as events or evidence E within Bayes' theorem; namely, D is to θ as E is to H. The final piece in Bayesian parameter estimation is the assertion that the problem of parameter *estimation* is best thought of as the problem of parameter *inference*, and the inference we wish to make is which free parameters of the model are most plausible given the data. The best-fit parameters are then just the most plausible; mathematically cast, $\theta_{\text{best}} \equiv \max\, p(\theta \mid D, M)$. But the expression $p(\theta \mid D, M)$ is, by the correspondences just established, readable as $p(H \mid E)$ – the posterior. It immediately follows that we can make an inference over the parameters via Bayes' theorem as

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)\; p(\theta \mid M)}{p(D \mid M)}, \tag{2.10}$$

where we have explicitly included the conditional dependence on the chosen model M. This can be recast as

$$\mathcal{P}(\theta) = \frac{\mathcal{L}(\theta)\,\pi(\theta)}{\mathcal{Z}}, \tag{2.11}$$

where $\mathcal{P}(\theta) \equiv p(\theta \mid D, M)$ is the posterior belief over the parameter values, $\mathcal{L}(\theta) \equiv p(D \mid \theta, M)$ is the likelihood, $\pi(\theta) \equiv p(\theta \mid M)$is the prior over the parameter values, and $\mathcal{Z} \equiv p(D \mid M)$ is known as the *Bayesian evidence* of the model. The Bayesian evidence goes by a number of other names, the most common of which is the *marginal likelihood*. Both these terms will be used in this work.

Note again that it is absolutely necessary to specify a prior over the model parameters to calculate the posterior. Prior beliefs are transformed into posterior beliefs by considering a hypothesis with respect to data (using the likelihood as the organ of consideration); the posterior thereby represents the strength of the evidence supporting each hypothesis, which is the very aim of the investigation, and which cannot be undertaken without the prior. This point is important enough that we can state, as a general rule, that a model in Bayesian inference *consists of specifiying* both *a likelihood function and a prior distribution.*

The Bayesian evidence, in its normalising role as the marginal likelihood, is given by the

integral

$$\begin{aligned} p(D \mid M) &= \int_{\theta \in \Theta} p(D \mid \Theta, M)\ p(\Theta \mid M)\ \mathrm{d}^n\theta \\ &= \int_{\theta \in \Theta} \mathcal{L}(\theta)\, \pi(\theta)\ \mathrm{d}^n\theta \end{aligned} \tag{2.12}$$

where n is the number of model free parameters. The form of equation (2.11) emphasises the fact that, for a given model and data-set, the posterior ratios depend only on the values of the likelihood and the prior, both in the numerator. For parameter estimation, the Bayesian evidence plays the role of a normalising factor, and can be ignored, since inferences about the model parameters depend only on the posterior ratios of the parameters.

By contrast, the Bayesian evidence is the central quantity of model selection (also known as *Bayesian model comparison* or *hypothesis testing*); it is important enough that we will name equation (2.12) the *evidence integral*. To understand its import, let us cast model selection in Bayesian terms, as the process of making inferences about which model in a set of models is most plausible given certain data. Extending our application of Bayes' theorem, given a finite set of models $\mathcal{M}$, we should like to treat each model $M \in \mathcal{M}$ as a hypothesis, and to be able to state that

$$p(M \mid D) = \frac{p(D \mid M)\ p(M)}{p(D)} \tag{2.13}$$

so that again, the best model is the one which is most plausible, or equivalently, has the highest posterior value. A little thought shows that, following the pattern from earlier, the terms here fulfil the requirements to act as counterparts to the terms in equation (2.2).

We note here that both Bayesian parameter estimation and Bayesian model selection are derived by making an appropriate substitution in the hypothesis term H in (2.2). Further, the expression $p(D \mid M)$ in the numerator of equation (2.13), the "model likelihood" is the same as that which occurs in the denominator of equation (2.10), namely, the Bayesian evidence.[ix] Bayesian parameter estimation and model selection therefore occupy, in a sense, adjacent rungs in a hierarchy of potential kinds of hypotheses. This notion will be extended further in §6.4 and appendix C when dealing with the common situation in cognitive science of having data about a generic phenomenon collected from multiple sources (say, data about response times collected from multiple subjects).

[ix] If you are unconvinced that the expression $p(D \mid M)$ refers to the same quantity in the contexts of both parameter estimation and model selection, we here show that it is. We wish to perform inference on a set of models $\mathcal{M}$ given data D using equation (2.13). For this to be possible in our formalism, we require that each model $M \in \mathcal{M}$ makes well-defined probabilistic predictions over the possible values of D; put differently, it must be possible to calculate the model likelihood, $p(D \mid M)$. We are treating a model M as the sextuple $(\Omega, \Psi, \Theta, \Xi, \Sigma, \Pi)$, using the notation of the previous section. Since a model's only variable aspect is its free parameters, and since at each value of its free parameters, a model by construction makes a well-defined prediction over the data – namely, $p(D \mid \Theta, M)$ – the model likelihood can *only* consistently be conceived and calculated as the parameter likelihood marginalised over the model's free parameters. But this is simply the quantity in equation (2.12), i.e., the Bayesian evidence in the denominator of (2.13).

Returning to the question of model selection proper, in practice, the set of statistical models being considered will often be discrete and finite, so the posterior ratios between models can be calculated pairwise between them. The posterior ratios of two models given the same data is given by

$$\frac{p(M_1 \mid D)}{p(M_0 \mid D)} = \frac{p(D \mid M_1)}{p(D \mid M_0)} \frac{p(M_1)}{p(M_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0} \frac{p(M_1)}{p(M_0)} \tag{2.14}$$

That is, the posterior ratio of two models is proportional to the ratio of their Bayesian evidences; most often, is an important enough quantity to be given its own name, the *Bayes factor*; symbolically, we will designate it as B_{10}. With this nomenclature, we can summarise the result above more succinctly:

$$\text{posterior odds} \quad \propto \quad \text{Bayes factor} \times \text{prior odds} \tag{2.15}$$

Often, the prior belief in each of the models will be the same, so that their posterior ratio simply *is* the Bayes factor, though on occasion greater consideration of the prior belief over the models may be required.

The Bayes factor, then, is a quantity of great moment for science: it is principled, mathematically rigorous, and direct measure of the extent to which the data support one model over another. It is principled, because it is based upon a sound account of how plausible reasoning should work; mathematically rigorous, for it uses nothing but the basic results of probability theory; and direct, because it answers the question of how credible one hypothesis is compared to another through an explicit metric of evidentiary support *per se*, not by proxy. Tying these concepts back to the original discussion of inductive inference, it can be said that the Bayes factor gives a mathematically rigorous solution to the central practical problem of inductive arguments which reach confirmatory conclusions; namely, how to measure the (relative) degree of support which an inductive argument lends to its conclusions. *A fortiori*, where it can be applied, it provides a solution to the fundamental problem of empirical science: how does one pass from the concrete and bounded knowledge embodied in data to knowledge of the underlying processes with any degree of confidence?

4:2 The interpretation of Bayes' theorem for statistical inference

The prior and posterior

The prior $\pi(\theta) = p(\theta \mid M)$ is (mathematically) a probability distribution over the whole range of a model's parameters that (conceptually) describes the researcher's relative confidence in the "true" model parameters falling on those values. Let us be precise about what it is to talk of the "true" value of a model's parameters: it is to say that, presuming that the model is an accurate description of the underlying data-generating process, the underlying process actually possesses those values for the parameters (or equivalent). The posterior, as the evidence-updated prior, tells

us the confidence we ought to have in particular parameters being those actually possessed by the underlying process, assuming the fidelity of the model as a description of reality.

Note that if we start with a uniform prior over a certain range of parameter values, the maximum *a posteriori* parameters (the parameters which maximise the posterior density function $\mathcal{P}(\theta)$) are simply the parameters in that range which return the highest likelihood, as one would expect: the parameters of the distribution from which the data were most probably drawn from are the parameters which are most likely to generate the data set, all else being unknown. When we choose a non-uniform prior, the prior simply acts as a weighting on the likelihood; this is nothing more than admitting into consideration any previous evidence, so that the most probable parameters are those which are most likely to generate both the data we have and the data encoded in the prior distribution. Note that since the posterior takes the data and model as given, the Bayesian evidence, which depends only on the data and the model, is a constant multiplicative factor for the purposes of calculating the posterior, which is thereby directly proportional to the product of the likelihood and the prior.

There is a useful quasi-physical way of looking at the prior (and the posterior) which will come into much use later. In the previous section, we discussed the notion of "parameter space" as a geometric picture of the abstract entity of the set of all possible combinations of parameter values: for a model with n free parameters, we imagined each parameter defining an axis, mutually orthogonal to the rest, in an n-dimensional "space" much like ordinary three-dimensional space. The limits and values of the parameter values also defined the limits and coordinates of the space, so that we could imagine the parameter space as a bounded n-dimensional volume measured by the parameters; since a cube is the like sort of bounded volume in three dimensions, we dub this entity the "hypercube."

The prior is a function defined over a model's parameters, so what is its interpretation inside the volume of the hypercube? In this construction, parameter space compactly represents all the hypotheses under consideration inside the hypercube, and we know we must distribute all available confidence across the hypotheses. By analogy with physical mass, volume, and density, we can picture this distribution by imagining confidence as a quantity of "probability substance" which fills the hypercube and has a total mass of one; the prior then specifies the density of the unit of probability mass at each point in the parameter space.

For the reasons discussed in the section on the general theory of Bayesian probability theory, the question of how to assign a prior in general will not be taken up at length here. That notwithstanding, *apropos* the current discussion, a few general notes and terms of art that will be of use later when assigning priors will be introduced here.

The first term emerges from the fact that very often, the probability distribution that is chosen for the prior will be one with a well-defined functional form dependent on certain parameters; for example, a common prior over a model parameter θ is a normal distribution with mean μ_θ

equal to what is seen as the most plausible value θ and standard deviation η_θ to express the range of plausible uncertainty. The parameters of the prior distribution – as opposed to the parameters of the model over which the prior is defined – are referred to as *hyperparameters*.

The second term of art to be introduced here the *conjugate* prior. For certain parameters of certain likelihood functions, it can be shown that there exists a probability density function, which, if chosen as the prior, will preserve its functional form in the posterior; in other words, when conjugate priors are used, prior updating to the posterior just involves updating the hyperparameters. So, if the prior distribution over a parameter θ is chosen to be the conjugate distribution $\pi(\theta) = f_{\text{conj}}(\theta \mid \alpha_0, \beta_0)$ where $\alpha_0, \beta 0, \ldots$ are the prior values of the hyperparameters, the posterior distribution can be expressed as $\mathcal{P}(\theta) = f_{\text{conj}}(\theta \mid \alpha_1, \beta_1)$ where where $\alpha_1, \beta 1, \ldots$ are the posterior values of the hyperparameters. Generally speaking, when a likelihood function allows the distribution which acts as the conjugate for the prior and posterior to be found, there exist formulae to allow for the posterior hyperparameters to be reckoned from the prior hyperparameters and data, making them convenient for use when the conjugate-possessing likelihood function is used.

Finally, in the absence of strong reasons to believe that the model parameters are centred on certain values, one should use an *uninformative* prior. An uninformative prior is one which aims to express, insofar as possible, complete ignorance about the value of the parameter(s) over which it defined. Conversely, an *informative* prior expresses an a priori preference for, or higher confidence in, certain values of the parameter (and so, in a sense, contains more "information" regarding the actual parameter values; the maximally informative prior is a Dirac delta function, which places absolute confidence in a particular parameter value and so makes it immune to updating). These terms are perhaps better construed not so much as mutually exclusive categories, but as opposing poles on the scale of how broadly distributed *a priori* confidence is in the hypotheses under consideration.

Note that the term "uninformative" here should not be taken to imply anything beyond the indication that the prior is an attempt to profess ignorance about true parameter values. Especially, there is no normative content to the phrase: both informative and uninformative priors can be used wisely. A well-formed informative prior is well-informed by previous experience; it is otherwise simply biased or prejudiced. Similarly, when there is evidence – in, say, previous research – to favour certain parameters, this should be taken into account in the prior; for an uninformative prior, which may seem attractive for being unbiased, would be inappropriately sceptical.

It is worth noting there is no broadly agreed upon and rigorous mathematical definition of an uninformative prior. However, while there is no exact agreement on what constitutes an uninformative prior in general – or even in many particular cases – all attempts to formulate uninformative priors are founded on "the principle of indifference," which states that every

possible outcome should be assigned equal weight. Disagreements regarding uninformative priors often revolve around defining what "weight" means exactly (for examples of this in practice, see §6.3:1); in practice, however, relatively uninformative priors can be approximated well enough by a prior which is uniform over a reasonable and broad range of possible parameters (which can be established through preliminary data analysis and modelling, common-sense constraints, previous experiments, and so on).

The likelihood

In general, our model M will, for each particular set of values of the parameters, predict a specific probability distribution of experimental outcomes. The likelihood at parameter point θ is defined as the probability of obtaining the observed data D under the distribution predicted by M at θ. Supposing then that we have D comprising data points $\{d_1, d_2, \ldots, d_k\}$, the likelihood is calculated as

$$\mathcal{L}(\theta) = \prod_{i=1}^{k} p(d_i \,|\, \Theta, M)\,. \tag{2.16}$$

Since $p(d_i\ p(\Theta\,|\,,)\,M) \leqslant 1$ for all d_i, the likelihood quickly becomes so small as to be virtually un-computable. Therefore, it is common to work with the more convenient log-likelihood:

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log\ p(d_i \,|\, \Theta, M)\,. \tag{2.17}$$

Clearly, given a particular model and data set, the likelihood is a function of the model parameters; we therefore denote the function as $\mathcal{L}(\theta)$, the *likelihood function* of model M. As shown above, the likelihood function in turn depends on what we shall refer to as the *prediction functions*, $p(d_i \,|\, \theta, M)$. Depending on the model, an explicit analytical form of the prediction function(s) may not be known or may not exist, in which case it is necessary to estimate it (or them). Further discussion about the calculation and estimation of (log-)likelihoods can be found in the Methods section; for now, we will take it as given that likelihood values can be straightforwardly obtained for our models of interest.

Conceptually speaking, the likelihood can be viewed as a measure of how surprising or expected the data is, with high likelihood indicating low surprise or high expectancy, and vice versa.

The Bayesian evidence and Bayes factors

We have established that, beyond its role as a normalising factor for the posterior, the Bayesian evidence plays a further role as a measure of relative evidentiary support for a model. Why should the Bayesian evidence $\mathcal{Z}$ tell us anything about the explanatory or descriptive power of a

(Based on) Jeffreys's suggestion			Kass and Raftery's suggestion		
B_{10}	$\log_{10} B_{10}$	*Evidence against* H_0	B_{10}	$2 \log_e B_{10}$	*Evidence against* H_0
1 to 3.2	0 to 0.5	Not worth more than a bare mention	1 to 3	0 to 2	Not worth more than a bare mention
3.2 to 10	0.5 to 1	Substantial	3 to 20	2 to 6	Positive
10 to 100	1 to 2	Strong	20 to 100	6 to 10	Strong
100	2	Decisive	> 150	> 10	Very strong

Table 2.1. Qualitative interpretations of the weight of evidence provided by the Bayes factor. As given in Kass and Raftery (1995), reproduced with permission. The Bayes factor B_{10} is here expressed as the ratio of an "alternative" hypothesis (1) to a "null" hypothesis (0) (note that the labelling of alternative and null is arbitrary in Bayesian terms). It is here assumed that the alternative has the greater Bayesian evidence, so that the Bayes factor is greater than 1 (and the log-Bayes factor greater than 0); the inversion is easily handled. See the text in §2.4:2 for important details regarding the nature and function of the information given in this table.

model? Let us turn our attention to equation (2.12), wherein $\mathcal{Z}$ is defined. We can see that the Bayesian evidence is the integral of the prior-weighted likelihood over parameter space. This means that the Bayesian evidence is nothing more, and nothing less, than the average likelihood of the model given the data, weighted by the prior. (Reference the discussion earlier about the marginal likelihood and the expectancy of the data.) Models with high Bayesian evidence are simply those with high average likelihood – in the view of such models, across a broad range of parameters, the data are not particularly surprising.

The intuition here is that overly simple models that fail to capture the data well will have low likelihood everywhere; since every value of the integrand will be low, so too will be the total integral – namely, the evidence. Complex but over-fitted models will have small regions of extremely high likelihood where the parameters have been sufficiently fine-tuned; however, being over-fitted, such models will have low likelihood at most points in parameter space, and so again, the overall evidence will be low. (The unnecessary parameters causing a very rapid drop in likelihood as the fine-tuned values are changed.) A simpler theory (*viz.* one with a more compact parameter space) will be granted grater evidence than a more complicated one, unless the complications significantly improve the model's ability to explain the data. The only way for a model to achieve high evidence is to have relatively high likelihood over an appreciable portion of its parameter space, which means all the parameters do useful work towards explaining the data.

Let us remark that the Bayesian evidence has no interpretation as an absolute quantity, except as a normalising factor for the posterior. We cannot speak, in this formalism, of how well the data supports a model *per se*; we can only speak of the extent to which the data supports one model *as compared to* another, as given by the Bayes factor.

Once a Bayes factor has been obtained, it is tempting to ask how significant the quantitative

statement of evidence it makes is, when set against our normal qualitative intuitions that evidence can weak, strong, decisive, and so on. Table 2.1 contains two commonly used schemes, drawn from Kass and Raftery (1995), which interpret the quantitative Bayes factor as heuristic qualitative interpretations of the weight of evidence afforded to one model over another in terms of the Bayes factor. The first table is a slight modification of a proposal by Jeffreys (1961), who suggested interpreting the Bayes factor in half-units on the base-10 log scale. The second interpretation (and it seems the more commonly used), is Kass and Raftery's own; it raises the threshold for strong evidence somewhat, and recasts the logarithmic Bayes factor as twice the natural logarithm. This transformation puts the Bayes factor on the so-called *deviance scale*, comparable to the familiar scales used in, for example, the Akaike information criterion; the deviance-scaled Bayes factor will be our principal means of representing the Bayes factor in this work.

Note that, in terms of Bayesian inference proper, the Bayes factor is the ultimate answer to the question of model comparison. The schemes given in table 2.1 are not calibrations of the Bayes factor to the scale provided by more intuitive notions of standards of evidence; they are calibrations of standards of evidence to the scale provided by the Bayes factor. In other words, the judgement of the significance of a given Bayes factor means stepping beyond the formalisms of Bayesian inference, and the judgement will depend critically on the context in which it is being made. The descriptive terms given in any such scheme ("substantial", "decisive", and so on) are therefore by nature approximate, and derived from the practice and the needs of the enterprise in which the scheme is employed. The schemes cited here formalise useful, normal standards of evidence for the purposes of scientific inquiry; the scheme that would decide what Bayes factor constitutes decisive evidence in a criminal trial (the legal standard of "beyond reasonable doubt") might look very different indeed.

Finally, we mention the issue of prior sensitivity. The Bayes factor is critically sensitive to the choice of prior, since – *contra* to the more common understanding of a model being defined by its likelihood function alone – the prior over the model parameters forms an integral part of the notion of a model in Bayesian statistics. While posteriors tend to converge quite quickly under a range of priors, the Bayesian evidence calculated in each case can differ quite markedly, resulting in divergent inferences on the basis of the Bayes factor. When the prior is vague or debatable, it is therefore common to perform prior sensitivity analyses, that is, to find the Bayesian evidence for a proposed likelihood function under numerous reasonable priors to establish the strength of the model selection.

Aside from traditional prior sensitivity analysis, in this thesis, the author introduces and uses another method of insuring against prior sensitivity that can work when certain relationships obtain between the maximum likelihood and Bayesian evidence; this method is covered in appendix D.

—ω—

CHAPTER 3

Nested sampling and MultiNest

Synopsis. The exposition of Bayesian inference (BI) in the previous chapter, while containing a profound set of theoretical principles for inference, is practically speaking worthless if it cannot be used to actually answer any questions of interest. The process of applying these principles to invent methods of actually performing inductive inference turns out to be difficult. Fortunately, research in the methods of BI have made significant advances along this front. We focus our attention here on MultiNest, the BI algorithm which will be used throughout this thesis. MultiNest is a sophisticated form of an approach to BI known as *nested sampling*, and was invented in the context of basic physics – particle physics, cosmology, and astronomy – fields in which the use of Bayesian methods is commonplace. In those fields, MultiNest has provided many-fold increases in the efficiency of the two key problems of BI, *model comparison* and *parameter estimation*, compared to competing techniques.

This chapter will provide a comprehensive exegesis of the MultiNest method from the basic ideas of nested sampling up to the particulars of the algorithm itself. Through this development, those familiar with other techniques of BI will be able to see the differences between them and make their own judgements; the unfamiliar will be able to gain an understanding of why MultiNest works, what its comparative (dis)advantages are in broad strokes, and how it will be useful for our later purposes.

1. THE PRACTICAL CHALLENGES OF BAYESIAN MODEL ANALYSIS

For the theory of Bayesian inference to be usefully applied, practical techniques for calculating posteriors and Bayesian evidences are needed. We will refer to the undertaking of these two tasks collectively as *Bayesian model analysis*. Bayesian evidence, in particular, is *prima facie* very difficult to calculate. In all but the simplest of cases, the evidence integral (2.12) is analytically intractable, necessitating the use of numerical methods to perform the computation. Numerical methods must surmount a number of difficulties if they are to effectively perform Bayesian model analysis. First, when evaluated directly, the evidence integral is cursed to become exponentially more computationally expensive as the number of model dimensions (*viz.* free parameters) increases. Furthermore, it is not uncommon for a Bayesian analysis to encounter difficulties because of complexities, originating in the form of the likelihood function, in the form of posterior

distribution itself. Posteriors may be *multimodal* – they may exhibit multiple peaks in parameter space – or may have *degeneracies* between model parameters, forming "ridges" in parameter space where the posterior value is the same or similar, which are particularly troublesome when curved. (Figure 3.3 showcases examples of these difficult posterior distributions.) If present, these features can introduce considerable inefficiencies into the calculation of the posterior and evidence; at worst, they can render methods incapable of producing anything other than outright poor estimates, or indeed any estimate at all.

Fortunately, theoretical advances in the algorithms for Bayesian model analysis have progressed far enough for the above problems to be circumvented within the bounds of modern computational capability. For the studies presented in later chapters, we use MultiNest, an algorithm for performing Bayesian model analysis developed for use in cosmology, astrophysics, and particle physics; we will later show how its use can be fruitfully extended into the cognitive sciences. MultiNest provides an efficient means of calculating the Bayesian evidence of a model directly, while making no assumptions regarding, so not being limited by, the form of the likelihood function of the model. It also results in accurate and informative posterior distributions, and is capable of effectively addressing the problems of multimodal and degenerate posterior distributions for models which have low-to-moderate numbers of parameters (Feroz et al., 2009; Feroz et al., 2013; Feroz & Hobson, 2008).

MultiNest is an elaboration on the basic technique of nested sampling, introduced by Skilling (Skilling, 2004,2006). Nested sampling is a Monte Carlo technique aimed primarily at efficiently calculating Bayesian evidences; posterior inferences are produced as a direct by-product of the evidence calculation. At the heart of the technique lies the reduction of the evidence integral into a much simpler one-dimensional form, consideration of which gives rise to the eponymous sampling method. This reduction, in essence, lets nested sampling explore those parts of the posterior distribution where the posterior is highest – and therefore which make the most meaningful contributions to the Bayesian evidence.

This guided process of random sampling is the central idea which reduces the computational difficulty of evidence evaluation in MultiNest. The crux of MultiNest is, through this sampling, to build, rapidly and accurately, a picture of the prior-weighted likelihood function $\mathcal{L}(\theta)\,\pi(\theta)$, the determinant object of both the posterior and the Bayesian evidence; furthermore, using the reduced form of the evidence integral, the collected samples can be used to approximate and evaluate the integral in an efficient manner. In slightly more detail, MultiNest works by sampling points across parameter space, weighted the prior, and calculating the value of the likelihood function at those points. It then finds, nested within the region covered by those points, a smaller region (or clusters of smaller regions) of parameter space with higher associated likelihood values (the greatest contribution to the Bayesian evidence coming from, and the best-fit parameters being found in, regions of parameter space with high likelihood). This process is iterated as

many times as necessary for the completion of an accurate picture of the whole prior-weighted likelihood function. Once this has been found, it is relatively straightforward to find the best-fit parameters, by taking the maximum *a posteriori* parameters, and the Bayesian evidence, by integrating over the function in the one-dimensional evidence integral, the form of which function can be approximated directly from the samples collected.

The remainder of this section will be a technical overview of how MULTINEST works. We start with an exposition of the basic ideas of nested sampling, before proceeding step-wise through various developments on the basic technique which culminate in the MULTINEST algorithm itself. In line with this chapter's aim to elucidate the theoretical basis of and justification for the approach applied in later chapters, the focus will be on building a strong conceptual understanding of how and why the chosen technique works.

For reasons of length, and since they are not directly relevant to the argument that follows, the fine technical details will be omitted; the interested reader is referred to the original papers for further information, especially Skilling (2004) and Feroz et al. (2009). Also note that there will be no extensive comparison made here between MULTINEST and other BI techniques, such as thermodynamic integration, since such a comparison lies well outside the scope of this work and would represent a massive undertaking in and of itself; practical comparisons of MULTINEST to other techniques will be the main subject of Chapter 5. For now, the focus is squarely on applying the principles of the previous chapter and demystifying the black box of later ones.

2. NESTED SAMPLING

2:1 Transforming the evidence integral: Λ

Recall from the previous section that Bayesian evidence for a statistical model M given by data D is found by the evidence integral

$$\begin{aligned} \mathcal{Z} \equiv p(D \mid M) &= \int_{\theta \in \Theta} \mathcal{L}(\theta)\, \pi(\theta)\ \mathrm{d}^n\theta \\ &= \int_{\theta \in \Theta} p(D \mid \theta, M)\ p(\theta \mid M)\ \mathrm{d}^n\theta \end{aligned} \tag{3.1}$$

where Θ is the set of n free parameters for the model, with n also being referred to as the *dimensionality* of the model. In all but the most elementary of cases, the evidence integral is analytically intractable, demanding the use of numerical techniques for its evaluation. However, it is also clear that with standard techniques of numerical integration, directly calculating the integral in (3.1) quickly becomes practically infeasible with increasing model dimensionality: since these techniques rely on dividing the domain of integration (the parameter space) into an n-dimensional mesh, and since the number of computations per evaluation is proportional to the

number of mesh elements, the number of calculations required to achieve a particular level of accuracy increases exponentially with the number of dimensions. Nested sampling is rooted in a recasting of the integral in (3.1) into a much simpler one-dimensional form. To arrive at this result, we start by reviewing the standard numerical procedure for calculating the value of the integral, whence the simplified version will suggest itself naturally.

A brief note before we proceed: the explanation will be illustrated by a straightforward two-dimensional example. Accordingly, in the description that follows, we shall keep in mind the following geometrical picture: the parameter space as a rectangular two-dimensional "field" in the plane of the x and y dimensions, with the prior, likelihood, and prior-weighted likelihood functions pictured as arbitrary, curving surfaces over the parameter field in the z dimension (as illustrated in figure 3.1). The evidence is then conceived as the volume under the prior-weighted likelihood surface down to the parameter field.

Obviously, this picture is not general; nevertheless, as discussed in 2.3:3, results in low-dimensional spaces extend naturally to higher-dimensional ones and facilitate a firmer grasp on the underlying concepts. In what follows, we will sometimes use the language of a two-dimensional embedded parameter space in a three-dimensional space containing the prior and likelihood surfaces, and use that picture to help ground our understanding, although it should be understood that the general case is meant.

Returning to our main theme – in a slightly simplified form, the standard procedure for numerical integration entails the following (illustrated in figure 3.1(a)):

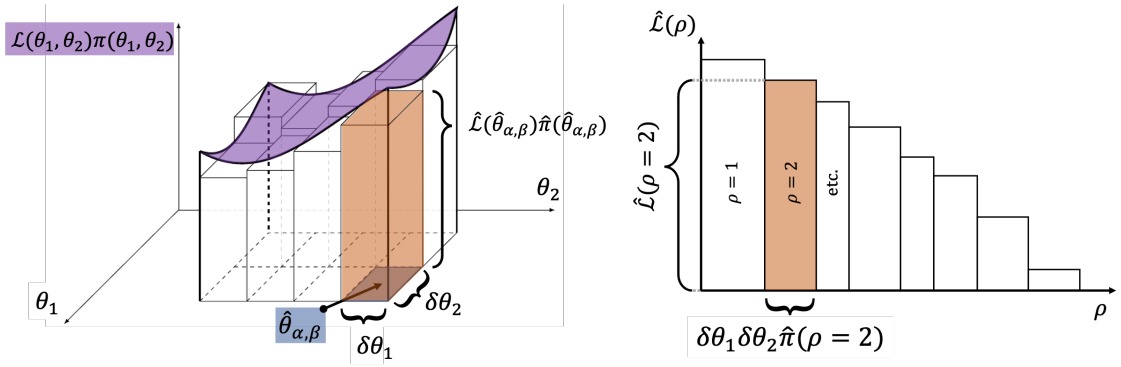
1. The parameter space is divided into a discrete mesh or grid by partitioning each parameter θ_i into small segments of width $\delta\theta_i$, and thereby the whole space into rectangular elements of side length $\delta\theta_1, \delta\theta_2, \cdots, \delta\theta_n$. The k^{th} segment of the i^{th} parameter encompasses the parameter values $(k-1)\,\delta\theta_i \leqslant \theta_i < k\delta\theta_i$ and will be denoted by $\hat{\theta}_i^k$, so that each element of the mesh is identified by coordinates $\hat{\theta} = \hat{\theta}(k_1, k_2, \cdots, k_n) = \left(\hat{\theta}_1^{k_1}, \hat{\theta}_2^{k_2}, \cdots, \hat{\theta}_n^{k_n}\right)$.

2. For each element of the mesh, a likelihood and prior and value are assigned:

$$\begin{aligned} \hat{\mathcal{L}}\left(\hat{\theta}\right) &= \hat{\mathcal{L}}\left(\hat{\theta}(k_1, k_2, \cdots, k_n)\right) \\ \hat{\pi}\left(\hat{\theta}\right) &= \hat{\pi}\left(\hat{\theta}(k_1, k_2, \cdots, k_n)\right) \end{aligned} \tag{3.2}$$

   These amount to discretised versions of those functions, i.e., they give an estimation of the value of the functions in those regions of parameter space covered by each element in the mesh.

3. The Bayesian evidence is approximated as the sum of the estimated contribution from every



(a) Standard mesh method **(b)** Likelihood-ranked

Figure 3.1. TWO INTEGRATION SCHEMES FOR ESTIMATING THE BAYESIAN EVIDENCE. (a) The (θ_1, θ_2) plane is the parameter space; the prior-weighted likelihood function (PWLF) $\mathcal{L}(\theta_1, \theta_2)\,\pi(\theta_1, \theta_2)$ is shown inside the limits of integration in the purple-shaded surface above the parameter plane. The Bayesian evidence is, by definition, the volume contained beneath the prior-weighted likelihood function inside the limits of integration. Standard mesh integration starts by dividing the parameter plane into small rectangles with side lengths $(\delta\theta_1, \delta\theta_2)$; a representative unit is highlighted in blue and labelled $\hat{\theta}_{\alpha,\beta}$; each unit is then assigned a a height, as a representative approximate value of the PWLF $\hat{\mathcal{L}}(\hat{\theta}_{\alpha,\beta}(\hat{\pi}(\hat{\theta}_{\alpha,\beta})$in the region it covers. The contribution to the Bayesian evidence from the region covered by a unit is then estimated as the volume of the resulting cuboid, shaded in red. The evidence is then calculated as the sum of the volumes of all such cuboids, summed in the order of of their appearance in the parameter plane. The true evidence value of the evidence is approached as the parameter plane is divided into ever-smaller units.(b) The three-dimensional units of (a) are ranked by the value of their associated likelihood $\hat{\mathcal{L}}(\hat{\theta}_{\alpha,\beta})$ and rearranged as a series of rectangles whose areas are equal to the volumes of the elements in the mesh integration approach. Clearly, the area under this curve is equal the volume under the PWLF, i.e., is equal to the Bayesian evidence.

mesh element:

$$\mathcal{Z} \approx \sum_{k_n} \cdots \sum_{k_2} \sum_{k_1} \hat{\mathcal{L}}\left(\hat{\theta}(k_1, k_2, \cdots, k_n)\right) \hat{\pi}\left(\hat{\theta}(k_1, k_2, \cdots, k_n)\right) \delta^n\theta \tag{3.3}$$

where $\delta^n\theta = \delta\theta_1\delta\theta_2\cdots\delta\theta_n$. That is, the evidence is seen as the sum of numerous small volumes, each consisting of a base "area" in parameter space and a height given by the prior mulitplied by the likelihood at the parameters where the volume is located.

Evidently, as the mesh becomes finer, the approximation becomes better; the true value of the evidence (the evidence integral) is obtained in the limit of the sum (3.3) as the various $\delta\theta_i$ tend to zero, that is, the mesh spacing becoming infinitesimally fine.

Let us to take a closer look at the expression in (3.3). Conceptually, the total value of the Bayesian evidence is being built up from contributions from the small elements $\mathcal{L}(\theta)\,\pi(\theta)\,\delta\theta_1\cdots\delta\theta_n$; the order in which the elements are being added here is determined by the ordering of the parameter values, but it is clear that we could perform this summation in any order and still arrive at the same result. Imagine, then, that the elements are instead summed in descending order of their

likelihoods. We label each element in the discretised parameter space by ρ, its rank by likelihood value, and then sum over that rank:

$$\mathcal{Z} \approx \sum_{\rho} \mathcal{L}(\rho)\, \pi(\rho)\, \delta^n\theta \tag{3.4}$$

This alternative is shown in figure 3.1(b).

What do we gain from this sorting? In (3.3), the geometric interpretation of the value of an element is its "volume" in an $(n+1)$-dimensional space, formed as the "area" of a discrete unit of parameter space $\delta\theta_i$ multiplied by a "height" in the remaining dimension determined by the product of the likelihood and the prior at that point. Subject to our sorting, a different geometric interpretation comes to the fore: the contribution to $\mathcal{Z}$ of an element is the area of a strip in two dimensions, with the height being given by the likelihood and the width being being given by the quantity $\delta\mu = \pi(\rho)\,\delta^n\theta$. But this quantity is just the prior mass contained in the small volume of parameter space pointed at by ρ, for it is the prior density at the point in space multiplied by the volume of parameter space that it occupies. Since we consider each element in the integration mesh once, and only once, the sum of all the $\delta\mu$ must be equal to the total prior mass, i.e., unity. Hence, we can view each $\delta\mu$ as a unit of *accumulated* prior mass.

From this, we see that by "unfolding" parameter space and the prior density in this way, there must exist a function relating likelihood to prior mass directly, $\Lambda = \Lambda(\mu)$. Therefore, rather than accumulating evidence by integrating contributions from elements of parameter space, we can integrate this function to obtain the evidence over the prior mass directly:

$$\mathcal{Z} = \int_{\theta\in\Theta} \mathcal{L}(\theta)\, \pi(\theta)\ \mathrm{d}^n\theta \longleftrightarrow \int_0^1 \Lambda(\mu)\ \mathrm{d}\mu \tag{3.5}$$

By consideration of the construction given above, it should be clear that the domain of the function is the interval $(0,1)$, since these are the values of prior mass the can be accumulated, and that the image of the function must be the interval $(0, \mathcal{L}_{\max})$, where $\mathcal{L}_{\max}$ is the largest value of the likelihood function. Moreover, Λ must be a monotonically decreasing function, i.e., $\Lambda(\mu)$ smoothly decreases as μ increases, with $\Lambda(0) = \mathcal{L}_{\max}$ and $\Lambda(1) = 0$.

If this function could be found or estimated, the evidence could in that way be evaluated by simple methods of numerical integration *no matter the dimensionality of the model*, thus bypassing one of the major practical problems of Bayesian model analysis. More specifically, it can be seen that, if one were able to evaluate the likelihoods $\Lambda_i = \Lambda(\mu_i)$, where the μ_i form a sequence of decreasing values

$$0 < \mu_M < \cdots < \mu_2 < \mu_1 < \mu_0 = 1 \tag{3.6}$$

the evidence could be numerically approximated using standard quadrature methods as

$$\mathcal{Z} \approx \hat{\mathcal{Z}} = \sum_{i=1}^{M} w_i \Lambda_i, \tag{3.7}$$

where w_i is a weight. We use the simple trapezium rule, so that $w_i = \frac{1}{2}(\mu_{i-1} - \mu_{i+1})$.

The reader may, at this point, be tempted to ask how exactly this reduction helps us computationally. After all, by rearranging the volumetric elements of (3.3) into the areal elements of (3.4), we do not obviously reduce the number of computations required: each element is still the product of a likelihood and unit of prior mass, and in the construction used, these still have to be computed for each element of parameter space for the function $\Lambda(\mu)$ to be approximated. For this reduction to be useful at all, then, it is still necessary to find a way to estimate the form of $\Lambda(\mu)$ *without* having to evaluate $\mathcal{L}(\theta)\pi(\theta)$ at every point in parameter space.

This is what nested sampling does. Before setting out the algorithm by which the approximation is performed, we will clarify the relationship between $\Lambda(\mu)$ (the likelihood over the prior mass) and $\mathcal{L}(\theta)$ (the likelihood over the model parameters); this will bestow Λ with a clear geometrical interpretation in parameter space that points the way to a more efficient approximation, which in turn will motivate the numerical algorithm.

2:2 Interpreting Λ in parameter space

Start by defining an infinitesimal unit of prior probability mass as

$$d\mu = \pi(\theta)\, d^n\theta. \tag{3.8}$$

Intuitively, this is the probability mass under the prior distribution that the true[i] model parameters are contained inside the small volume $d^n\theta$ of parameter space around the point θ, which has probability density $\pi(\theta)$. Following that, let us also define the following function:

$$\mu(\lambda) = \int_{\Theta_\lambda} \pi(\Theta)\, d^n\Theta \qquad \text{where } \Theta_\lambda = \{\Theta \;:\; \mathcal{L}(\Theta) > \lambda\} \tag{3.9}$$

The set $\Theta_\lambda = \{\Theta \;:\; \mathcal{L}(\Theta) > \lambda\}$ selects a region of parameter space, specifically the one composed of all points in parameter space which, when evaluated against the data, return a likelihood value λ or greater. On the assumption that the likelihood function varies smoothly with the parameters, these regions are bounded by contour lines on the likelihood surface of equal likelihood λ; we will call these the *isolikelihood contours*. (See figure 3.2.) Conceptually, therefore, $\mu(\lambda)$ is the prior probability mass contained within the regions within parameter space where the likelihood is greater than λ. Moreover, each λ picks out a particular isolikelihood contour, and small variations

[i] Recall the discussion earlier: "true", *assuming that the proposed model does, in fact, hold in reality.*

in λ therefore produce nested shells of isolikelihood contours in parameter space. Thus, we have a way of associating accumulated prior mass, unique regions of parameter space, and likelihood values.

From this definition, a few things immediately follow. First, since λ represents a likelihood value, it must be the case that λ lies between zero (zero likelihood being assigned to parameter values which cannot reproduce the observed data, and negative likelihoods being impossible) and the maximum likelihood that our model can obtain on our data; that is, $0 \leqslant \lambda \leqslant \mathcal{L}_{\max}$. Second, $\mu(0) = 1$, since all the prior probability mass must be contained within the region where the likelihood is greater than zero, that is, across the whole of parameter space. Third, as λ approaches $\mathcal{L}_{\max}$, the probability mass contained within the isolikelihood contours defined by λ must approach zero, since the region contained within the iso-likelihood contours shrinks to contain only the maximum likelihood point(s); stated mathematically, $\lim_{\lambda \to \mathcal{L}_{\max}} \mu(\lambda) = 0$. Finally, μ must be a monotonically decreasing function of λ: for $\lambda_1 > \lambda_2$, the probability mass contained within the iso-likelihood contour set by λ_1 must be less than the mass contained within the contour set by λ_2, since the region of parameter space contained in Θ_{λ_1} must be a subset of that contained in Θ_{λ_2}.

It is now clear that we can define Λ as the inverse of μ:

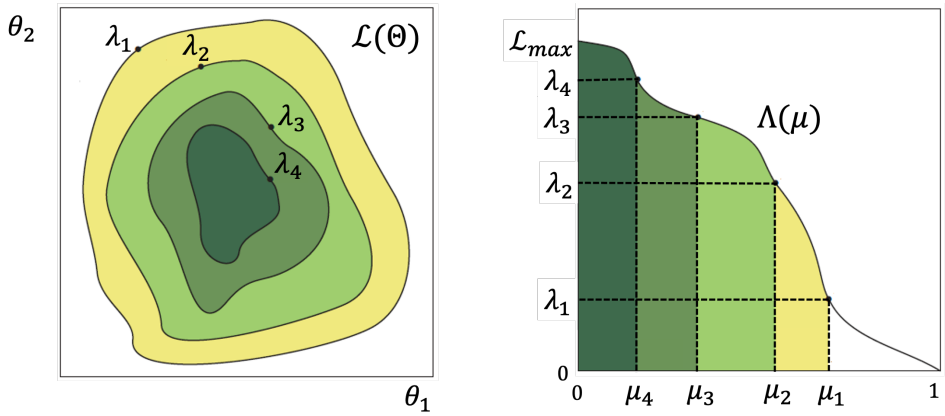
$$\Lambda(\mu) \;:\; \Lambda(\mu(\lambda)) = \lambda \tag{3.10}$$

That is, $\Lambda(\mu)$ returns the likelihood value λ whose isolikelihood contour in parameter space contains μ amount of prior mass. As expected, Λ must be a monotonically decreasing function of μ – the more mass that is contained within an isolikelihood contour, the larger the region of parameter space it needs to cover, and therefore the lower the likelihood value that defines its bound.

2:3 The idea of nested sampling: Direct integration over the prior mass

With this understanding of what $\Lambda(\mu)$ betokens firmly in mind, let us return to the problem of estimating the function in a way that does not depend on likelihood sorting. Recall from equations (3.5) and (3.7) that, to estimate $\mathcal{Z}$, we require a set of m samples consisting of paired points (μ_i, Λ_i) where $0 < \mu_m < \cdots < \mu_2 < \mu_1 < \mu_0 = 1$. Now, by its very definition, the regions of parameter space which dominate in their contribution to the integral for $\mathcal{Z}$ will be coextensive with wherever the bulk of the posterior is found. These regions will typically occupy only that small portion of parameter space wherein high likelihood values are concentrated.[ii] In terms of prior mass

[ii] There are two reasons for this. First, if the likelihood is to be useful, it *should* significantly narrow the likely range of parameters admitted into the prior. Second, ratios in the value of the likelihood at different points tend to be much greater than those in the prior (which itself is often flat), thereby producing more pronounced differences in the value of the prior-weighted likelihood function and so the evidence.



(a) Isolikelihood contours in parameter space

(b) The function $\Lambda(\mu)$

Figure 3.2. RELATING THE LIKELIHOOD TO PRIOR MASS DIRECTLY. (a) A contour plot of a two-dimensional likelihood function. The lines represent isolikelihood contours with likelihood λ_i. The interior of each isolikelihood contour picks out its own particular subspace of the parameter space, as indicated by the shading, with the region picked out by $\lambda' > \lambda$ being nested inside the region picked out by λ. **??**. Illustrating how the function Λ relates likelihoods to prior mass directly. The area of the strip shaded in a particular colour corresponds to the prior mass contained by integrating the prior over the same-coloured region of parameter space, as indicated in (a). This figure is a modified version of figure 1 in Feroz et al. (2013), reproduced with permission.

enclosed, then, the posterior mass will be narrowly distributed over a small range of prior mass values; by virtue of the means by which we accumulate that prior mass, this will be found in the lowest reaches of the prior mass range $(0, 1)$. Concomitantly, the posterior mass will be quite broadly distributed over $\log(\mu)$. Therefore sampling linearly in $\log(\mu)$ rather than μ, we set one half our sequence of samples as

$$\mu_0 = 1, \quad \mu_1 = r_1, \quad \mu_2 = r_1 r_2, \quad \cdots, \quad \mu_t = r_1 r_2 \ldots r_t, \quad \cdots, \quad \mu_m = r_1 r_2 \ldots r_m \tag{3.11}$$

where each r_t lies between zero and one.

The obvious sampling method would be to sample this prior mass value directly, choosing r_t randomly and uniformly from the interval $[0, 1]$ in each successive iteration to arrive at μ_t. If we had the likelihood sorting, we could then assay it to find the corresponding values θ_t and $\mathcal{L}(\theta_t)$ to obtain Λ_t. But this, of course, betrays our purpose, which is precisely to obviate the need for the likelihood sorting. However, a little consideration reveals the existence of a totally equivalent sampling procedure, one in which we sample parameter space rather than the prior mass directly. Suppose that on each iteration, we sampled points in parameter space weighted by the prior, until a point was found conforming to the constraint $\mathcal{L}(\theta_t) > \mathcal{L}(\theta_{t-1})$; now, because of the relationship established between isolikelihood contours and prior mass accumulated, this point *must* satisfy $\mu_t < \mu_{t+1}$. But this satisfies our basic condition (3.6) for the series of samples from which our function estimate is formed. Sampling in μ and θ respectively, each interval in

the sampling process represents successively lower-valued points in the range $[0, 1]$, and nested shells of isolikelihood contours, each with higher likelihood value.

So, we need not sample from the prior mass *directly* to find successively smaller values of μ from which to form out estimate of Λ. In collecting samples as θ_t, though, we lose the information of which equivalent μ_t was sampled; but if we could estimate the corresponding μ_t, we would forgo the need to any likelihood sorting at all. Fortunately, we can do exactly this. Suppose, on the t^{th} iteration of our sampling procedure, we were to take N points all satisfying the likelihood constraint and select the "worst" point (the one with lowest $\mathcal{L}$ and highest μ) as our t^{th} sample. Knowing that $\mu_0 = 0$ and $\mu_t = r_t \mu_{t-1}$, to estimate μ_t, we need only estimate r_t. But, because we are selecting N points independently and randomly, each r must be distributed according to $p(r) = Nr^{N-1}$ in $[0, 1]$ – that is, according the the probability distribution of the largest of N samples drawn uniformly from the interval $[0, 1]$. The mean and standard deviation of $\log(r)$ are $E[\log r] = -\frac{1}{N}$ and $\sigma(\log r) = \frac{1}{N}$; since each r_t is independent, by plugging these back in to our formula for μ_t, we see that the prior mass after t iterations is expected to shrink such that

$$\log \mu_t \approx -\frac{t \pm t}{N} \tag{3.12}$$

whence we can approximate

$$\mu_t = \exp -\frac{t}{N} \tag{3.13}$$

as was required.

We have now obtained a method for calculating the Bayesian evidence by reducing it to a numerically tractable one-dimensional integral, and sampling parameter space using nested isolikelihood contours to estimate the integral. This sampling process, by concentrating its search on those regions of parameter space with high likelihood, significantly reduces the number of likelihood calculations required to find the evidence, while still providing a detailed picture of the informative parts of the posterior, allowing for considerable gains in computational efficiency when calculating the evidence and the posterior.

2:4 The algorithm for nested sampling

From the above, we can surmise the following general algorithm for performing nested sampling:

1. Set the iteration counter to $t = 0$.
2. Draw N "active" or "live" samples $u_k \in \Theta$ from the full prior $\pi(\theta)$, so that the initial prior mass $\mu_0 = 1$. The set of active samples $\{u_k, 1 < k \leqslant N\}$ is the "active set."
3. Find the likelihoods of the active samples $\mathcal{L}(u_k)$ and remove the sample with the lowest likelihood on this iteration from the active set. Denote the likelihood of the discarded point by $\mathcal{L}_t$.

4. Replace the sample just removed from the active set with a new one, u', drawn from the prior (as in step 1), subject to the constraint that its likelihood be greater than that of the point just removed; that is, $\mathcal{L}(u') > \mathcal{L}_t$.

5. As was shown earlier, we can take $\mu_t = e^{-t/N}$. Having found μ_t and $\mathcal{L}_t$, the contribution to the total evidence of this iteration can be readily calculated by (3.7).

6. Increment t and repeat steps 3-6 until the evidence is determined to a specified precision.

The precision of the evidence value (which governs the termination of the algorithm) is set as follows: At iteration t, the maximum possible remaining evidence is estimated by assuming that all the points inside the isolikelihood contour at t have likelihood equal to that of the most likely point found so far. That is, we estimate the remaining evidence as $\Delta\mathcal{Z}_t = \mu_t\mathcal{L}_{\max}$, where $\mathcal{L}_{\max}$ is the maximum likelihood from the current set of active samples. The algorithm is terminated when $\ln(\Delta\mathcal{Z}_t)$ reaches some specified value, usually 0.5.

As the above makes clear, the primary sampling parameters which govern the execution of the nested sampling algorithm are the number of live or active points N and the tolerance $\ln(\Delta\mathcal{Z}_t)$ which specifies the stopping criterion.

2:5 Final evidence estimate, errors, and posterior inferences

Once the the algorithm is terminated, the evidence estimate as given in equation (3.7) is refined by adding a final increment calculated from the terminal set of active points. This increment is given by

$$\Delta\hat{\mathcal{Z}} = \sum_{k=1}^{N} b_{M+k}\,\Lambda_k \tag{3.14}$$

with $b_{M+k} = \mu_M/N$ for all k.

The dominant source of error in the evidence estimate obtained through nested sampling is due to the fact that each prior mass value μ_t is a random variable, per the development in equations (3.11)-(3.13); this prior mass assignment error far outweighs the discretisation error accrued during the numerical integration (3.7) done to obtain the evidence (Feroz & Hobson, 2008). Fortunately, the prior mass assignment error can be easily estimated as the relative entropy of the full sequence of samples, the formula for which will be given below.

The author notes here that, since prior mass assignment is the only source of random uncertainty in the evidence calculation, the prior mass assignment error approximates the sampling error (that is, the standard deviation in the evidence estimates that would be obtained through multiple runs), thus obviating the need for multiple inference runs to establish the uncertainty in the evidence estimate.

Taking the prior mass assignment error into account and combining it with the results of

(3.7) and (3.14), the final estimate of the log-evidence is given by

$$\ln \hat{\mathcal{Z}} \;=\; \ln\left(\sum_{i=1}^{M} w_i \Lambda_i \;+\; \sum_{k=1}^{N} b_{M+k} \Lambda_k\right) \;\pm\; \sqrt{\frac{H}{N}} \tag{3.15}$$

where H is the *negative relative entropy*, defined as

$$H \;=\; \int \ln\left(\frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\mu}\right) \mathrm{d}\mu \;\approx\; \sum_{i=1}^{M} \frac{w_i \Lambda_i}{\hat{\mathcal{Z}}} \ln\left(\frac{\Lambda_i}{\hat{\mathcal{Z}}}\right). \tag{3.16}$$

Posterior inferences can be directly generated from the full sequence of samples collected. A little thought reveals that each sampled point represents a point in the posterior distribution, given by

$$\mathcal{P}_j = \frac{w_j \Lambda_j}{\hat{\mathcal{Z}}} \tag{3.17}$$

where the index j, running from 1 to $M + N$, covers all the samples collected.

3. THE MULTINEST ALGORITHM

3:1 The limitations of vanilla nested sampling

While the essential idea of nested sampling is quite ingenious, the "vanilla" version of the method presented above suffers from severe deficiencies. Most pressing are two closely related problems: those of maintaining sampling efficiency and of dealing with difficult posterior geometries.

To understand the first of these, we note that by far the most computationally demanding step in nested sampling is to draw a new sample from the prior that is acceptable under the hard constraint $\Lambda' > \Lambda_t$. Consider carrying out the naïve procedure of drawing from the full prior at each iteration: this results in an ever-decreasing acceptance rate for new samples, as the ever-increasing likelihood constraint (and ever-shrinking isolikelihood contour) limits the satisfactory regions in the prior. Given the extent of the bottleneck this causes, especially in later stages of sampling, further developments of nested sampling have focused on sidestepping this pitfall.

How is this sidestep achieved? Well, it is clear that if one knew – at least roughly – the isolikelihood contour associated with a given discarded point, one could limit new samples to be drawn from within those bounds, thereby focusing the search for acceptable samples in those regions of parameter space where such those samples are most likely to be found. The driving force in the evolution of the techniques of nested sampling, then, has been *to find better methods of approximating the bounds of the isolikelihood contour at each iteration.* So long as the bound well approximates the true isolikelihood contour, the acceptance rate will remain high. Conversely,

a poor approximation will reintroduce the exact same problem of sampling inefficiency, since the utilised bound will be mostly comprised of regions of parameter space with unsatisfactory likelihood values.

3:2 Intermediate developments: Ellipsoidal and recursive clustered nested sampling

The most straightforward development in this vein is simple *ellipsoidal* nested sampling Mukherjee et al. (2006). In this method, the isolikelihood contour at each step is approximated by an n-dimensional ellipsoid; the exact form of the ellipsoid is fixed by the covariance matrix of the set of active samples and enlargement factor f which accounts for the possibility that acceptable likelihood points lie somewhat outside the region. Evidently, such an approach is satisfactory when the isolikelihood contour at each step does, in fact, conform relatively closely to an ellipsoid; this is the case for simple unimodal posterior distributions, but the approach is not well adapted to posterior distributions which deviate significantly from this morphology. Such posteriors, however, are common in practice. To deal with them, methods are needed which relax the assumptions of this simple approach with regard to the shape of the posterior (and by extension the isolikelihood contours).

Practice shows that there are two flavours of posterior distribution which call for further modification to the contour estimate, *multimodal distributions* and those exhibiting pronounced curving *degeneracies*. These flavours can be mixed. Figure 3.3 illustrates these kinds of difficult posterior.

Multimodal posterior distributions have multiple, distinct regions of parameter space have high posterior values (each such region being a "mode"). Distributions of this kind can be dealt with by a straightforward extension of simple ellipsoidal nested sampling known as *recursive clustering*: by identifying discernible clusters of well-separated active points as the sampling process continues, each cluster can be interpreted as belonging to a different mode, and once so identified, be treated independently. Shaw et al. (2007) introduced and implemented recursive clustering by identifying distinct cluster of live points using K-means clustering. (See figure 3.4.)

Curving degeneracies are trickier to deal with. A "degeneracy" is an elongated region of parameter space which has uniformly high values, representing a kind of correlation between parameters such that, by changing the parameters within the degenerate region, the likelihood remains the same. Thus, a curving degeneracy means that the form of the isolikelihood contour resembles not an ellipse, or multiple ellipses, but a (multidimensional) banana, sigmoid, toroid, or some other complicated form. Such distributions will never be well approximated by a simple ellipsoidal bound, and therefore demand more sophisticated methods to set their bounds.

Taken together, the ability to deal with multimodality and curving degeneracies amounts to the ability to deal with more or less arbitrary posterior functions. This is the demand met by MultiNest.

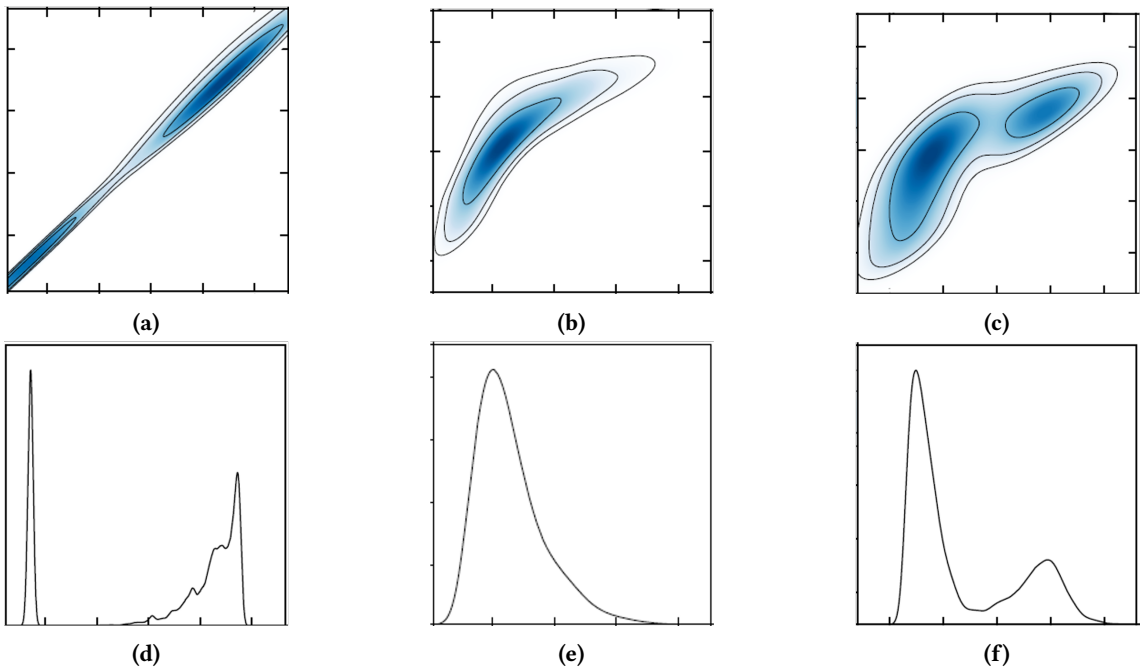


Figure 3.3. EXAMPLES OF DIFFICULT POSTERIORS. Drawn from cognitive models applied to real reaction time data, from preliminary work done by the author not reported here. The figures in the top row show contour plots of the posterior distribution marginalised across two model parameters, for a variety of models and data; the bottom row shows the corresponding one-dimensional marginal posteriors for the parameter along the abscissa in the 2d plots. (a), (d) A model which exhibits both strong multimodality and strong, though non-curving, degeneracy. (b), (e) A model with a moderate curving degeneracy. (c), (f) A model exhibiting multimodality with mild features of curving degeneracy.

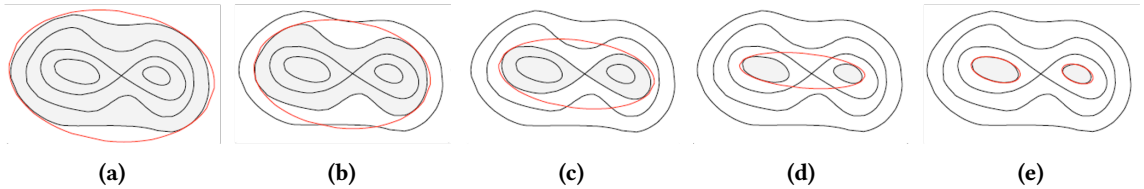
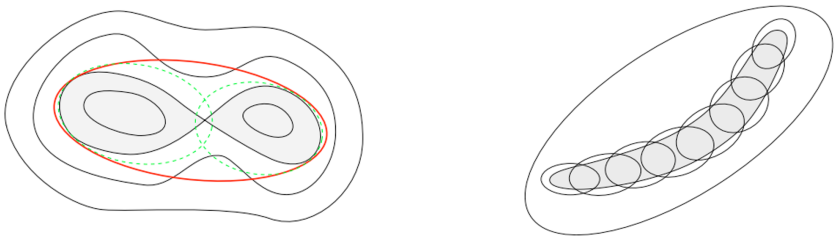


Figure 3.4. INCREASING SAMPLING EFFICIENCY FOR MULTIMODAL POSTERIORS. A sketch showing how allowing an ellipsoidal bound to split into two increases the efficiency of nested sampling, for a two-dimensional example. The black lines indicate isolikelihood contours in the likelihood function. In (a), the bound (in red) from which new samples are drawn largely covers regions of parameter space which have higher likelihood; this leads to efficient sampling. As the algorithm nests inwards in steps (b)-(d), the acceptance rate rapidly falls as the bound becomes stuck around two small modes in parameter space. Allowing the bound to split and sample independently from each mode, per (e), restores the sampling efficiency. This figure is a modified version of figure 2 in Feroz et al. (2009), reproduced with permission.



(a) For multimodal distributions **(b)** For curving degeneracies

Figure 3.5. OVERLAPPING ELLIPSOIDS ALLOW FOR BETTER APPROXIMATIONS OF THE ISOLIKELIHOOD CONTOUR. Sketches to illustrate the following: (a) If we are limited to non-overlapping ellipsoids (red line, cf. figure 3.4(c)), and if two modes overlap, the ellipsoidal bound which attempts to approximate the isolikelihood contour will be poor, leading to sampling inefficiency when new live points are drawn. A bound drawn from two overlapping ellipsoids (green line) can provide a much better approximation. (b) By extension, *multiple* overlapping ellipsoids could be used to provide bounds to approximate isolikelihood contours with arbitrary forms, as with this curving degeneracy.This figure reproduces figures 4 and 5 in Feroz et al. (2013), with permission.

3:3 MULTINEST'S REFINEMENTS TO NESTED SAMPLING

As noted, the impetus for the MULTINEST algorithm is the problem of how to construct approximate isolikelihood contours from a set of active points when the isolikelihood contours are not necessarily amenable to approximation by ellipsoids. The starting point for MULTINEST's conceptual development is to notice that, if the ellipsoids corresponding to different modes overlap, sampling from one ellipsoid that encloses all the live points can be quite inefficient. In such cases, multiple *overlapping* ellipsoids present a better approximation to the isolikelihood contour; this is illustrated in figure 3.5(a)). The idea of overlapping ellipsoids being used to construct bounds to approximate the isolikelihood contour not only improves sampling efficiency in the multimodal case; it also solves the problem of dealing with curving degeneracies, as shown in 3.5(b).

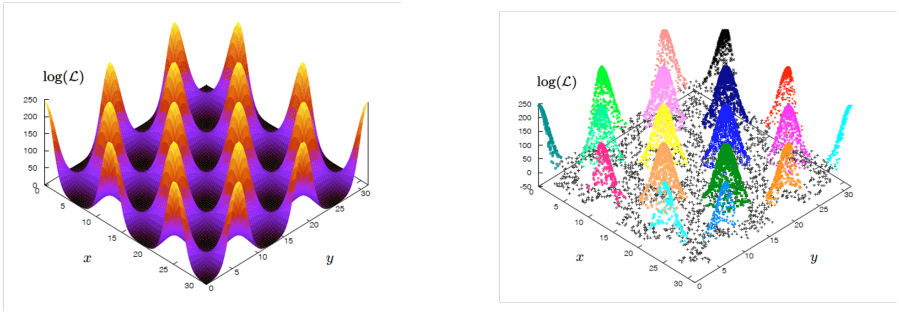
This is the basic idea exploited in Feroz and Hobson (2008) and further developed in Feroz et al. (2009), where MULTINEST proper was introduced. The basic method MULTINEST uses to implement this ideas is, at each iteration, to split the set of active points into a series of (potentially overlapping) partitions, and then to form an individual ellipsoid around each partition; the region contained inside the isolikelihood contour is approximated by the union of the individual ellipsoids, and the next replacement live point is sought inside that union. To find the optimal decomposition of the active points into partitions and thence ellipsoids, an "expectation-maximisation" approach is employed, which seeks (essentially) the minimise the sum of volumes of the ellipsoids (with an additional constraint relating to an analogue of the enlargement factor f mentioned above in reference to Mukherjee et al.'s basic ellipsoidal nested sampling).

Examples of the bounds formed by MULTINEST's decomposition algorithm are shown in figure 3.6. As can be seen, the expectation-maximisation approach to ellipsoidal decomposition algorithm grants considerable flexibility to the isolikelihood contour approximations formed



(a) With non-overlapping ellipsoids **(b)** With a torus

Figure 3.6. RESULTS OF MULTINEST'S DECOMPOSITION ALGORITHM. The ellipsoids (in red) that result when the algorithm is applied to 1000 samples (in black) drawn from: (a) Two non-overlapping ellipsoidal regions. (b) A torus.This figure reproduces figure 3 in Feroz et al. (2009), with permission.



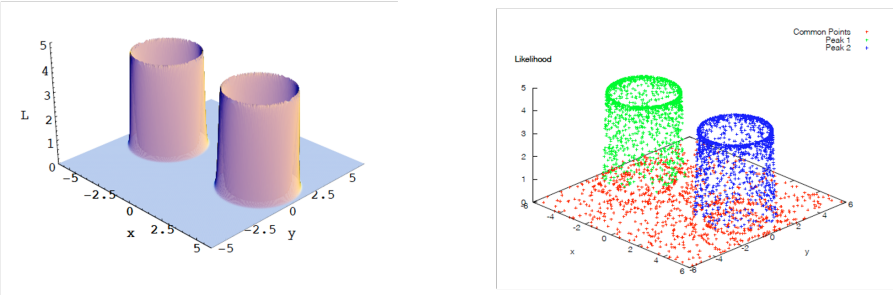
(a) True likelihood function **(b)** MULTINEST's exploration

Figure 3.7. RESULTS OF MULTINEST FOR THE EGG-BOX LIKELIHOOD FUNCTION. (a) The true form of the likelihood function. (b) Dots denoting the points with the lowest likelihood at successive iterations of the MULTINEST algorithm. Different colours denote points assigned to different modes as the algorithm progresses.This figure reproduces figure 5 in Feroz et al. (2009), with permission.

during posterior exploration. Notice that the irregularly-shaped posterior mode of figure 3.6(b) is broken into a relatively large number of small but overlapping ellipsoids, while the smooth, near-Gaussian posterior modes of figure 3.6(a) are kept whole. It is also worth remarking that the decomposition algorithm automatically enables the identification of distinct posterior modes, which can be taken as non-overlapping partitions in the decomposition; so identified, the distinct modes can then be isolated and each left to undergo its own independent nesting sampling procedure.

3:4 Evidence of MULTINEST's efficacy in application

To round off the exposition of MULTINEST, we offer a very brief review of two toy models used to validate MULTINEST, as presented in Feroz et al. (2009). The first, the "egg-box model", has a likelihood function with numerous modes in parameter space. The results of MULTINEST's exploration of the posterior, contrasted with the true posterior, are shown in figure 3.7. The second, the "Gaussian rings" model, has a likelihood function which is, in a sense, maximally difficult: it consists of two modes, each of which is composed of a very thin, highly curved



(a) True likelihood function **(b)** MULTINEST's exploration

Figure 3.8. RESULTS OF MULTINEST FOR THE 2D GAUSSIAN RINGS LIKELIHOOD FUNCTION. (a) The true form of the likelihood function. (b) Dots denoting the points with the lowest likelihood at successive iterations of the MULTINEST algorithm. Different colours denote points assigned to different modes as the algorithm progresses.This figure reproduces figure 6 in Feroz et al. (2009), with permission.

region of degeneracy. The results of MULTINEST's exploration of this posterior in two dimensions is shown in figure 3.7(b); most impressively, MULTINEST was applied to this problem with up to thirty dimensions with good results. For both toy problems, MULTINEST returned accurate log-evidence estimates with low error (less than half a unit of log-evidence). Further details, including exact comparisons of the log-evidence obtained, the formulae for the toy likelihood functions, and efficiency comparisons to earlier techniques, can be found in Feroz et al. (2009).

3:5 Other nested sampling techniques and the choice to use MULTINEST

We close this chapter with a brief discussion to justify making MULTINEST the nested sampling algorithm of choice for the remainder of this thesis. First is its overwhelming success in the fields of astrophysics and cosmology, where it was initially developed, as evidenced by the nearing 3000 citations Feroz et al. (2009) has received and by the many successes attested to in application Trotta et al. (2013). Second, the "next generation" of nested sampling after MULTINEST, the POLYCHORD algorithm, is specifically designed to deal with very high-dimensional models (Handley et al., 2015). The scaling of computational expense with number of model dimensions d favours MULTINEST at d and POLYCHORD at higher d (> 100 or so in examples involving Gaussian likelihood functions). The modest number of dimensions in the models used here – not exceeding seven – counsels for the use of MULTINEST. Comparisons of MULTINEST with non-nested sampling techniques are made in Chapter 5.

—ω—

CHAPTER 4

The simple reaction time task

Synopsis. This chapter turns its attention to the cognitive process at the heart of this work, simple responding. Reaction time (RT) tasks, wherein subjects must respond as quickly as possible to stimuli while abiding by certain rules, are crucial to cognitive science, since RT is taken to reveal important information about the cognitive processes engaged by the task. Amongst RT tasks, the simplest is probably the (aptly named) *simple reaction time* (SRT) task, which requires only that subjects respond as quickly as possible to the appearance of an unchanging stimulus after a short stimulus-free "foreperiod". The SRT derives its importance both from its wide analogy to practical situations, and its fundamental place as a prototype to all other RT tasks. Further, despite its simplicity, data obtained from the SRT exhibits a number of important and complex features which remain poorly understood. The purpose of this chapter is to review these phenomena and the hypotheses that have been espoused to account for them, and subsequently to highlight current issues in our understanding of the process of simple responding.

In particular, this chapter covers two major phenomenological aspects of the SRT, the response time distribution and so-called *foreperiod effects*. In regard of the first of these, the chapter outlines the main features of the SRT's RT distribution in both its PDF and hazard function representations, as well as numerous existing models of RT distribution. In regard of foreperiod effects, the chapter covers two important and related effects: the *foreperiod distribution effect*, whereby mean RT is modulated by the duration and overall distribution of foreperiods in the task, and the *asymmetric sequential foreperiod effect*, whereby trial-to-trial changes in FP duration influence mean RT. The chapter closes by considering two theoretical syntheses that are required to advance our knowledge of the processes of simple responding, and thereafter defines the issues that will be tackled in later chapters.

1. WHY STUDY SIMPLE REACTION TIME?

Let us open this chapter with two broad and uncontroversial statements. First, if a science can be characterised by the form of data which it commonly utilises, then much of cognitive science and cognitive psychology could be characterised as the study of response time; it is "psychology's ubiquitous dependent variable" and has been collected, calculated, and studied since the earliest days of quantitative psychology (Luce, 1986; Woodrow, 1914). Second, amongst the basic principles of cognitive science are that there exists an internal, and not directly observable, mental structure; that the organisation and operation of the components of that structure can be inferred from external phenomena, especially behaviour; and that a well-formed theory of this structure can be used to account for the observed phenomena. Thus, the promise – or at least

the premise – of the study of response times is that they can be used to draw indirect inferences about the characteristics of that underlying mental structure.

The basic experimental paradigm for collecting response time data is that of the *reaction time task*. In this paradigm, subjects are presented with a series of reaction signals to which they are required to give appropriate responses; the signals and appropriate responses vary according the particular experiment at hand. Apart from this chapter's eponymous simple reaction time task, common species in this class include the choice reaction, the go/no-go, and the stop signal (or countermanding signal) tasks. It is a broad class, encompassing experiments designed to probe aspects of cognition as varied as decision making, response inhibition, attention, and memory, but its members are unified by the common demand for participants to respond as quickly and accurately as possible to the target stimuli. Reaction time tasks – especially the three mentioned above – have been used in a great deal of significant experimental, theoretical, and computational work, and this corpus arguably represents much of our most detailed understanding of the processes of cognition.

The ubiquity of reaction time tasks, their central role in cognitive science, and the advanced developments made in their theoretical examination underlie our choice to study such tasks in general. Within that broad compass, we choose to narrow our focus further, centring it on the simple reaction time task narrowly (and one-choice reaction time tasks more broadly, though we shall speak mainly of the SRT). In the SRT, only one reaction signal is presented per trial and the subject's sole aim is to respond as quickly as possible to its appearance (in common with other one-choice tasks); trials occur at a high rate. We restrict ourselves in this manner for three reasons. First, the SRT can be considered the prototype of all reaction time tasks, the rest of which can be seen as extensions or generalisations of this one task; it thereby derives unique significance from its direct relation to all other reaction time tasks. Second, the cognitive processes underlying SRT performance are of inherent interest. Third, amplified by the import of the first two points, there is considerable scope for novel and meaningful insights in the study of the SRT. Let us say something more about each of these reasons and draw out their interrelations.

We have stated that the SRT can be considered the prototype of the whole class of reaction time tasks; let us explain what precisely is meant by that statement. It was mentioned earlier that reaction time tasks vary widely, both in form and emphasis. This variety is deliberate: the foundational presumption of reaction time tasks, *qua* experimental paradigm, is that under systematic variation of the reaction stimuli, the manner of their presentation, and the rules governing responses to them, the nature and relation of the cognitive processes thus differentially demanded can be brought out. Nevertheless, the *sine qua non* of all reaction time tasks is for a subject to respond appropriately to a stimulus; and in that essential aspect, the SRT represents the most elemental combination possible of stimulus and response. Only one stimulus exists, and only one response; no cognitive operations are required to successfully perform the task beyond

perceiving the response stimulus and deciding to respond. (That is, in principle – it is not to say, as we shall see, that no other processes take place. But perception and decision-making are the principal, imperative attributes of task performance.) There is no need, for example, to determine *which* stimulus is present, or decide which of many responses is appropriate.

Every other reaction time task, whatever modifications it makes to this basic scheme, will *inter alia* also require the perception of a reaction stimulus and a decision to respond; there could otherwise be no meaningful task. The inverse cannot be said: the imperative attributes for proper performance in other reaction time tasks do not, in general, carry over to others. For example, the inhibition of responses required in the stop signal task has no counterpart in the choice reaction task; likewise, the identification of the reaction stimulus, essential to the choice reaction task, plays no role in a go/no-go task. Given this unique and fundamental relationship the SRT has to all other reaction time tasks, it seems reasonable to state that a complete understanding of the response process in the SRT is a prerequisite to a complete understanding of the almost-indubitably more sophisticated response processes employed in other reaction time tasks. Working out the full significance of this position will be taken up as a major theme later.

Apropos the second reason on our list, while the SRT is, by name and in description, a simple task, the phenomena it exhibits and the mechanisms underlying SRT performance are not. There are two related lines of reasoning here, one empirical and the other theoretical. On the empirical branch, there are strong, consistently observed effects – which we shall refer to collectively as *foreperiod effects* – that persist across various experimental manipulations and that indicate the presence of a latent and dynamic process of preparation in the period prior to onset of a reaction signal. This preparatory process is directly implicated in the decision-making process. Most importantly, mean response time is heavily affected by both the length and distributions of the "dead" intervals (the aforementioned foreperiods) preceding the signals. The empirical evidence from which we can infer the presence of complex mechanisms underlying SRT performance will be covered in more detail in §4.5.

On the theoretical branch, at heart, the task is a test of two aspects of cognition which are of fundamental importance: prediction and attention. To perform the task well, a subject must be able to predict reasonably well when the next response stimulus is likely to appear; she must also be able to sustain attention in order to be ready to detect the signal when it does, in fact, appear. There is a potentially complex relationship between these two processes: attention is costly, so the ability to accurately predict the onset of the response stimulus is valuable insofar as it allows one's attention to be dropped when the signal is unlikely to appear; but attention must first be paid to the incoming stimuli for accurate predictions to be formed. Prediction, of course, means applying past experience to form expectations of the future, so basic mechanisms of memory and learning are also implicated.

If one takes seriously the notion that the cognitive processes underlying SRT performance are

involved in other reaction time tasks, this complexity has obvious bearing on our understanding of more complex tasks. Since the process of signal detection and the decision to respond are governed by sophisticated cognitive mechanisms, and since those mechanisms are likely involved in other, more elaborate tasks, any proper understanding of these tasks must incorporate a satisfactory account of these more basic processes. As a corollary, we use more elaborate tasks to probe higher-order cognitive processes, such as signal identification and discrimination; it is difficult to see how we can isolate and control the desired object of study without being to disentangle the effects of the more basic processes.

Finally, it was claimed that there is considerable scope for novel and meaningful insights in the SRT. Let us set that statement into context. Experimental work on the SRT has been performed since the earliest days of experimental psychology (Woodrow, 1914), when it was viewed as a fundamental means of measuring attention; there was also a high degree of experimental interest in the period following the Second World War up to the eighties which yielded numerous important empirical findings, effectively flanked by two review papers, Teichner (1954) and Niemi and Näätänen (1981). After a relative hiatus in the latter eighties and nineties, interest in the topic continued – see, for example, the work of Los and co. (Los and Van den Heuvel, 2001; Los et al., 2001, 2014, 2017); of Vallesi and colleagues, (Vallesi and Shallice, 2007; Vallesi, 2010; Vallesi et al., 2007, 2014, 2013); and of Steinborn and co. (Steinborn et al., 2008, **Steinborn&al2010**; 2009; Steinborn and Langner, 2011, 2012) – but appears to have been overtaken when compared to studies of, e.g., choice reaction time.

Importantly, the largely experimental nature of the work on this task means that only a few attempts have been made to explicitly mathematically model the cognitive processes underlying performance in this fundamental task. Of the mathematical work that has been done, the majority has focused on explaining the distribution of response times, often not specifically in the SRT. Notable examples in this vein include the work of Hohle (1965), who introduced the ex-Gaussian model of response time; that of Burbeck and Luce (1982) and later Smith (1995), who worked on the notion that responses are driven separately by both the absolute level and changes in the level of response stimuli; and that of Schwarz (2001), who introduced the ex-Wald model of RT. Furthermore, a few relatively independent attempts have been made to understand and model the SRT, rooted in them modelling of the two-choice forced reaction task (Ratcliff & Van Dongen, 2011). By focusing on the modelling of RT distributions, this research has largely set aside the modelling of foreperiod effects. Meanwhile, though qualitative theoretical accounts of the mechanisms underlying foreperiod effects have been forwarded alongside experimental work, few if any of these have had substantial mathematical content.

The incompleteness of research on the SRT has two aspects, then. The first is absolute: there are fundamental aspects of SRT performance which remain ill understood. Mathematical models which capture most of the known empirical data are lacking in this most fundamental of tasks.

The second is relative: this lack of understanding and research effort, at least in the last few decades, seems especially pronounced when compared to these similar, but more complex, tasks, for which comprehensive empirical and (especially) theoretical-mathematical work has been done; the choice reaction task, with the widely adopted Ratcliff drift-diffusion model, and the stop signal task with Logan's independent horse race model, come to mind (Logan & Cowan, 1984; Ratcliff & McKoon, 2008). Beginning to remedy this shortfall will be a major theme in later chapters, especially chapters 8 and 9.

The remainder of this chapter constitutes an introductory survey of the literature on the SRT. The purpose here is not comprehensiveness, but simply to cover the main findings and ideas which are required to follow the argument of later chapters. The first section will describe the experimental paradigm and its manipulations. The second and third sections will cover the well-established behavioural phenomena that have been recorded in the SRT, and the theoretical constructs that have been put forward as explanations of those phenomena. The final section will cover common mathematical models of response time, which will form the basis of the theoretical studies covered in later chapters.

2. DESCRIPTION OF THE TASK

2:1 Experimental design

Basic description

The simple reaction time task (SRT) belongs to a broader class of tasks known as *one-choice reaction time task*, the distinguishing feature of which is the choice structure (or lack thereof) in responding: the only decision being made is whether to respond. All tasks in this class consists of a series of independent trials; on each trial, the subject performing the task is presented with *response signal* (RS), the sensory stimulus to whose appearance they must respond, and is offered an opportunity to respond. The sole aim of each trial is to respond as quickly as possible to the RS once it appears, but not before (a so-called *speeded response*).

The various one-choice RT tasks are further distinguished by their temporal structure. For example, in the classic "vigilance" task, stimulus onsets are widely spaced in time – say, twelve events in twenty minutes – and in the "psychomotor vigilance test", the stimuli appear more frequently – several per minute (Ratcliff & Van Dongen, 2011). The simple reaction time task is defined by the short times between of stimulus presentations, ranging from a few hundred milliseconds to not more than a few seconds. The SRT will dominate our attention in this work; we mention the other one-choice tasks for later context, and because numerous considerations that we make in the SRT can be extended to the other tasks.

To perform the SRT, the subject is familiarised with the RS prior to starting. It usually takes

the form a simple visual or auditory stimulus, for example, a shape appearing on a screen, the change in luminance of a patch of pixels, or the onset or offset of a pure tone against background noise, to give some common examples (see Burbeck and Luce, 1982; Smith, 1995; Los et al., 2001, 2014, 2017; Ratcliff and Van Dongen, 2011). During the task, each trial has an identical structure (figure 4.1). Within a trial, the RS is preceded by a *foreperiod* (FP) of variable duration, during which only an unchanging background signal is present; the onset of the FP is marked by a *warning signal* (WS), a brief even that clearly marks the start of a trial. A trial may be terminated in a number of ways: by subject response, by the lapse of a pre-determined amount of time, or a combination of the two (most trials being terminated by response, with the remainder being terminated after a timeout period). The timecourse of single trial in the SRT therefore comprises the WS, the FP immediately preceding the presentation of the RS, the duration of the RS, and the terminating event. Trials may be separated by an intertrial interval, usually on the order of a few seconds or shorter in length.

Some terminology to refer to within-trial events is useful. The *critical moments* (CM) are moments in which it is possible that the RS will appear; it is clear that the CMS depends upon the distribution of foreperiods being used. The moment at which the RS is actually presented on a given trial is called the *imperative moment* (IM).

The variation of experimental parameters

Aspects which can be varied within this experimental design are the format of stimulus presentation, the presence of performance feedback, and the distribution of foreperiods; of these, the last is most important for our purposes. We will now briefly discuss the parameters that can varied within each of these aspects, and their effects on the task's cognitive demands.

RESPONSE SIGNAL. The response signal remains consistent throughout a given run of any simple RT task; this is the defining difference between simple and other RT tasks. The response signal is If the stimulus if visual, it will be presented in a known location, e.g., in the centre of the screen. The stimulus which constitutes the RS can itself can be modified along a number of dimensions. First, the intensity of the signal can be increased or decreased, with more more salient signals provoking faster responses. Second, the transitions of stimulus from present to absent can be smooth or sudden. Finally, the response signal can be either the onset or the offset of the distinctive stimulus. The termination of the signal can occur after a fixed period of time, which may be long or short, or may be conditional on a subject's response (with some potential timeout limit).

PERFORMANCE FEEDBACK. Subjects may be given feedback on their performance; little research which is concerned with the fundamental mechanisms underlying SRT performance addresses the question of how this feedback may affect performance, so this experimental manipulation will not be considered any further.

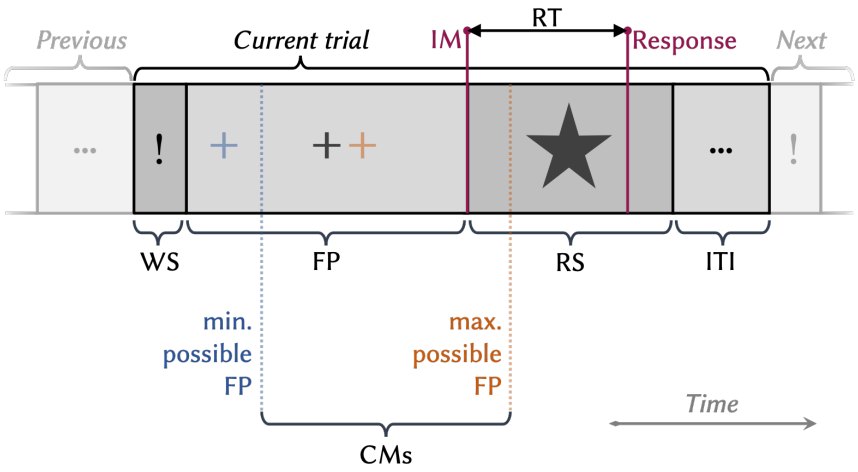


Figure 4.1. A SCHEMATIC OF THE TIMECOURSE OF A SINGLE TRIAL OF THE SIMPLE REACTION TIME TASK. Abbreviations: WS – warning signal (may not be present); FP – foreperiod (usually variable, may vary continuously or discretely); RS – response signal; ITI – intertrial interval (may not be present); CM critical moment; IM – imperative moment; RT – response time. See text for full explanations of the terms.

FOREPERIOD DISTRIBUTION. Variation in the distribution of foreperiods is probable the single most important experimental manipulation for the SRT, at least for our purposes. The major distinction to be made with regards to foreperiod distribution is between constant and random foreperiod designs. In the constant FP variant, the foreperiod on each trial is the same, typically on the order of 500 to 2500 ms. In random FP designs, the foreperiod for each trial is drawn IID[i] from a pre-determined distribution. Common distributions are the exponential, uniform, and anti-exponential distributions; these are usually shifted so that there is a fixed minimum FP length, typically on the order 500 to 2000 msecs. The distribution may be realised continuously or discretely, with most research opting for discretised versions.

The presence of random (as opposed to constant) FPS is the main driver of the attentional and decision-making demands of the SRT: uncertainty about the timing of the imperative moment forces sustained attention to the location of the RS and an ongoing decision process about whether the RShas in fact appeared, thus warranting a response. Given a random FP design, the chief effect of choosing an FP distribution is to change the information carried by the time elapsed since the start of the trial; the discussion of the significance of this proposition will be deferred until §4.5.

2:2 Task measures

The principal measure of the SRT is response time (RT, also called response latency), measured from the imperative moment; the secondary measure is whether the trial was successful or failed.

[i] Independently and identically distributed, i.e., the value of the FP drawn on one trial does not depend on the value drawn on any other, and each one is drawn from the same distribution.

A successful trial occurs when a response is given after (and only after) the imperative moment and before trial cessation. A trial can be be failed by responding either too early or too late. If a response is given too soon – say, during the foreperiod – the response will be recorded as *premature*. If no response is given by the end of the stimulus presentation period, the response will be recorded as a *late*. In general, only the first response on a trial is counted.

It is worth noting here the (putative) distinction between *reactionary* and *anticipatory* responses. Reactionary responses are those made as a direct consequence of the presence of the response signal; that is, they are precipitated by its appearance. Anticipatory responses occur when a subject gives a response because they expect the occurrence of a RS at a certain time, not because they have reacted to its actual appearance. If the distinction above is upheld, anticipatory responses will be classified as premature.[ii]

The above definitions leave scope for some ambiguity on two matters: first, whether a response recorded during a foreperiod is a late response to the preceding RS or a premature response to the succeeding one, and second, how to discriminate between slow premature responses and fast reactionary ones, which may be recorded as RTs of similar magnitude. In practice, however, the simplicity of the SRT renders the job of sensibly classifying responses possible with a relatively simple scheme, and the extremely high proportion of hit trials means that very few trials are difficult to classify. The technical question of how to determine to which trial a response belongs, and whether it is a hit, premature, or late, will be taken up in the general methods section of chapter 7; for now, it will suffice to say that it can be done, and that any such processing will only affect a slight minority of the total data.

The remainder of this chapter is split into four parts. The first reviews the major empirical findings regarding the distribution of response time – its typical form and variations under changes in the response stimulus. will cover mathematical models of response time that have been used to account for simple RT data. The second will cover mathematical models of response time that have been used to account for this simple RT data. The third lays out set of findings relates to the complex effects of foreperiod length, distribution, and sequence on mean RT. The final section summarises the current state of knowledge regarding the SRT.

3. THE DISTRIBUTION OF RESPONSE TIMES

As has already been discussed, the primary dependent variable of the SRT is response time. Response times vary unpredictably within individuals, even under identical conditions; for that reason, RT is usually treated as a random variable, and the analysis of RT is focused on the *distribution* of response times as a whole. In order to properly describe RT, we now introduce some

[ii] The distinction between anticipatory and reactive responses will be explored further in the analysis of the phenomenon I refer to as *corrected responding*, reported in Chapter 7

formal mathematical concepts and terminology regarding the representation of distributions in general; their significance will become evident as the text unfolds.

The characteristics of a distribution can be represented in numerous ways, mathematically speaking; while each representation in principle carries the same information about the underlying distribution, they emphasise different facets. Two representation are of particular importance when considering distributions of response times, the familiar probability density function (PDF) and the *hazard function*, which – as we shall see – form natural companions in this application. Given the distribution of a random variable T, the PDF $f(t)$ of the distribution specifies the proportion of observations for which the random variable obtains a value in the small interval $(t, t + dt]$. Integrating the PDF gives the cumulative density function (CDF) $F(t)$, which gives the probability that T will assume a value below t; the CDF's complement is the survivor function $S(t) = 1 - F(t)$, which gives the probability that T will obtain a value of *at least* t. The hazard function, also known as the conditional density function, is formally defined as the ratio of the PDF to the survivor function,

$$h(t) := \frac{f(t)}{S(t)}. \tag{4.1}$$

Conceptually, this function gives the conditional probability of T obtaining a value in the small interval $(t, t + dt]$, given that it has not obtained a value below t. It can be readily shown that the survivor and hazard functions are related by the equation

$$h(t) = -\frac{d}{dt} \log\, S(t) \tag{4.2}$$

and the PDF and hazard functions by

$$f(t) = h(t)\, \exp\left(-\int_0^t h(x)\, dx \right). \tag{4.3}$$

The hazard function is a natural representation of a distribution when one thinks of it as arising from a probabilistic process unfolding in time, with the distribution's random variable representing the time at which some event in that process occurs (such as its termination); this is because the hazard function constitutes the answer to the question, "What is the probability of the event of interest occurring momentarily, given that it has not occurred yet?" Transposing these definitions to the discussion of response times, the PDF gives the probability of a response occurring in the small interval $(t, t + dt]$, and the hazard function gives the probability that a response will occur in the brief moment $(t, t + dt]$ given that it has not already occurred by time t.

The formal aspects now dispensed with, let us turn to the description of actual RT distributions. We start by giving summaries of the main findings, before displaying numerous examples from which they are abstracted.

In common with other response time tasks, such as the choice reaction task, response

latencies in the SRT are distributed in a stereotyped manner. The chief qualitative attributes of the SRT's density function are an early peak of fast responses and a long right tail of slower responses. The exact location of the peak and the length of the tail vary from subject to subject; within subjects, the peak and spread of the distribution vary systematically according to stimulus and general factors such as wakefulness. In all cases, however, the general form of the distribution remains the same.

When comparing distributions, the PDF representation makes differences in the early peak visible at the expense of eliding them in the tail, because the relatively low probability of events in the tail compared to the peak makes differences in the tail imperceptible. Distributions with quite distinct behaviour in the tails can thus come to seem similar. Representing the distribution through its hazard function reemphasises the differences in tail, since the function shows the conditional, not the absolute, probability of responding once the later times have been reached.

In general, hazard functions of two kinds have been reported in the SRT. The first of these has an early peak, followed by decay to a constant level. The second rises more slowly to a maximum which it maintains, perhaps with a slight dip towards the end. The early-peaked form of the hazard function tends to follow from high-intensity stimuli; the steadily rising form from low-intensity or ramped-up stimuli.

Figures 4.2 and 4.3 display a sample of the reported findings from which the above general conclusions are drawn. For details of each, see the individual figure captions. Note the consistency of the results: in both experiments, high-intensity stimuli produce hazard functions with an early peak and a long, high tail; at lower intensities, the early peak is subdued or absent entirely.

With these features of the RT distribution in mind, we now turn our attention to the mathematical models which have been put forward as descriptions of the same.

4. MATHEMATICAL MODELS OF RESPONSE TIME AND THE SRT

4:1 Shifted Wald model

The shifted Wald distribution has been used as a model of response time since the earliest days of mathematical psychology (Stone, 1960). The basic view of this model is that response time in the SRT can be decomposed into the sum of two parts: a constant time δ_t and a variable time t distributed according to the Wald (also known as the inverse Gaussian) distribution. The variable t is nominally identified with duration of the decision-making – that is, the cognitive – processes involved in responding; and the constant δ_t as a summary variable for all the non-decision-making components, for example, motor response time. We deal first with the aspects of the distribution putatively relating to decision latency.

In the abstract, the Wald distribution arises from considerations regarding the motion of a

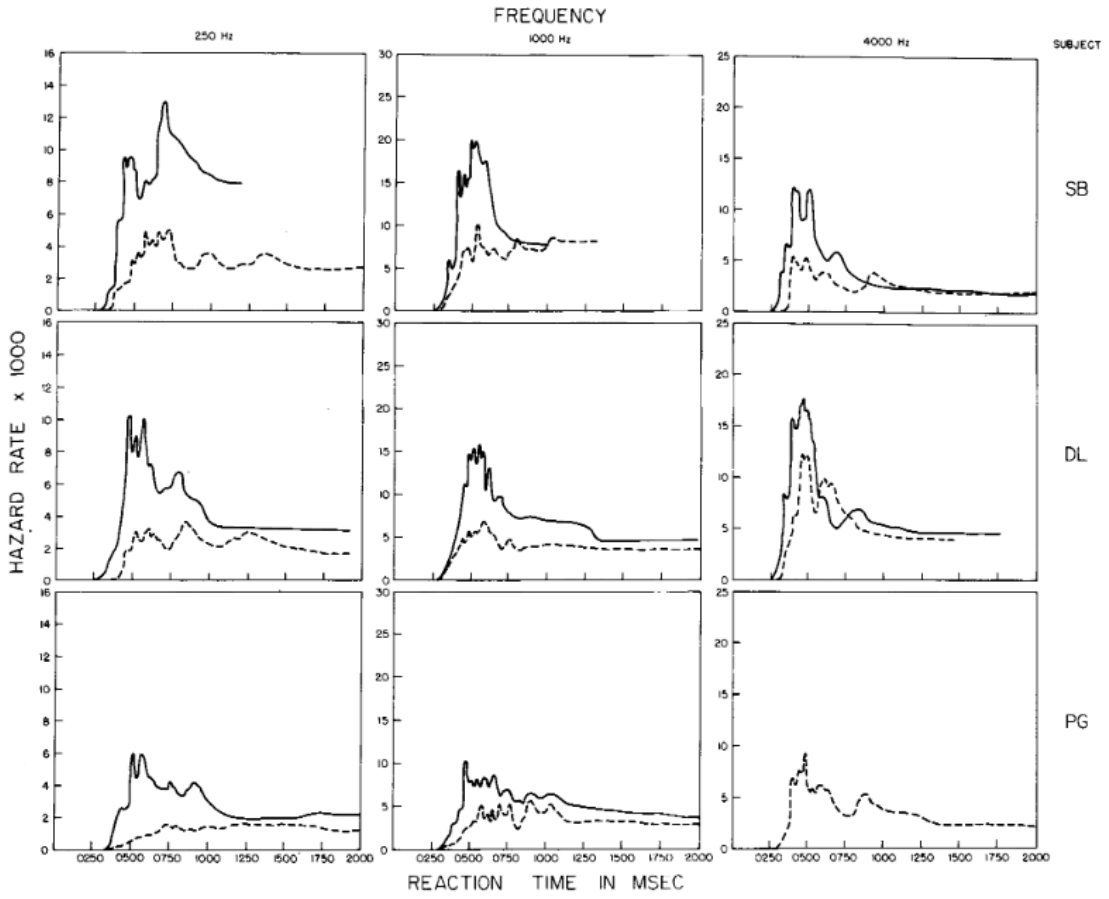


Figure 4.2. Representative response time distributions and hazard functions reported in the literature. Hazard functions reported in Burbeck and Luce (1982). In this experiment, the RS was an on offset in a pure-tone signal presented against background noise. The foreperiods were distributed exponentially with a three-second mean after a 200ms delay, and there were six experimental conditions, formed from the combination of three frequencies for the pure-tone signal and two intensity levels for each signal. Each row represents a subject's results, each column a frequency condition; the solid and dotted lines represent the high and low intensity conditions. Note the early peak in the high intensity condition, attenuated or absent in the low.

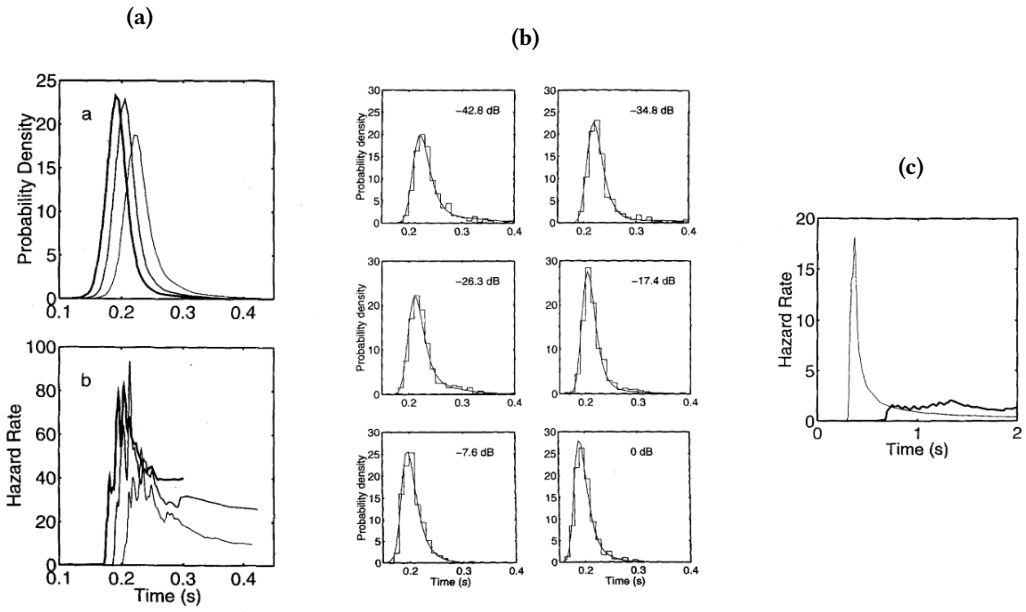


Figure 4.3. FURTHER REPRESENTATIVE RESPONSE TIME DISTRIBUTIONS AND HAZARD FUNCTIONS REPORTED IN THE LITERATURE. From the experiments reported in Smith (1995), reproduced with permission. All data represents around 400 trials.(a) Response time distributions and corresponding hazard functions for a single subject under a range of experimental conditions. The light line shows the hazard function for detecting a 0.63-contrast, low-frequency (2.6 c/deg), abrupt-onset patch. The heavy line shows the hazard function for a 1.0-contrast, high-frequency (13.20 c/deg) patch that was ramped on over 285 ms.(b) Histograms of response time distributions for a single subject under different experimental conditions. Smooth curves represent the fit of a model to the data.(c) Two hazard functions for a single observer. The light line shows the hazard function for detecting a 0.63-contrast, low-frequency (2.6 c/deg), abrupt-onset patch. The heavy line shows the hazard function for a 1.0-contrast, high-frequency (13.20 c/deg) patch that was ramped on over 285 ms. Note again the strong early peak in the high-intensity stimulus, absent in the other.

particle in one dimension, where the motion is governed by two factors. Denote the position of the particle at time t by $x(t)$. The first factor governing the motion of the particle is a constant drift, which drives the particle along x at a constant rate v. The second factor is random, Brownian motion, which can be thought of as a perturbation to the particle position by an increment $\eta\xi(t)\sqrt{dt}$ in any arbitrarily small interval of time dt. The quantity $\eta\xi(t)\sqrt{dt}$ is the size of the perturbation at t, relative to $t - dt$; $\xi(t)$ is drawn independently at each time t from a standard normal distribution, so that the total perturbation, because of the coefficient η, is characterised by a zero-centred normal distribution with variance η^2.[iii] From this set of conditions, the Wald distribution emerges as the solution to the following problem: given a particle starting at $x(t = 0) = 0$, and a particular threshold point along x, which we denote θ, what is the probability that the first time the particle meets that threshold is t? It can be shown that the distribution of crossing times is given by

$$f(t \mid v, \eta, \theta) = \frac{\theta}{\eta\sqrt{2\pi t^3}} \exp\left[-\frac{(\theta - vt)^2}{2\eta^2 t}\right]. \tag{4.4}$$

Note that v, η and θ are only ever expressed in ratios of one another. This entails that there are effectively only two *free* parameters in this distribution, for setting two of the three possible ratios v/θ, θ/η and η/v fixes the third, which in turn fixes the distribution. The form given in eq. (4.4) is thereby overparameterised. One can simplify the distribution by setting the value of one any of the parameters to some convenient value, forming the base unit in which the two parameters left free are expressed; the distribution can then be further reformulated by defining new parameters in terms of the free ones and performing the relevant substitutions. Applying this sort of reduction by defining $\mu = \theta/v$ and $\sigma = \eta/\theta$ gives rise to the following sufficient parameterisation of the Wald model:

$$f(t \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(t - \mu)^2}{2\mu^2\sigma^2 t}\right]. \tag{4.5}$$

Note that this parameterisation is equivalent to setting $\theta = 1$ and measuring the drift rate and perturbation standard deviation of the diffusion process in units scaled by the initial distance to the threshold. As the scaled inverse of the drift rate v, μ gives the mean first passage time across the threshold, that is, the mean decision time. The parameter σ controls the contribution of the random diffusion relative to the drift process; small values of σ mean that the constant drift dominates, culminating in narrower and less skewed resultant distributions of decision time, whereas larger values entail more involvement from the random diffusion factor, making the resultant decision time more variable. A schematic of the shifted Wald is shown in figure 4.4.

From a formal point of view, the choice of parameterisation is arbitrary; however, substantive

[iii] Formally, a Wiener process.

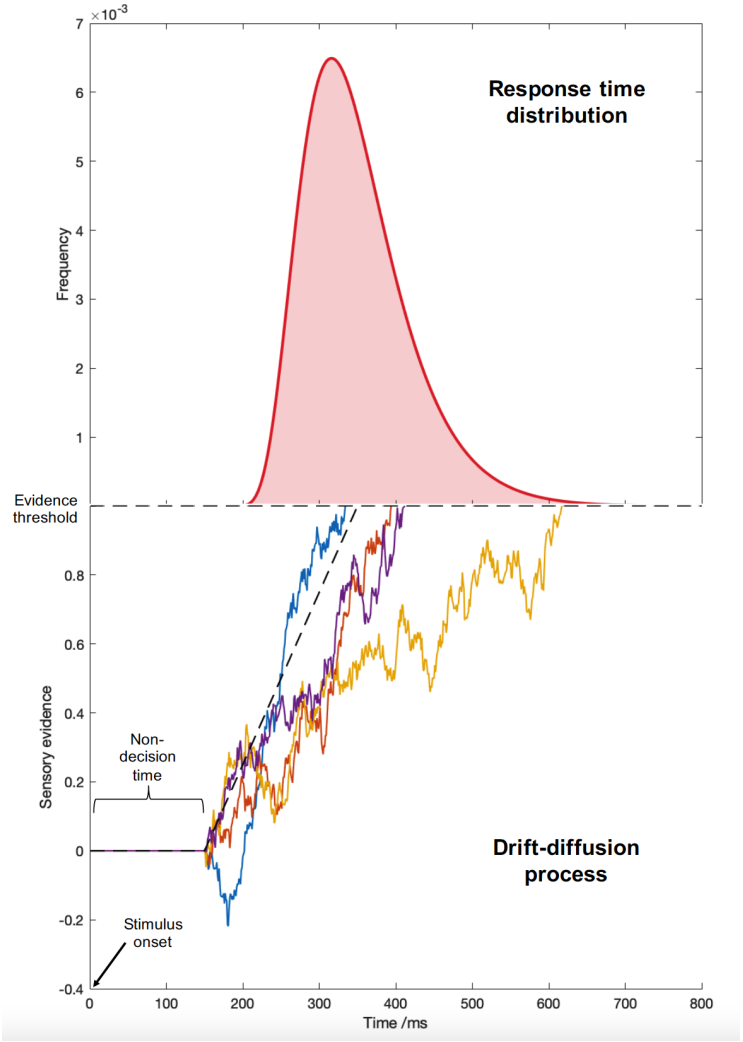


Figure 4.4. A SCHEMATIC OF THE MECHANICS OF THE SHIFTED WALD MODEL. The bottom panel of the figure shows the drift-diffusion process. The dashed line represents the mean drift rate of sensory evidence accumulation; the coloured lines are sample paths with the noisy diffusion process added. The top panel shows the RT distribution that results from the aggregate of threshold passing times across many samples. As can be seen, the noisy diffusion process pushes some paths to reach the threshold faster than average, and others slower; because the faster-accumulating samples must still reach the threshold in positive time, but slower-accumulating ones can stay under the threshold indefinitely, the characteristic long right tail of slow responses in RT distributions is recovered.

meaning is often ascribed to particular parameters, which in turn affects the choice of parameterisation, as will be discussed shortly. To engage in this discussion, we will first digress to make a conceptual reading of the distribution and its generating process as it relates to decision-making.

This distribution, which at first glance possesses all the necessary features of a response time distribution (once δ_t is incorporated), can be taken as a mere mathematical description of the observations, with no deeper meaning than that. However, a cognitive interpretation of the Wald distribution as an account of the decision-making process is often given. In this view, the motion encoded by $x(t)$ and the passage of the threshold θ represent, at least approximately, the workings of a real cognitive process; in this process, the "motion" x represents the accumulation of a quantity, often called *sensory evidence*, which encodes the degree of response-relevant information that the subject has obtained. When enough sensory evidence has accumulated to pass the threshold θ, an irrevocable decision to make an overt response is passed (and not before). Clearly, in this view, ν and η represent the rate and noisiness of this evidence-accumulation procedure.[iv]

Now recall that only two of the three parameters are required to fully specify the distribution of decision latencies. The usual choice is to keep η fixed at some arbitrary level, so that the distribution is expressed in terms of $\nu' = \nu/\eta$ and $\theta' = \theta/\eta$. Seen through the lens of the above interpretation, this imbues appealing substance to the values of the parameters ν' and θ'. The value of θ' represents the subject's internal criterion for the amount of sensory evidence required to respond: low values of θ' mean a lax criterion, which result in faster responses from less information; high values entail a strict criterion and correspondingly slower responses. In turn, ν' represents the rate at which response-relevant can be extracted from the sensory information being presented; thus, easy-to-detect signals correspond to higher values of ν' and quicker responses than hard-to-detect ones. The choice of η as the scaling factor can be explained away by saying that it represents a natural unit of measurement, since it is presumably a constant quantity intrinsic to the processes underlying sensing and cognition. This interpretation opens up the promise that precise quantitative descriptions of the underlying cognitive process can be inferred, allowing for the direct measurement and comparison of cognition in a variety of circumstances.

As this interpretation illustrates, while different parameterisations of the same model are, of course, formally equivalent, they may reflect different preferences regarding the interpretation of the model and its predicted distribution with respect to the data. Assuming the parameterisation above, for example, reflects a bias towards viewing a model as a direct description of the underlying cognitive process, with the task of parameter estimation being to provide measurements of these unobservable imputed processes. Despite the appeal of this interpretation, we resist using the

[iv] The term "evidence" unfortunately happens to apply to two basic concepts involved in this study: Bayesian evidence, the measure of the support that data gives to a particular model, and sensory evidence, the theoretical quantity upon which decisions are made. We flag this elision here and trust the reader to understand which concept is meant by the context.

parameterisation which it suggests, and will instead adopt the form in equation (4.6). Justifications for this choice can be found in §6.2:4.

The treatment of the non-decision time δ_t is considerably more straightforward. Formally speaking, the *shifted* Wald distribution implies nothing more than subtracting δ_t from t, to capture the offset of the response time from the decision time, meaning that instances of the variable t in the PDF need only be replaced by $t - \delta_t$. The complete PDF for response times t of the shifted Wald distribution we will adopt is therefore

$$f(t \mid \mu, \sigma, \Delta) \;=\; \frac{1}{\sqrt{2\pi\sigma^2 (t-\Delta)^3}} \exp\left\{ -\frac{[(t-\Delta)-\mu]^2}{2\mu^2\sigma^2(t-\Delta)} \right\}. \tag{4.6}$$

[v].

A few notes for future reference: In what follows, where a variable X is referred to as being Wald-distributed, written as $X \sim \text{Wald}(\mu, \sigma, \delta_t)$, we will take it to mean according to the form in equation (4.6) unless otherwise specified. Also, we remark now that the issue of parameterisation will become important again when discussing the selection of suitable priors for the model.

4:2 Ex-Gaussian model

The exponentially modified Gaussian distribution, commonly called the ex-Gaussian and introduced by Hohle (1965), also models response latency as the sum of two variables. In contrast to the shifted Wald model, though, both are taken to be random variables, with one exponentially and one normally distributed. That is to say, under the ex-Gaussian model, total response time can be decomposed into the sum $t = t_{\text{Exp}} + t_{\text{Gauss}}$ where t_{Exp} is a random variable distributed exponentially with decay (or mean) parameter τ, and t_{Gauss} is a random variable drawn from a normal distribution with mean μ and variance σ^2. It can be shown that distribution of t under these assumptions cleaves to the PDF

$$f(t \mid \tau, \mu, \sigma) = \frac{1}{\tau} \exp\left[\frac{1}{\tau}\left(-t + \mu + \frac{\sigma^2}{2\tau} \right) \right] \Phi\left[\frac{t - \mu - (\sigma^2/\tau)}{\sigma} \right], \tag{4.7}$$

where $\Phi(\cdot)$ is the CDF of the standard normal function. The mean and variance of the Gaussian component determines the position and spread of the main peak of the distribution; the exponential component gives it a characteristic positive skew; and the ratio τ/σ determines the dominant component and thereby the overall shape of the distribution. Note that the ex-Gaussian model does not need a shift parameter, like δ_t in the Wald distribution above, since its function is performed by and absorbed into the mean of the normal component.

Numerous psychological interpretations have been applied to the ex-Gaussian model; all

[v] Note the slight abuse of notation: the variable t now stands for response time where it was used for decision time in the discussion around equations (4.4)-(4.6)

essentially involve associating the exponential and Gaussian components with at least two separate cognitive processes which act in series to produce a response, be they motor response and decision-making; or change and level-detectors in sensory processing; or the action of sustained and transient response channels. The motivations and justifications for these interpretations are beyond the scope of this piece; readers are referred to Hohle (1965) and Stone (1960) as a starting point if they wish to learn more.

4:3 Ex-Wald model

The exponentially-distributed Wald (ex-Wald) model, like the ex-Gaussian model, assumes that response times are generated through a serial process, with the total response time being the sum of the durations of two stochastic processes Schwarz (2001). Compared to the ex-Gaussian, the ex-Wald keeps the exponential component, and substitutes a Wald-distributed random variable in place of the Gaussian one. Taking the exponential component to be $t_{\text{Exp}} \sim \text{Exp}(\tau)$, where τ is to be read as the decay parameter, and the Wald component to be $t_{\text{Wald}} \sim \text{Wald}(\nu, \sigma, \theta)$, we have $t = t_{\text{Exp}} + t_{\text{Wald}}$. The distribution that arises from this convolution can be shown to be

$$f(t \mid \tau, \nu, \theta\,;\, \sigma) = \frac{1}{\tau} \exp\left[-\frac{1}{\tau} + \frac{\theta(\nu - \kappa)}{\sigma^2}\right] F(t \mid \kappa, \nu, \theta) \;, \tag{4.8}$$

where we define $\kappa \equiv \sqrt{\nu^2 - 2(\sigma^2/\tau)}$ and where

$$F(u \mid \lambda, \rho, \omega) = \Phi\left(\frac{\lambda u - \omega}{\rho\sqrt{u}}\right) + \exp\left(\frac{2\omega\lambda}{\rho^2}\right) \Phi\left(-\frac{\lambda u - \omega}{\rho\sqrt{u}}\right), \tag{4.9}$$

$\Phi(\cdot)$ once again being the CDF of the standard normal function. As with the Wald distribution, the distribution is uniquely identified by any two of the ratios ν/θ, θ/σ, and σ/ν, so one of those three individual parameters can be set to an arbitrary value, and the exact distribution fixed by the remaining two, without loss of generality. Following the decision for the Wald distribution, and as indicated above, we choose henceforth to fix σ. Alike to the ex-Gaussian model, the positions and spread of the main peak of responses is determined by the Wald component, with the exponential component providing a further positive skew.

If one wishes to give cognitive gloss to the model, one may interpret the Wald component as representing the decision-making stage, where it is understood to occur in essentially the way described above for the Wald model, i.e., as the accumulation of "sensory evidence" to a decision threshold; the exponential component is then introduced to summarise the effects of all processes succeeding (and possibly preceding) the decision stage, such as the motor response stage. In effect, this model forgoes the shift parameter of the basic shifted Wald model in favour of an additional exponentially distributed random variable. Note that this interpretation provides a theoretical justification for the absence of a shift parameter.

4:4 Ratcliff one-choice drift-diffusion model

The Ratcliff one-choice drift-diffusion model (ROCDD, Ratcliff and Van Dongen, 2011) assumes, like the shifted Wald model, that response time is the sum of a decision-making process, modelled as the noisy accumulation of sensory evidence, and a non-decision time, putatively covering the duration of any outstanding sensory and motor response processes. Unlike the shifted Wald model, however, the Ratcliff drift-diffusion model presumes that there exists random trial-to-trial variability in the values of the parameters governing the drift-diffusion process. Specifically, the drift rate on each trial is supposed to be drawn from a normal distribution with mean $\bar{v}$ and variance η^2, and the non-decision time is supposed to be drawn uniformly from the range $\left[\bar{\delta}_t - s, \bar{\delta}_t + s\right]$. As with the shifted Wald model, a decision to respond is made when the sensory evidence reaches a particular value, the fixed decision boundary θ.

There is no closed-form solution to the ROCDD. The means by which the PDF of this distribution can be calculated will be taken up further in later chapter, when the calculation of the log-likelihood is required.

5. FOREPERIOD EFFECTS

5:1 Overview

Across many studies and decades, strong and consistent relationships between FP duration and FP distribution with mean RT have been observed; these relationships persist irrespective of many details of experimental design, and form an interrelated family of phenomena that we shall call *foreperiod effects* (Gordon, 1967; Requin and Granjon, 1969; Nickerson and Burnham, 1969; Näätänen, 1971, 1975; Mo and George, 1977; Niemi and Näätänen, 1981, Polzella et al., 1989; Los et al., 2001, 2014, 2017 – to give a few). Since RT and *ipso facto* mean RT are taken to indicate something of the workings of the cognitive processes governing responding, these phenomena are *a fortiori* taken to reveal the presence of a process of *temporal preparation* in reaction time tasks.

In this literature, two conclusions stand out as being particularly important. The first is that participants in the SRT appear to be sensitive the statistical features of the overall foreperiod distribution (FPD) when performing an SRT task: the FPD influences the temporal dynamics of stimulus onset probability (SOP) over the course of a trial, and mean RT tracks the changes in SOP according to the FPD used. Second, participants appear to be subject to short-term learning effects whereby the experience of recent foreperiods influences their preparedness to respond at different critical moments in subsequent trials; in particular, recent experience of long foreperiods leads to slowed responding when later trials have short foreperiods. The remainder of this section

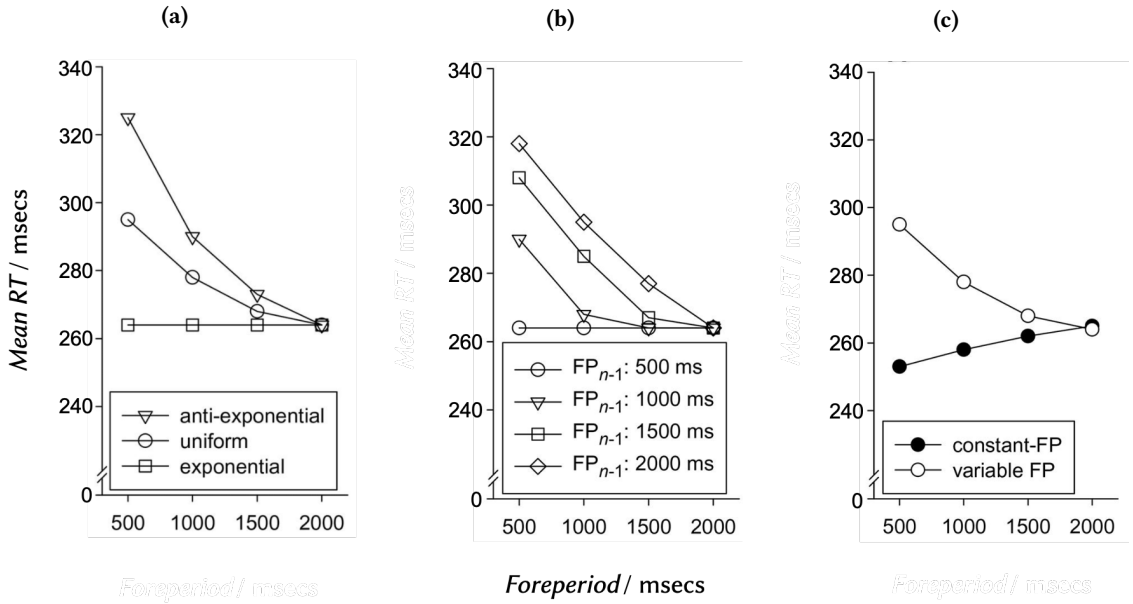


Figure 4.5. FOREPERIOD EFFECTS IN THE SIMPLE REACTION TIME TASK. (a) Mean response time as a function of within-trial foreperiod duration in three different random-foreperiod designs. (b) Mean response time as a function of within-trial foreperiod duration in a random-foreperiod design with uniform FPD, with trials segregated according to the previous trial's FP, and FP curves plotted separately for each partition. (c) Mean response time as a function of within-trial foreperiod duration in a random-foreperiod design with a uniform FPD compared to mean response time in constant-foreperiod design. Reproduced from Los et al. (2014) with permission.

will cover the empirical findings which form the basis of these conclusions, and the hypotheses that have been forwarded by way of explanation. Interested readers are directed to the excellent review by Los et al. (2014) for further details and references.

5:2 Empirical findings

Foreperiod distribution effects

The first member in the family of foreperiod phenomena reveals itself in two steps. The first establishes that mean RT (on a given trial) is a non-trivial function of foreperiod duration (on a that trial) across a range of FPDs; the second shows that, while these foreperiod-mean RT curves (henceforth simply *foreperiod curves*) are superficially different between FPDs, the foreperiod curves are related to one another in a systematic way.

Following Los et al. (2014), to explore these relationships further, we consider a series of three SRT designs. Each design has discrete foreperiods distributed across four critical moments at 500, 1000, 1500, and 2000 msecs; the designs differ in the FPD, with the discretised versions of uniform, exponential, and anti-exponential distributions considered.

Figures 4.5(a) shows the typical foreperiod curves obtained across these designs. Now, a uniform FPD will result in relatively fewer trials with short FPs than an exponential FPD; and an anti-exponential FPD will have even fewer short FP trials than a uniform FPD. As can be seen, then, the greater the preponderance of trials with long FPs in an FPD, the slower the average response to trials with short FPs will be. Furthermore, on those trials with the longest FP – where the IM occurs at the last possible CM, the mean response time across all the FPDs is equal. We will call this set of relationships the *foreperiod distribution effect*. Before we begin to interpret these results, though, let us examine the second major empirical phenomenon found in the data.

Asymmetric sequential foreperiod effects

The second member in the family of foreperiod effects emerges across trials. Rather than plotting FP curves using all the RTs collected during a task, the RTs are first segregated according to the FP on their preceding trials. In the designs we are considering, this results in four partitions of the data, one for each of the possible CMs. The FP curve is then plotted for each partition separately.

We will consider the results of this procedure under the uniform FPD, shown in figure 4.5(b). As can be seen there, the greater the FP on the preceding trial, the more pronounced the FP curve is on the current trial; indeed, when the FP on the preceding trial is the shortest possible, there is no FP effect (in the sense of no distinct relationship between mean RT and FP duration) on the current trial at all.

This amplifying or diminishing effect of the preceding trial's FP on the current FP curve has two notable features. The first is that it entails that there is a strong *sequential effect* of foreperiod on mean RT: long foreperiods on the preceding trial increase mean RT on the current trial. Second, as between the FPDs in the FPD effect, so between the partitions in the sequential foreperiod effect: on those trials with the longest FP, the mean response time across all the partitions is equal. That is, the overall increase in mean RT with increasing prior FP duration is driven by large variability in RT on trials with short FPs while long ones remain unaffected. In this way, this sequential effect is also *asymmetric*, for it does not affect long-FP trials in the same way as it does short ones.

Note that, in this discussion, the data being considered are aggregate data across subjects; indeed, this is how FP effects are often presented in the literature. It may be objected that the apparent FP effects might be the result of the process of aggregation, perhaps due to a few outliers or some other mixture effect, rather than being indicative of the actually object of interest: the effect of the task's temporal structure on *individual* performance. This objection will be dealt with in the study of empirical data in chapter 7, where it will be shown that the aggregate effects are in fact reflections of individual effects.

5:3 Theoretical discussion

Trace conditioning and the effects of recent history

How are these effects to be explained? To begin with, it is worth noting that the asymmetric sequential foreperiod (ASF) effect can provide an at-least partial explanation of the FPD effect. Suppose that the results of figure 4.5(b) hold generally, namely, that trial-to-trial long-to-short FP transitions produce considerably slower mean RTs, while long-to-long, short-to-long, and short-to-short produce less or no slowing. As the FPD shifts to include a smaller proportion of short FPs – say, going from an exponential to a uniform, or a uniform to an anti-exponential distribution – then the probability of a trial with a short FP being preceded by one with a long FP necessarily increases. By default, then, mean RT on short-FP trials will on average become slower as the shift in FPD is made, while the mean RT on longer-FP trials will slow less or not at all; which is congruent with the observations constituting the FPD effect.

This observation lies at the heart of the *trace conditioning* model of foreperiod effects (Los & Van den Heuvel, 2001), which posits that each critical moment is associated with a variable conditioned strength. The higher the conditioned strength of a CM, the faster the participant responds if the IM coincides with it. Each trial then leaves a "trace" on this conditioned strength according to three learning rules. First, the conditioned strength decreases as the CM is approached and passed without incident, with the momentary rate of decrease proportional to the distance from the CM ("extinction"). Second, when the RS is presented and responded to, the conditioned strength around that moment increases, with the momentary rate of increase proportional to the distance from the IM ("reinforcement"). Finally, after the IM has passed, the conditioned strength remains relatively unchanged. The changes in the trace then propagate to the next trial; the magnitude of the changes is assumed to be quite large, so that the experience of the most recent trials has a great effect on the trace of current trials.

These learning mechanisms combine to produce the kind of ASF effect discussed earlier: early CMs are driven to extinction after a few trials in which they are passed, slowing their associated responses, while later CMs remain relatively unchanged. Thus, trials with long FPs followed by ones with short FPs produce slowing, while the other sequences of long and short FPs produce less pronounced effects.

However, work exists which shows that the FPD effect cannot be entirely accounted for through the action of the ASF effect. Los and Agter (2005) found that the ASF effect contributes less than 20% to the FPD effect across a range of conditions. Vallesi et al. (2013) found that increasing the intertrial interval in a uniform FPD design significantly attenuates the sequential effect while leaving the overall FP curve relatively untouched, buttressing earlier work which also uncouples the FPD from the ASF effects (Vallesi & Shallice, 2007; Vallesi et al., 2007). Indeed, as we are about to see, there is a global effect of foreperiod distribution which is difficult to reconcile with the

essentially short range nature of trace conditioning.

Foreperiod distributions and sensitivity to hazard

The hazard function of a distribution $h(t)$ – the probability of an event occurring at t, given that it has not occurred by t – was introduced earlier in connection to response time distributions. The concept of hazard now returns in connection to the foreperiod distribution effect. Reminding ourselves of the results in figure 4.5(a), it can be seen that the FP curves in the various FPD conditions bear an inverse relationship to the hazard functions of those distributions.

To see this, consider performing an SRT with a uniform FPD; suppose that one knows the distribution of foreperiods and have a timer to keep track of the time elapsed on a given trial. As the foreperiod progresses, what should one expect regarding the probability of stimulus onset (SOP) as the critical moments pass? Continuing our four-CM design from above, evidently, when the trial starts, there is a 1/4 chance of the imperative moment occurring at the first CM. If the first CM passes without incident, however, the SOP for the next CM increases to 1/3. This process repeats such that, if one passes all the CMs except the last, one can be certain that the IM will occur at the next CM – that is, SOP increases to 1. In other words, with a uniform FPD, the hazard rate of the IM occurring at the next CM steadily increases throughout the trial. When the same logic is applied to apply to the exponential distribution, the hazard rate is constant; with the anti-exponential distribution, the hazard rate increases faster than with the uniform. Thus, there is a strong inverse relationship between mean RT and the hazard function of the FPD. The exponential case is particularly compelling, since the hazard function of the exponential distribution is by nature flat, and the FP curve is also.

Taken by itself, the hazard function of the FPD is a mere statistical fact. The fact that this relationships exists is given psychological theoretical content by positing that the process of bringing oneself into a state of preparedness to respond to RS onset is cognitively costly, while proper task performance – *vis-à-vis* quick responding – is rewarding. To optimise the trade-off between cognitive effort and task performance, preparedness is modulated by hazard: one is more ready to respond, and therefore responds quicker, when that readiness is more likely to result in a quick response rather than idle waiting Näätänen (1971), Vallesi and Shallice (2007).

As useful as hazard is in explaining the FPD effect, it suffers from a major empirical defect in that cannot explain the ASF effect, since it pertains only to performance over the timecourse of a single trial. Furthermore, hazard suffers from a major conceptual defect. For hazard to modulate preparedness, the FPD – as a global object over all FPS – must be known; but the only way for a subject to learn an FPD is through task performance, through exposure to particular instances. This raises the issue of how subjects are able to determine the FPD in order to be able to utilise hazard during task performance.

6. TWO PROBLEMS OF SYNTHESIS IN THE THEORY OF SIMPLE RESPONDING

Putting this all together, then, we see that the known facts about SRT performance require at least two major theoretical advances before the claim of that a relatively complete understanding of the cognitive processes underlying simple responding has been achieved.

The first is to reconcile the long-term and short-term effects in the foreperiod effects in a single cohesive framework, by finding a basic set of mechanisms can synthesise the FPD and ASF effects across the whole range of FPD conditions, preferably in a way that makes sense of the apparent effects of learning and attention imputed by the data. To this end, a number of proposals have already been advanced, including the dual-process model of Vallesi and Shallice (2007), which posits the existence of independent short-term, history-based mechanisms to account for the ASF effect and long-term, hazard-based mechanisms to account for the FPD; and the multiple trace theory of Los et al. (2014), which extends the trace theory of conditioning to include a memory component that could allow for the learning of the overall FPD, and thus for sensitivity to hazard.

However, none of the extant proposals are fully developed, and all are focused on explaining the relationship between foreperiod and *mean* RT. As we have also seen, though, the *distribution* of response times is an important part of the phenomenology of simple responding. Moreover, in a kind of symmetry to the fact that current models of foreperiod effects do not account for RT distribution, current models of RT distribution do not account for foreperiod effects. The second theoretical advance required is therefore to provide a synthetic account of both RT distribution and ASF effects.

We will not, alas, achieve this double synthesis in this work. In particular, this work will not focus on advancing theoretical models which can organically encompass the whole family of foreperiod effects. However, later chapters will provide synthetic models of RT distribution and FP effects, and it will be argued that the use of sophisticated techniques of Bayesian inference, such as MULTINEST can catalyse the scientific advances required to achieve a comprehensive theory of simple responding. We begin this argument in the next chapter.

—ω—

Part II

Investigations

Let every student of nature take this as a rule — that whatever his mind seizes and dwells upon with peculiar satisfaction, is to be held in suspicion.

— Francis Bacon, Novum Organon

CHAPTER 5

Demonstrating the basic viability of MultiNest

Abstract. A key conclusion of this thesis is that modern methods of Bayesian inference based on the concept of nested sampling, and the MultiNest algorithm in particular, ought to be adopted for broad use with cognitive models and data for both model comparison and parameter estimation. Two necessary (if not sufficient) criteria that must be met for this claim to be defended are, first, to prove that the proposed technique is capable of achieving the desired results – namely, producing accurate and precise estimates of models' Bayesian evidences and posterior parameter distributions – and, second, to show that it does so in a superior way to rival methods according to relevant desiderata. These two criteria, which we shall term *efficacy* and *competitiveness*, are chosen since they are the primary ones in the assessment of any practical tool: if the basic facts about these matters cannot be established, the usefulness of a proposed technique cannot be evaluated, and if a technique is found wanting in either respect, further discussion of it is rendered moot.

Previous studies have demonstrated the effectiveness of MultiNest when faced with a variety of numerically challenging problems; in this chapter, I provide an extension of and complement to these studies by offering the first report of tests of MultiNest when applied to a model where the posterior and Bayesian evidence are analytically tractable. This approach sets the technique against the only absolute benchmark available, and therefore strictly evaluates MultiNest's fundamental capacities. Through this undertaking, I demonstrate that MultiNest is capable of producing extremely accurate posterior and Bayesian evidence estimates, as well as useful error estimates. A second test assesses MultiNest's performance in a cognitive modelling context, attacking a previously studied toy problem in model selection involving the linear ballistic accumulator model. This second test finds MultiNest's estimates to be in general agreement with earlier ones. These findings supplement and support the already-convincing literature on MultiNest's efficacy.

In both these cases, I consider MultiNest's competitiveness by comparing its performance against rival methods of Bayesian inference for which the same tests have been performed; the discussion is tailored to the cognitive modelling community, so the comparison focuses on techniques already put forward in this field, especially thermodynamic integration and steppingstone methods. The assessment of competitiveness is done along a number of practical dimensions, such as computational cost, ease-of-use, and accuracy of estimates. This comparison finds that MultiNest offers numerous advantages over the rival techniques considered, and suffers from no serious drawbacks by comparison; in sum, it is argued that it is an accurate, efficient, general purpose, all-in-one and easy-to-use method for model inference. Taken together, these results establish the basic viability of MultiNest for use in cognitive modelling; as such, they provide the elementary justification for its use in later chapters and ground the extended demonstrations of the method's effectiveness and competitiveness therein.

1. INTRODUCTION · EFFICACY AND COMPETITIVENESS

Statistical models – that is, formal mathematical accounts which specify the precise probability of observing certain data – act as keystones in almost every analysis of the data of cognition. More, even: in many sub-fields of cognitive psychology, detailed statistical models constitute our current best theoretical understanding of the processes of cognition. Take, for example, the study of human decision making, where the theoretical framework *du jour* posits that decision-making is driven by the stochastic accumulation of a quantity, known as *sensory evidence*, which, upon reaching a threshold, triggers an irrevocable decision to respond (one way or the other, as the case may be). Ratcliff's drift-diffusion model (Ratcliff et al., 2016) and Brown and Heathcote's linear ballistic accumulator (Brown and Heathcote, 2008) model of choice-reaction tasks are canonical examples in this *evidence accumulator model* framework.

The success of such models in explaining various aspects of the observed data has stimulated the growth of the idea of *cognitive models* (see, for example, the book by Lee and Wagenmakers (2013) or the more recent one by Farrell and Lewandowsky (2018)). Cognitive models are statistical models of behavioural phenomena which do not aim merely to represent the data, but to provide precise mechanistic accounts of processes of mind such as "perception, memory, decision making, and other fundamental aspects of cognition" (Annis & Palmeri, 2018). This is a lofty and laudable aim; arguably, its pursuit is essential if we are to make progress in understanding the not-directly observable processes of cognition. However, whenever theoretical proposals can be formulated as statistical models – as with the cognitive models just mentioned – the scientific researcher must contend with two challenges endemic to all such models: *parameter estimation* and *model selection*. For cognitive models to fulfil their promise, then, adequate methods of surmounting these challenges are required.

Let us restate these challenges briefly (which were extensively covered in §2.3-2.4). The problem of parameter estimation is to find, for a given model, the values of its parameters which allow it to most closely represent the data to which it is being applied. The second challenge is deeper. While it is a basic activity of science to propose theoretical explanations of the data under study (in just the manner that has been illustrated), it is in turn a basic problem of science that data is never able to uniquely determine which of these explanations provides the best account of the data to hand (cf. appendix A). In general, and as has already been intimated with the examples given above, a number of plausible theoretical accounts can be constructed for any given data set. The challenge of model selection which follows is to determine the model from a pool of candidates that provides the best account of the data. Collectively, the solution of the problems of parameter estimation and model selection is referred to as *model inference*.

In response to these challenges, cognitive psychology has increasingly taken up the use of Bayesian methods of model inference (Annis and Palmeri, 2018; Vandekerckhove et al., 2015) amongst an upsurge in the use of Bayesian statistical methods in the field more generally (Andrews

and Baguley, 2013; Heck et al., 2020; van de Schoot et al., 2017; Vandekerckhove et al., 2018). The Bayesian approach provides a principled, unified, and coherent answer to the question of how to make reasonable inferences from data; it elegantly unifies the problems of parameter estimation and model comparison by casting them both as special cases of hypothesis evaluation and, in fact, they are seen to occupy adjacent levels in a hierarchy of hypotheses. The Bayesian framework was explained in detail in chapter 2, and is presumed to be understood here. For reference, a brief recap of key points pertinent to the present discussion is contained in the footnote.[i]

While the Bayesian account of inference is theoretically compelling, its application presents daunting practical obstacles. In the vast majority of cases, the posterior and the marginal likelihood (or Bayesian evidence) cannot be analytically computed; moreover, as covered in Chapter 3, estimating the posterior and (especially) the marginal likelihood are numerically difficult tasks. Concomitantly, recent work has proposed the use of numerous advanced methods for the estimation of marginal likelihoods for cognitive models. To give two more recent examples, Annis et al. (2019) have put forth thermodynamic integration (TI) and steppingstone (SS) methods as candidates for efficient Bayesian evidence estimation with cognitive models (see also the developments in Evans and Annis (2019)), and Gronau et al. (2017) have recommended bridge sampling (BS) methods, especially the Warp-III development thereof.

Now, in the context of the two tasks of model inference, these proposals rely on the use of an independent Markov Chain Monte Carlo (MCMC) method to sample the posterior, from which samples the Bayesian evidence is calculated.[ii] Since these approaches rely on MCMC posterior sampling, they inherit the weaknesses of the MCMC sampler used; these may include the need for a period of burn-in; high autocorrelation between samples, difficult-to-assess convergence; the need for multiple chains to establish validity, ensure mixing, and overcome "stuck" chains; and inefficiencies in the face of difficult posterior distribution morphologies. While many problems

---

[i] Parameter estimation is achieved by finding the so-called *posterior parameter distribution*. The posterior is a probability distribution over the model's parameter space, and it encodes the relative quality of various parameter values *qua* accounts of the data to which the model is being applied. The posterior at a given set of parameters θ is proportional to both (a) the likelihood of observing the data under the model at θ and (b) the *prior probability distribution* at θ, i.e., the posterior is proportional to the prior-weighted likelihood function (PWLF). The prior is the pre-application counterpart to the posterior, and it exists to takes into account any existing information about the relative quality of different parameter values.

The PWLF is not a proper probability distribution, since its integral does not in general come to unity. Integrating the PWLF over parameter space yields the correct normalising factor to obtain the posterior proper. This factor is called the *Bayesian evidence* or *marginal likelihood* of the models being considered. The ratio of Bayesian evidences of two models is known as the *Bayes factor*; it turns out that this Bayes factor, when multiplied by the prior probabilities over those same models (usually taken to be equal – this model prior is *not* the prior over the parameters), determines the relative quality of the models. The Bayes factor is often measured in terms of the so-called *deviance scale*, which is twice the natural logarithm of the Bayes factor proper. A common table for the qualitative interpretation of Bayes factor is given by Kass and Raftery, 1995 and reproduced in §2.4:2.

We note that, for the purposes of posterior estimation, the PWLF, as the unnormalised posterior, contains all the same information as the fully normalised posterior; for Bayesian model selection, though, the normalisation calculation is critical.

[ii] For the unfamiliar, a brief introduction to Markov Chain Monte Carlo sampling is provided in appendix E.

may be solved on an *ad hoc* basis for a particular problem – say, by choosing a different MCMC sampler – these techniques readily require the setting of a number of highly technical operational parameters at both the MCMC sampling and evidence-evaluation levels. Furthermore, evidence evaluation is typically an order of magnitude more costly than posterior parameter inference. Taken together, these features make existing proposed techniques of Bayesian model inference technically demanding.

An alternative approach to Bayesian inference is that of *nested sampling*, originally due to Skilling (Skilling, 2004, 2006) and fully described in chapter 3. This approach is not based on MCMC sampling, and, in its modern developments such as MULTINEST, has a number of attractive features: it provides simultaneous inferences for both model selection and parameter estimation (that is, it returns both the Bayesian evidence and posterior estimates of a model applied to data); it requires no special tuning of complex sets of sampling parameters; and it efficiently and automatically deals with more or less arbitrary posterior distribution morphologies. Indeed, in the fields of cosmology and particle physics within which it was developed, MULTINEST has afforded multiple-order-of-magnitude increases in the efficiency of marginal likelihood estimation compared to thermodynamic integration methods.

Despite its promise, nested sampling has in general been overlooked in the cognitive modelling literature: Annis et al. (2019) mention NS, but dismiss it on the authority of certain studies (Friel and Wyse, 2012; Liu et al., 2016) which claim that NS is less efficient than certain alternatives. However, the studies upon which this dismissal of NS techniques is based all relied upon the now (and then) outdated implementation of nested sampling due to Skilling. The later developments in NS covered in Chapter 3 significantly (conservatively, a hundred-fold) improve on the efficiency and accuracy of the basic NS technique (Feroz et al., 2009).

The aim of this chapter is to begin the demonstrative argument, which will extend throughout the rest of this thesis, that cognitive science could benefit from adopting advanced nested sampling techniques. The touchstones of the analysis here will be *efficacy* and *competitiveness*. "Efficacy" is taken to mean the intrinsic ability of a technique to properly perform its given function – so, in the case of the model selection techniques considered herein, to accurately and precisely evaluate the Bayesian evidence. "Competitiveness" means the assessment of the advantages and disadvantages of one technique compared to another when referred to chosen practical desiderata. Simply put, the aim is to demonstrate whether the tool does its job well, and how simple it is to use.

To achieve this aim, two simulation studies based on prior literature will be carried out to test MULTINEST's capabilities. For the first of these, we turn to the study conducted by Xie et al. (2011) which introduced the steppingstone (SS) estimator Annis et al. advocate. (The results of this study, in combination with reviews cited above, appears to form part of the basis of their advocacy.) Xie et al.'s study, which also considers the method of thermodynamic integration (TI), tests the

ability of these methods to produce accurate and precise estimates of the marginal likelihood by comparing the results obtained through the application of the estimators to a situation in which an analytical expression for the marginal likelihood is available. In the work of §5.2, I run the assessments presented by Xie et al. (2011) on MultiNest, with extensions to the depth and rigour of the analysis performed; further, owing to MultiNest's broader range of capabilities as regards posterior estimation and the provision of error estimates, the breadth of the test will be expanded to include these concerns.

The results of the assessments show that MultiNest is capable of producing highly accurate and precise estimates of the marginal likelihood and posterior distribution, as well as useful one-shot estimates of the error in the log-evidence (at least in the test case considered). The results further show that the marginal likelihood estimates obtained are of the same degree of accuracy and precision as those obtained by the TI and SS estimators, but at increased efficiency. Furthermore, and importantly, since these tests involve comparing the results obtained via MultiNest to analytically available benchmarks, these tests establish the basic validity of the technique; this study constitutes the first report of this basic result.

The second test case applies MultiNest to solve a problem of model selection involving two versions of the aforementioned linear ballistic accumulator (LBA) model, extending the . The LBA and this particular model selection problem have been used in numerous previous studies on estimating marginal likelihoods in cognitive models (Annis et al., 2019; Evans and Annis, 2019; Evans and Brown, 2018), studies use TI and SS methods, allowing for clear comparisons to this previous work. These results from this second test show that MultiNest produces comparable outputs to the ML estimation methods studied in earlier research, while offering a number of technical advantages.

In conjunction with broader considerations of the theory and practice of the various techniques inspected, the results outlined above are used to argue for MultiNest's potential usefulness to cognitive science; these arguments will be reinforced in later chapters by the theoretical cognitive-psychological discoveries made through the application of the approach.

2. FIRST TEST · A SIMPLE GAUSSIAN LIKELIHOOD

2:1 The justification for and aims of the test

The tests of MultiNest presented herein will be based upon one performed by Xie et al. (2011), who introduced the steppingstone algorithm for marginal likelihood (ML) estimation.[iii] The test case was a model consisting of a simple Gaussian likelihood with a normal prior over the mean

[iii] A reminder that the marginal likelihood is also referred to as the "model evidence", the "Bayesian evidence", or – where context makes the meaning clear – simply the "evidence". This chapter will employ this menagerie of terms freely and interchangeably.

parameter (with the variance assumed to be known), applied to a Gaussian-generated dataset with $n = 100$ data. This case was chosen because a model (*viz.* likelihood and prior) of this type is susceptible to the exact analysis of the ML – and, incidentally, the posterior – and the purpose of the paper was, in the authors' own words, "above all about estimating the numerical quantity known as marginal likelihood with accuracy", the logic being that methods which do not accurately estimate the evidence cannot be relied upon for the subsequent task of model comparison. Further details of the procedure as carried out by Xie et al. will be covered in subsection §5.2:2 below, alongside the details of the procedure the author carried out.

This exemplar is adopted as the the test case for MultiNest for two reasons. First, following the logic of Xie et al. outlined above, the presence of an exact expression for the ML allows for the fundamental accuracy and precision (AP) of MultiNest to be ascertained, and if MultiNest promises to give accurate and precise measures of the ML for complex models, it *must* be able to do so to a high degree in one of the simplest possible cases. (While the author does not doubt that such tests have been performed, to his knowledge, no such basic test has been reported previously.) Thus, in the language of the introduction, this case allows for the most direct test possible of the fundamental efficacy of MultiNest *qua* estimator of Bayesian evidence. The second reason for following Xie et al. is that the TI and SS methods tested in their papers have been prominently proposed for use in cognitive science (Annis et al., 2019; Evans and Annis, 2019); rerunning the same tests with MultiNest will allow for constructive comparisons to be made between alternative techniques, shedding light on the question of competitiveness.

The analysis as originally undertaken was designed solely to test the AP of the ML estimates provided by the SS and other methods. On account of certain capabilities of MultiNest not shared by other ML estimation approaches, and in the spirit of providing a comprehensive test of its efficacy, the scope of the test undertaken here will be greater: with two further assessments performed. The first extension will be to assess the AP of the log-evidence uncertainty estimates provided by MultiNest. (As the reader may recall from §3.3, ordinary MCMC methods require multiple runs to establish the uncertainty in the obtained ML estimates, whereas MultiNest provides an uncertainty estimate with each run.) Also, unlike SS and the others, MultiNest does not rely on an independent MCMC method to sample the posterior, but automatically provides posterior samples through its own nested sampling approach. The second extension will thus be to establish the accuracy of the posterior estimate and parameter inferences obtained through those samples. Taken together, these three assessments constitute a test of MultiNest's efficacy across all of its basic functions, and allow for a comprehensive view of its competitiveness relative to TI and SS methods.

2:2 General methods

Fit data, model, and exact solution

The dataset $\mathbf{y} = (y_1, y_2, \dots, y_n)^{\mathrm{T}}$, which we shall designate the *fit data*, at the basis of the test consists of $n = 100$ samples drawn from a normal distribution with zero mean and unit standard deviation. The same dataset as Xie et al. was used, recovered from the location given by them in their paper.

To model the fit data, a Gaussian likelihood function with mean parameter μ and standard deviation parameter τ is chosen, so that the likelihood of the data is given by

$$\mathcal{L}(\mathbf{y} \,|\, \mu, \tau) = \prod_{i=1}^{n} \phi\left(\frac{y_i - \mu}{\tau}\right) \tag{5.1}$$

where $\phi(\cdot)$ is the standard normal PDF. An inflexible prior distribution is chosen over τ, with the prior mass concentrated at $\tau = 1$ – that is, the standard deviation is treated as a known quantity. The prior distribution over μ is selected to be a normal distribution with its mean hyperparameter set to $\xi = 0$ and standard deviation hyperparameter to $\eta = 1$.

It is a standard result (DeGroot, 1970) that, for the model (likelihood function and prior) given above, the Bayesian evidence of data $\mathbf{y}$ is given by

$$\mathcal{Z}(\mathbf{y}) = \left(2\pi n \sigma^2\right)^{-\frac{n}{2}} \left(\frac{\eta^2}{\sigma^2} + 1\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\left[\frac{s^2 + \overline{y}^2}{\sigma^2} + \frac{\xi^2}{\eta^2} - \frac{\left(\frac{\overline{y}}{\sigma^2} + \frac{\xi}{\eta^2}\right)^2}{\left(\frac{1}{\sigma^2} + \frac{1}{\eta^2}\right)} \right] \right\} \tag{5.2}$$

where

$$\sigma^2 = \frac{\tau^2}{n}, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \text{and} \quad s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right). \tag{5.3}$$

The above formulae are used to calculate the exact value of the ML of the Gaussian model applied to the fit data.

Given prior hyperparameters ξ and η, the posterior distribution over μ after observing fit data $\mathbf{y}$ is normally distributed (still) with updated posterior hyperparameters

$$\begin{aligned} \xi_{\text{post}} &= \left(\frac{n}{\tau^2} + \frac{1}{\eta^2}\right)^{-1} \left(\frac{\xi}{\eta^2} + \frac{\sum_{i=1}^{n} y_i}{\tau^2}\right), \qquad \text{and} \\ \eta_{\text{post}} &= \left(\frac{n}{\tau^2} + \frac{1}{\eta^2}\right)^{\frac{1}{2}}. \end{aligned} \tag{5.4}$$

Generating the analysis data: The study of Xie et al.

Having calculated the exact ML, Xie et al. then went on to apply the HM, TI and SS methods to estimate the same quantity, from which estimates the analysis of the accuracy and precision of each method was performed. We deal first with the handling of the latter two methods, which requires a brief detour into certain technicalities; a similar exposition is then repeated for the HM estimator.

The TI and SS methods (which we shall jointly call the "thermodynamic methods") rely on so-called *power posteriors*. In general, for a model with parameters θ applied to data $\mathbf{x}$, a power posterior is defined as

$$\mathcal{P}_\beta(\theta \mid \mathbf{x}) = \frac{\mathcal{L}(\mathbf{x} \mid \theta)^\beta \, \pi(\theta)}{p(\theta)} \tag{5.5}$$

where $\mathcal{L}(\cdot)$ is the likelihood function of the model, $\pi(\cdot)$ is its prior distribution, and β is the so-called *temperature*. The temperature varies between 0 and 1; it is a scaling power for the likelihood component of the posterior function, so that the $\beta = 0$ power posterior corresponds to the pure prior function, and the $\beta = 1$ power posterior corresponds to the pure posterior function.

Thermodynamic estimators of the marginal likelihood require samples to be taken from power posteriors at various temperatures. More specifically, the temperature scale is split into K partitions (or *rungs*); power posterior samples are then to be collected at the $K + 1$ values of β (at the top and bottom of each rung) for the TI method and K values of β (at the bottom of each rung) for the SS method. The power posterior samples so collected are then used to estimate the ML according to certain formulae (see Xie et al. (2011) for further details).

Note that the rungs need not be of equal width, that is, the β values bounding each rung need not be equally spaced. The number and spacing of rungs chosen are together called the *temperature schedule*. In general, the choice of temperature schedule governs the sampling procedure, affecting the samples collected and thereby the resulting ML estimate.

For each of the thermodynamic methods, Xie et al. considered four different temperature schedules, arising from the combination of two different K values ($K = 50$ and $K = 100$) and two different spacings of β ($K + 1$ values spaced uniformly from 0 to 1, and $K + 1$ values spaced at uniform quantiles of $\mathrm{Beta}(0.3, 1)$). At each rung, 2000 power posterior samples were collected; since the general form of the power posterior is analytically known in this case, the samples were drawn directly from the distribution. For the $K = 50$ conditions, this resulted in 102,000 (power) posterior samples drawn for the TI method and 100,000 for the SS; for $K = 100$, there were 202,000 and 200,000 respectively.

By contrast, the harmonic mean estimator does not require power posteriors, only draws from the full posterior; a simple formula exists to estimate the ML from these samples (Newton & Raftery, 1994). Since the number of posterior samples fundamentally affects the convergence behaviour and precision of all these estimators, and to allow for parity in the comparisons between

the thermodynamic methods and the HM, there were two conditions under which the HM was calculated, one using 102,000 posterior samples, the other 202,000 (that is, equivalent to the number used for the $K = 50$ and $K = 100$ conditions used with TI). Since HM does not entail the temperature scheduling attendant to the use of power posteriors, these were the only two conditions considered for this method.

A quick word on terminology: for the sake of clarity and consistency, and in line with the term "temperature schedule," we will henceforth refer to a particular set of sampling parameters (*viz.* parameters which control an estimator's behaviour: number and spacing of rungs, number of samples collected, etc.) as a *sampling schedule* or simply a *schedule*. Thus, in total, Xie et al. considered ten method-schedule conditions, four schedule conditions each for the thermodynamic methods, and two for the HM method.

Finally, since each estimator, by virtue of the randomness of the MCMC samples on which it relies, returns a somewhat different result each time it is used (under the same schedule), multiple replications are of course required to establish the AP of a method. To this end, Xie et al. ran 1000 replications for each method-schedule condition. The sets of 1000 samples of (log-)ML estimates obtained under each method-schedule condition form the basis of the analysis of the AP of each method, so we will refer to these sets of samples, both individually and collectively, as the *analysis data*.

Generating the analysis data: Present study

MultiNest's principal sampling parameters are the number of live points used in sampling (N_{live}) and the tolerance (c_{tol}), which fixes the stopping criterion. (Refer to §3.3 for details.) To assess MultiNest's performance and allow for constructive comparisons to the thermodynamic methods described above, six different sampling schedules are considered, with approximately 1000 independent replications performed per schedule (bar for one). The basic ideas behind and motivations for the schedules are set out below; for details of the precise sampling parameters used and the number of replications performed for each schedule, refer to table 5.1 and its accompanying text.

Before the schedules are laid out, we note that two kinds of data are collected for each schedule: analysis data (analogously to the use of the term above for the data collected by Xie et al.), consisting of the log-marginal likelihood estimate (LMLE), "autogenous" log-marginal likelihood error estimate (ALEE, cf. §5.3:2 for the terminology), and posterior samples (p^*) for each run; and *performance data*, consisting of the number of posterior samples collected and the number of likelihood evaluations performed on each run. The analysis data is used in the three assessments described below to establish the efficacy of MultiNest. The performance data is presented alongside the sampling parameters in table 5.1 and will be attended to again in the discussion, where, in conjunction with the analysis data, it will be used to assess the

Sampling schedule	N_{live}	c_{tol}	N_{runs}	$\overline{N}_{\text{post}}$	$\overline{N}_{\text{like}}$
Standard sampling	1000	0.5	995	4400 ± 42	10 747 ± 254
High precision	80 000	0.001	228	623 123 ± 716	1 758 784 ± 8060
Mimic $K = 100$, Beta	32 400	0.05	911	205 921 ± 239	556 449 ± 3550
Slack $K = 100$, Beta	32 400	0.5	1099	142 550 ± 254	345 132 ± 1434
Mimic $K = 50$, Beta	16 000	0.05	1152	101 689 ± 269	274 711 ± 2459
Slack $K = 50$, Beta	16 000	0.5	1025	70 405 ± 180	170 402 ± 1079

Table 5.1. SAMPLING PARAMETERS AND PERFORMANCE DATA PER SCHEDULE UNDERTAKEN FOR MULTINEST. The first column gives the designation of the sampling schedule; the second, the number of live points used in the schedule; the third, the value of the c_{tol} tolerance parameter used by MULTINEST to determine its stopping criterion. The fourth column lists the number of replications performed per schedule; the outputs of these runs constitute the analysis data, as described in the text. The last two columns comprise the performance data. The fifth column gives the average number of posterior samples output, collapsed across all the runs in a schedule, alongside the standard deviation uncertainty; the sixth column gives similar information for the total number of likelihood evaluations performed (the difference being due to rejections of sampled points by the MULTINEST algorithm). Note that the values in the last two columns are rounded to the nearest whole number. The technical meaning of the sampling parameters was covered in §3.3

.

competitiveness of MULTINEST compared to thermodynamic methods.

The first schedule undertaken will be referred to as the "standard sampling" schedule, using the default values for N_{live} (1000) and c_{tol} (0.5). This output of this schedule will indicate what ordinary performance of the MULTINEST algorithm looks like.

The second "high-precision" schedule is designed to exercise an extreme test of MULTINEST's capabilities. By setting the number of live points very high (80 000), and the tolerance very low (1×10^{-3}), the log-evidence estimates returned will have extremely high precision, and the number of posterior samples collected will be exceptionally large. Any deviation from optimal performance will therefore be obvious in the resultant analysis data, whether as bias in the value of the log-evidence towards which the sampling procedure is converging, or divergence in the distribution of posterior samples from the true posterior.

Two further schedules, the "mimic" schedules, have sampling parameters tuned to closely imitate the performance – measured by the number of posterior samples and the SE obtained – of the steppingstone Beta-spaced $K = 50$ and $K = 100$ schedules respectively ($N_{\text{live}} = 16\,000, 32\,400$ and $c_{\text{tol}} = 0.05$). Sampling under these schedules is undertaken to demonstrate that, using the number of posterior samples as a metric of efficiency, MULTINEST is capable of performing as efficaciously as the SS method.

A brief word on how the sampling parameters for the mimic schedules were set. The uncertainty estimate, or ALEE, calculated by MULTINEST at the end of each run is equal to $\sqrt{H/N_{\text{live}}}$, where H is a quantity known as "relative entropy" (Feroz, 2008). Now, this uncertainty value is an estimate of the standard deviation – that is, the standard error (SE) – of the LMLES

that would be obtained across repeated runs. We can therefore set the number of live points required to mimic the performance of the SS method *vis-à-vis* the SE of the obtained LMLEs as $N_{\text{live}} = \left(\sqrt{H}/\text{SE}\right)^2$. We have the desired SE from the results of the SS method from Xie et al., and use the output of the standard sampling schedule to estimate H. The tolerance parameter c_{tol} was set by trial and error until the number of posterior samples obtained roughly matched those used by Xie et al.

The final two schedules, the "slack" schedules, retain the same number of live points as the mimic schedules, but increase the tolerance (to 0.5), leading to fewer iterations at the tail end of the run as the algorithm assumes convergence sooner. These schedules can therefore help to establish whether MULTINEST is capable of *better* performance than the thermodynamic methods of Xie et al., in the sense of providing ML estimates of comparable accuracy and precision at lower computational cost.

2:3 Assessment A: Accuracy and precision of Bayesian evidence estimate

Methods

PREAMBLE. The purpose of the first assessment is to establish the AP of the log-marginal likelihood estimates returned by MULTINEST and to provide data pertinent to comparing its performance to that of the alternative estimators. We begin by noting that every log-ML estimate (the LMLEs) returned by an estimator method operating under a particular schedule[iv] can be viewed as an IID sample from a distribution, the method-schedule's full log-ML estimate distribution, which would be obtained in the limit of the number of LMLE samples tending to infinity. We also remark that accuracy is to be understood as a function of the bias of the method – namely, the average difference of the LMLEs from the true value of the log-evidence – and precision is to be understood as a function of the spread of the LMLEs. More pointedly, the accuracy and precision are directly related to the mean and variance of the distribution of the errors in the LMLEs.

To quantify the AP for the SS, TI, and HM methods, Xie et al. calculated the mean across the 1000 LMLE samples for each method-schedule condition, as well as the standard error (SE) and root mean square error (RMSE). Denoting the set of LMLEs from a given method-schedule condition as $\mathcal{Z}_{\text{est}}$, and the true (analytical) log-marginal likelihood as $\mathcal{Z}_0$, the SE and the RMSE are given by

$$\text{SE} = \sqrt{\text{Var}\left(\mathcal{Z}_{\text{est}}\right)}, \quad \text{and} \quad \text{RMSE} = \sqrt{\text{Var}\left(\mathcal{Z}_{\text{est}}\right) + \left(\overline{\mathcal{Z}}_{\text{est}} - \mathcal{Z}_0\right)^2}. \tag{5.6}$$

For the RMSE, the first term under the square root is error which can be attributed to variance in the sampled LMLEs; the second term, error attributable to bias in the LMLEs. Therefore accurate – namely, unbiased – estimators will have RMSE equal to SE, and precise estimators will have low SE.

[iv] That is, the estimates obtained by repeatedly applying a given method to the same model and the same data with the same sampling parameters.

A short *intermezzo* is warranted before the full suite of methods applied to the MULTINEST analysis data is covered. While the RMSE and SE of the analysis data jointly quantify the accuracy and precision of the method which gave rise to that analysis data, in principle, two further factors complicate matters. First, since they are calculated from finite samples, the RMSE and SE are themselves but *estimates* of the true values of those quantities (which theoretically are those derived from the method-schedule's full log-ML estimate distribution). It follows that, whatever the means by which accuracy and precision are quantified from the sampled LMLEs, the resultant numbers are themselves subject to uncertainty. As ever with estimated quantities, properly assessing the import of the calculated quantities of accuracy and precision requires taking this uncertainty into account.

The second matter concerns the accuracy estimate. While the object with regards to the precision is simply to quantify it (and its uncertainty), with regards to the accuracy, we also wish to judge whether the analysis data gives evidence of bias in the method. Simply stating the RMSE and SE, while certainly indicative (and potentially *prima facie* convincing) for the question of whether bias exists, does not *per se* answer it. Rather, to establish the answer, it is necessary to rigorously interpret the obtained estimates using the tools of statistical inference.

Since the aim here is to provide a rigorous demonstration of MULTINEST's capacities, it should be understood that the methods presented below are performed with these stringent ends in mind. (This concludes the *intermezzo*.)

A1. BASIC DESCRIPTIVE STATISTICS. To assess the accuracy of MULTINEST's ML estimation process, the same statistics that Xie et al. calculated from their analysis data are calculated from the analysis data obtained from MULTINEST under each schedule condition. To round out the basic descriptive statistics pertinent to the matter at hand, the means and ranges of the absolute errors in the LMLEs from each schedule condition are also calculated, since these provide useful alternative characterisations of the accuracy and the precision of the estimates.[v] The standard error of the mean (SEM) quantifies the uncertainty in the accuracy estimate, so is also calculated for each schedule's LMLEs. These basic descriptive statistics are presented in table 5.2, alongside the data from Xie et al.'s original work. Supplementing this, histograms of the LMLEs returned by MULTINEST are shown in figure 5.1.

A2. FULL POSTERIOR DESCRIPTIONS OF ACCURACY AND PRECISION. To provide a more sophisticated quantification of the accuracy and precision of MULTINEST *qua* ML estimator, we return to the notion that characterising the AP of any such estimator is, from the mathematical point of view, a problem of estimating the distribution of the errors derived from the LMLEs. Now, it happens to be that every set of LMLE samples obtained from MULTINEST (across the schedules used here) can be treated as normally distributed, as established by Kolmogorov-Smirnoff tests (p-values

[v] The mean of the absolute errors is equivalent to the the absolute error of the mean in the ML estimate.

reported in figure 5.1). Since the LMLE distributions can be straightforwardly transformed to the absolute error distributions by subtracting the true value of the log-evidence, the problem of quantifying MULTINEST's AP, and the uncertainties in the AP, is reduced to that of estimating the mean and standard deviation parameters of the absolute error distribution as modelled by a normal likelihood. This is a straightforward problem, then, of posterior parameter inference.

To perform the inference, we start by transforming the analysis data to express the errors in the ML estimates as already mentioned; that is, for each method-schedule condition, the transformed analysis data $\epsilon = (\epsilon_1, \epsilon_1, \ldots, \epsilon_n) = \mathcal{Z}_{\text{est}} - \mathcal{Z}_0$ is formed, where n is the number of samples collected under the schedule. We posit that the transformed analysis data obeys a normal likelihood with mean parameter μ_ϵ (which corresponds to the bias) and standard deviation parameter σ_ϵ (which relates to the precision). Now, it is generally accepted that for a normal likelihood, an uninformative prior over the mean is uniform over the entire real line, and an uninformative prior over the standard deviation is proportional to the inverse of the same.[vi]

For a normal likelihood with these uninformative priors, it can be shown that the marginal posterior distribution over μ_ϵ takes the form of a location-scale Student's t-distribution with degrees-of-freedom hyperparameter $\nu = n - 1$, location hyperparameter $\lambda = \overline{\epsilon}$, and scale hyperparameter $\xi_t = s/\sqrt{n}$, where $s^2 = \frac{1}{n-1}\sum_i (\epsilon_i - \overline{\epsilon})^2$. The marginal posterior distribution over σ_ϵ^2 (note the square) is a scaled inverse chi-squared distribution with degrees-of-freedom hyperparameter $\nu_\chi = \nu = n - 1$ and scale hyperparameter $\xi_\chi = s^2$ (Gelman et al., 2013).

Note that this model essentially treats the mean absolute error and SE statistics reported in table 5.2 as random variables that need characterisation. The marginal posterior distributions for the accuracy and precision of the ML estimates for each schedule are shown in figures 5.2 and 5.3.

A3. ANALYSES FOR EVIDENCE OF BIAS. So much for merely *quantifying* the accuracy and precision. The more important question with respect to the fundamental aims of this study is whether the ML estimates returned by MULTINEST exhibit evidence of bias, namely, inaccuracy. This is a problem of hypothesis testing or model comparison, pitting the the null hypothesis that there is no bias against that alternative that there is. We follow three different hypothesis comparison protocols to test these hypotheses, make the desired inference, and check for consistency. Note that, across all the hypothesis comparison protocols, the formation of the hypotheses will postulate that the likelihood which models the error distribution is normal, as done above in sub-assessment A2,

The first hypothesis comparison protocol used is a posterior ratio test, based on the marginal posteriors obtained in the manner described for sub-assessment A2 and displayed in figure 5.2; for each schedule, the null hypothesis tested is that $\mu_\epsilon = 0$, versus the alternative the $\mu_\epsilon = \max_{\mu_\epsilon}\{p(\mu_\epsilon \mid \epsilon)\}$. The second protocol is a likelihood ratio test of the same null hypothesis versus

[vi] Or equivalently, the prior over the standard deviation in uniform on the log-scale. Such priors are improper – they do not integrate to unity as a proper probability distribution must. This impropriety does not in general cause problems, and that it does not will be taken for granted here.

the alternative that $\mu_\epsilon = \max_{\mu_\epsilon}\{\mathcal{L}(\epsilon \mid \mu_\epsilon)\}$. The third is a "Bayesian t-test" of the null versus the alternative that $\mu_\epsilon \neq 0$ (Rouder et al., 2009).

The posterior and likelihood ratio tests are chosen since they provide information regarding the relative probability of (a) the hypothesis that the distribution of LMLEs provided by MULTINEST under a given schedule condition is centred on the true log-evidence value against (b) the hypothesis that the LMLE distributions is centred on the most probable value given the observed LMLEs.[vii] The likelihood ratio test presumes that the standard deviation of the population distribution is equal to that of the observed sampling distribution; the posterior ratio test, through the prior over σ_ϵ, takes into account the uncertainty in this parameter. In particular, since the prior distribution over $\sigma_\epsilon \propto 1/\sigma_\epsilon$, the posterior ratio test takes into consideration that the true σ_ϵ is much smaller than that observed, thus giving higher prior weight to the notion that MULTINEST's precision is even higher than might be expected, and therefore that the observed absolute mean error is in fact due to bias and not just sampling variance. In this manner, the posterior ratio test represents prior scepticism of a lack of bias in MULTINEST, and is a stringent test of the accuracy of the method.

The likelihood ratio and posterior ratio protocols only test *point* hypotheses regarding the bias – that is, they only compare the hypothesis that the bias μ_ϵ is equal to one specific value over another, and do not test whether there is evidence that MULTINEST is biased in any way whatsoever (that is, comparing $\mu_\epsilon = 0$ against the whole range of hypotheses $\mu_\epsilon \neq 0$). This possibility is covered by the Bayesian t-test. The technical details of this test will not be covered here; see Rouder et al. (2009) for details. The crucial point for our present purpose is that the test returns the Bayes factor of the null over the alternative for the assumed likelihood and with an uninformative prior.

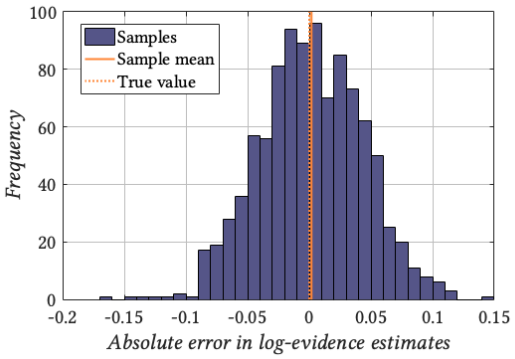
Results and discussion

A1. BASIC DESCRIPTIVE STATISTICS. Table 5.2 presents the basic descriptive statistics for all the method-schedule conditions tested, including those performed by Xie et al. Figure 5.1 shows the histograms of the estimates obtained from each sample. It seems reasonable to summarise the findings contained therein thus: the ML estimates returned by MULTINEST in this test case are accurate, precise, and reliable. The claims of this summary statement are justified with respect to the basic analysis results directly below.

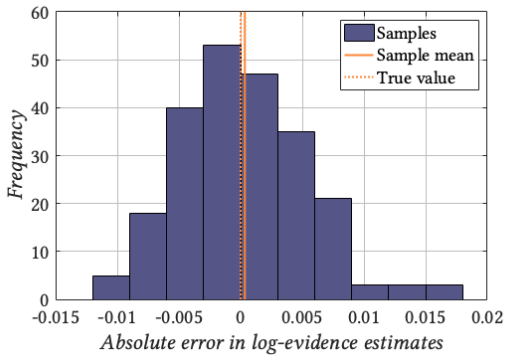
First, supporting the claim that MULTINEST is capable of providing highly accurate estimates of the model log-evidence: Apart from the standard sampling schedule, all the other schedules provide average estimates correct to within less than one part in a thousand in units of log-evidence. Even with the standard sampling schedule, which collected many fewer posterior samples than its counterparts, the average log-evidence estimate is correct to within two parts in a hundred in log-evidence units. Since meaningful differences in log-evidence between models are

[vii] That is, the mean value of the LMLEs, since the prior is flat in the posterior ratio case.

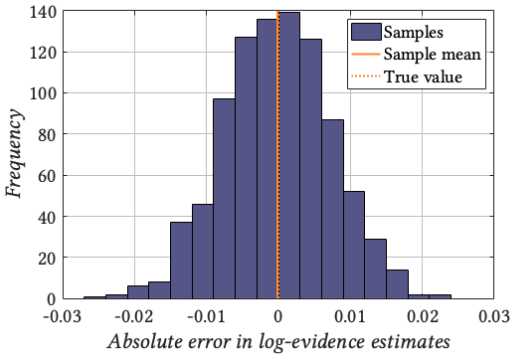
(a) Standard sampling ($p = 0.815$)



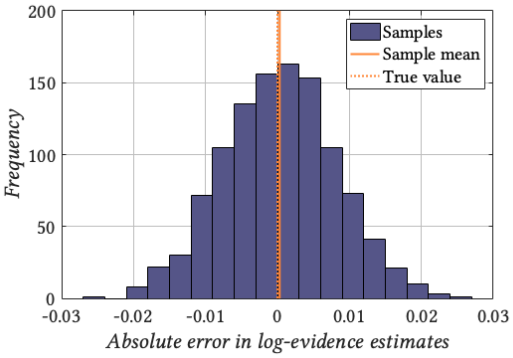
(b) High precision ($p = 0.858$)



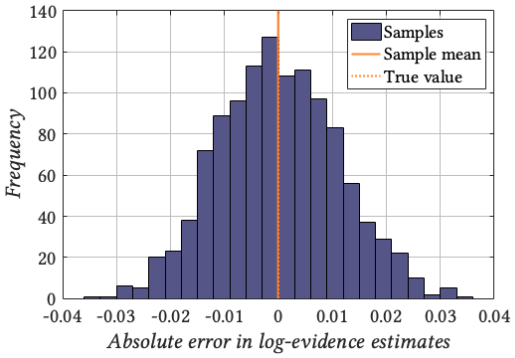
(c) Mimic $K = 100$ Beta ($p = 0.998$)



(d) Slack $K = 100$ Beta ($p = 0.980$)



(e) Mimic $K = 50$ Beta ($p = 0.622$)



(f) Slack $K = 50$ Beta ($p = 0.381$)

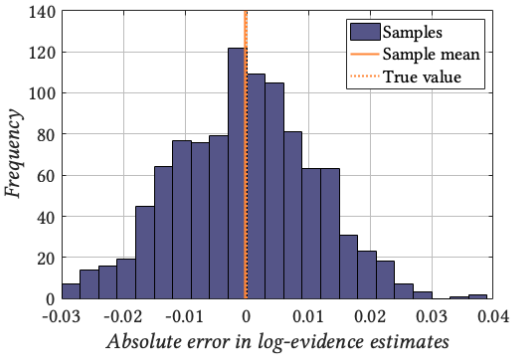


Figure 5.1. HISTOGRAMS OF LOG-EVIDENCE ESTIMATES RETURNED BY MULTINEST. Each histogram shows the estimates obtained across 1000 repeated runs of MULTINEST running the stated sampling schedule. The values along the x-axes are given as absolute errors from the analytical log-evidence, in log-evidence units. The solid vertical line shows the mean log-evidence estimate (across all 1000 runs) and the dotted vertical line picks out zero-error value. Note the differences in scale along the axes for each schedule. The labels also include the p-value obtained by applying a Kolmogorov-Smirnoff test for the normality of the sample (mean and standard deviation set by the sample). For further information regarding the definition of each schedule and the exact number of samples collected, refer to table 5.1.

Method	*Sampling schedule*	*Mean* ML	*Absolute error*				RMSE	SE	SEM
			Mean	*Minimum*	*Maximum*	*Range*			
Exact	-	−147.9161	-	-	-	-	-	-	-
Harmonic mean	$n = 202\,000$	−147.020	−0.896	-	-	-	0.9479	0.3081	0.00974
	$n = 102\,000$	−146.991	−0.925	-	-	-	0.9979	0.3179	0.01005
Thermodynamic integration	$K = 100$, Beta	−147.917	0.001	-	-	-	0.0079	0.0078	0.00025
	$K = 100$, Uniform	−147.955	0.039	-	-	-	0.0413	0.0146	0.00046
	$K = 50$, Beta	−147.921	0.005	-	-	-	0.0123	0.0115	0.00036
	$K = 50$, Uniform	−148.053	0.137	-	-	-	0.1384	0.0218	0.00069
Steppingstone	$K = 100$, Beta	−147.916	0.000	-	-	-	0.0074	0.0074	0.00023
	$K = 100$, Uniform	−147.916	0.000	-	-	-	0.0135	0.0135	0.00043
	$K = 50$, Beta	−147.916	0.000	-	-	-	0.0105	0.0105	0.00033
	$K = 50$, Uniform	−147.916	0.000	-	-	-	0.0162	0.0162	0.00051
MULTINEST	Standard sampling	−147.9144	0.0018	−0.1627	0.1456	0.3082	0.0424	0.0424	0.00134
	High precision	−147.9158	0.0003	−0.0119	0.0163	0.0282	0.0051	0.0051	0.00034
	Mimic $K = 100$, Beta	−147.9163	−0.0001	−0.0260	0.0224	0.0484	0.0074	0.0074	0.00024
	Slack $K = 100$, Beta	−147.9158	0.0003	−0.0261	0.0244	0.0504	0.0079	0.0078	0.00024
	Mimic $K = 50$, Beta	−147.9162	−0.0001	−0.0357	0.0334	0.0691	0.0112	0.0112	0.00033
	Slack $K = 50$, Beta	−147.9164	−0.0003	−0.0296	0.0380	0.0677	0.0113	0.0113	0.00035

Table 5.2. BASIC DESCRIPTIVE STATISTICS OF LOG-EVIDENCE ESTIMATE SAMPLES FOR EACH METHOD-SCHEDULE CONDITION. The mean Bayesian log-evidence, RMSE, and SE for the HM, TI, and SS methods are excerpted from Xie et al. (2011). The mean absolute error of the samples for each schedule of those methods has been straightforwardly calculated by taking the difference of the reported mean ML with the true log-evidence. Similarly, for those methods, the SEM was calculated as the SE divided through by $\sqrt{1000}$, for the reported 1000 sampling replications. Any seeming discrepancies are due to rounding errors. The results of the MULTINEST schedule assessments are reported to four decimal places, compared to three for the excerpted data, to maintain consistency with the RMSE and SE values (which are in the same units). Note that the mean absolute errors are stated to higher precision in table 5.3. All values given in units of log-evidence. Cells with a dash represent unavailable or insensible data. For descriptions of the sampling parameters applied in each schedule, refer to the text (§5.2:2) and table **??**; for further information about the characteristics of the MULTINEST samples obtained, refer to **??**.

on the order of one unit of log-evidence (cf. table 2.1), this is well within the bounds of usefulness. Moreover, there is – at least at first glance – no bias in these estimates, as seen by the equality of the RMSE and SE of the samples; this conclusion *viz.* lack of bias is further supported by the fact that no mean log-evidence estimate falls more than about 1.25 SEM units off the true value.

Second, that the estimates provided in this case are precise: The SE measures the standard deviation uncertainty, and so the precision, of the estimator; the table shows that this uncertainty is on the order of four-hundredths of a unit of log-evidence for the standard sampling schedule, and a five- to ten-thousandths of a unit for the other schedules. Furthermore, the greatest errors observed in any one run of MULTINEST are on the order of 0.15 units of log-evidence for the standard sampling schedule and on the order of a few hundredths of a unit for the remaining schedules. In light of the scale of meaningful scale of log-evidence differences stated above, this degree of precision is satisfactory for the purposes of model selection.

Finally, that the estimates provided are reliable: This is meant in the sense that the estimates are relatively stable across replications. Aside from the acceptable variance in the ML estimates detailed above, the reliability is seen in the tight (on the scale set by the SE) and the relatively symmetric ranges of the errors in the individual estimates. These findings show that, across all the thousands of ML estimates collected, there were no outliers and no skewness in the estimates obtained; extreme aberrations and other anomalous artefacts therefore seem unlikely, and the estimate obtained is therefore probably close to the true one. Note that this conclusion is further bolstered when considered in conjunction with the findings regarding the reliability of the error estimates in the next assessment.

A2. FULL POSTERIOR DESCRIPTIONS OF ACCURACY AND PRECISION. The posterior distributions of the accuracy (bias) and precision, using the model detailed above, are shown in figures 5.2 and 5.3; key numerical information about the posterior distributions given in tables 5.3 and 5.4. These results reinforce the key findings of sub-assessment A1, namely, that the estimates of the log-evidence returned by MULTINEST (at least in this test case) are accurate and precise. The key extension that this full posterior description introduces is to take into account the uncertainty in the estimates of the bias and precision for each schedule.

The question of accuracy, under the terms of this sub-assessment, can be addressed thus: If a method is accurate, the posterior distribution of the bias estimator μ_ϵ will be centred on zero, or, at the very least, the zero value will fall within the main bulk of the posterior distribution, which by definition describes the likely values of μ_ϵ. Attending to the plots in figure 5.2, we see that this condition is met by every schedule considered. A complementary reading of these results is that the bias estimates provided in table 5.2 (as the mean absolute errors in the LMLEs) are subject to but little uncertainty, since the standard deviations of the bias posteriors are relatively small; which means that the small values of the bias estimates almost certainly indicate that MULTINEST exhibits at most small bias in estimation, at least in the case tested; but that what uncertainty

	Posterior parameters			*Distribution statistics*	
Sampling schedule	λ	ξ_t	ν	MAPP	SD
Standard sampling	1.774×10^{-3}	1.343×10^{-3}	994	1.774×10^{-3}	1.344×10^{-3}
High precision	3.478×10^{-4}	3.387×10^{-4}	227	3.478×10^{-4}	3.402×10^{-4}
Mimic $K = 100$, Beta	-1.320×10^{-4}	2.448×10^{-4}	910	-1.320×10^{-4}	2.451×10^{-4}
Slack $K = 100$, Beta	3.081×10^{-4}	2.366×10^{-4}	1098	3.081×10^{-4}	2.369×10^{-4}
Mimic $K = 50$, Beta	6.247×10^{-5}	3.294×10^{-4}	1151	6.247×10^{-5}	3.296×10^{-4}
Slack $K = 50$, Beta	-2.679×10^{-4}	3.518×10^{-4}	1024	-2.679×10^{-4}	3.522×10^{-4}

Table 5.3. PROPERTIES OF THE MARGINAL POSTERIOR DISTRIBUTION OVER THE ML ESTIMATE BIAS. The first column designates the sampling schedule. The next three columns provide the hyperparameters of the Student's t-distribution which describes the marginal posterior over the bias parameter μ_ϵ, giving the location, scale, and degrees-of-freedom hyperparameters respectively. The fifth column states the MAPP value of the bias, which is equal to the location parameter and is also the mean and mode of the posterior; the sixth column states the standard deviation of the posterior distribution, namely, the uncertainty in the MAPP estimate. Note that the mean absolute error presented in table 5.2 is identical to the posterior location hyperparameter and MAPP value presented here. Note also that, since the degrees-of-freedom parameter is large, the posterior distribution is very close to normal. For plots of the posteriors, see figure 5.2.

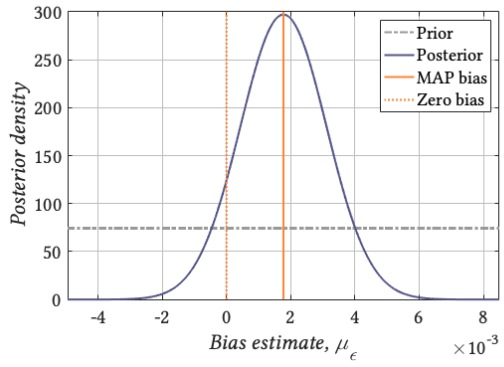
does exist is enough to encompass the possibility that there is no bias in the LMLEs obtained.

The question of precision can be similarly addressed: the posterior distribution over σ_ϵ informs us of the likely values of the precision *qua* standard deviation of repeated samples of MULTINEST's LMLEs operating under the same schedule, with the mode of the distribution representing the most likely value of the precision and the distribution's standard deviation representing uncertainty in the true value. As with the μ_ϵ case, the posterior distribution over σ_ϵ confirms that the precision estimate provided in table 5.2 (as the SE) is subject to only a little uncertainty, with the bulk of the posterior falling very close to MAPP estimate when considered on the "meaningful" scale of log-evidence units outlines above. Therefore, the conclusion drawn regarding accuracy in the results holds when taking the uncertainty in the precision into account.

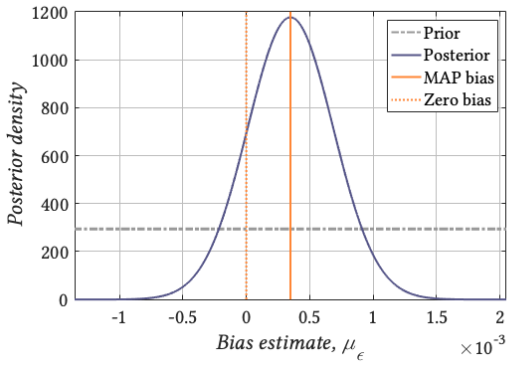
A3. ANALYSES FOR EVIDENCE OF BIAS. The results of the posterior ratio, likelihood ratio, and Bayesian t-tests are presented in table 5.5. Inspecting these results, it is clear that for each set of analysis data – that is, for every sampling schedule – there is no evidence of bias in the LMLEs returned, confirming what was suggested in the findings of the previous two sub-assessments.

The strongest support for the hypothesis that the estimates are biased is found under the posterior and likelihood ratio tests for the data returned from the standard sampling schedule; even there, the weight of evidence comes out to less than two units on the deviance scale, which is to say, "barely worth a mention". (This is equivalent to roughly 5:2 odds in favour of the alternate hypothesis against the null.) When the alternative hypothesis is broadened to include the possibility that there may be bias of any extent whatsoever, per the Bayesian t-test, there is *de facto* no evidence to suggest that MULTINEST is a biased estimator of log-evidence.

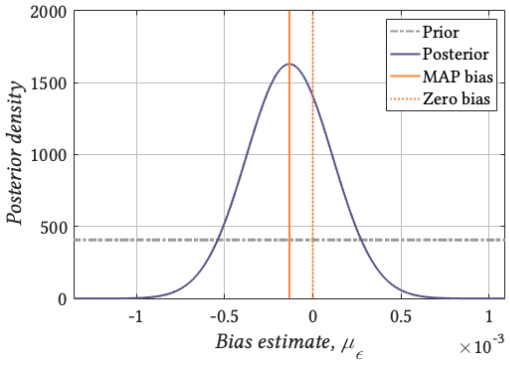
(a) Standard sampling



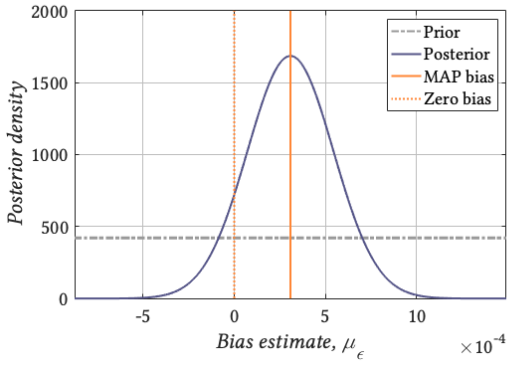
(b) High precision



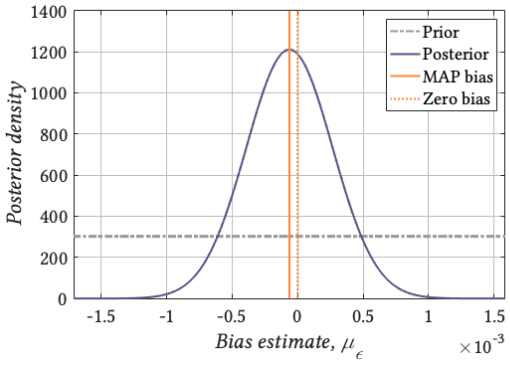
(c) Mimic $K = 100$ Beta



(d) Slack $K = 100$ Beta



(e) Mimic $K = 50$ Beta



(f) Slack $K = 50$ Beta

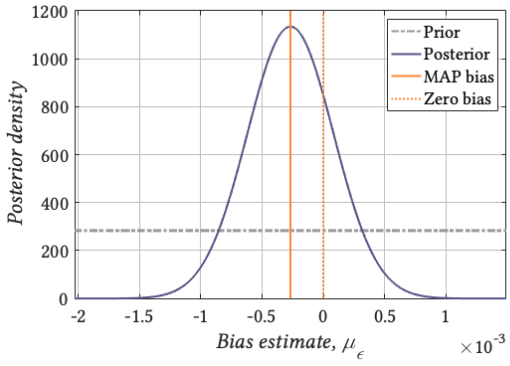
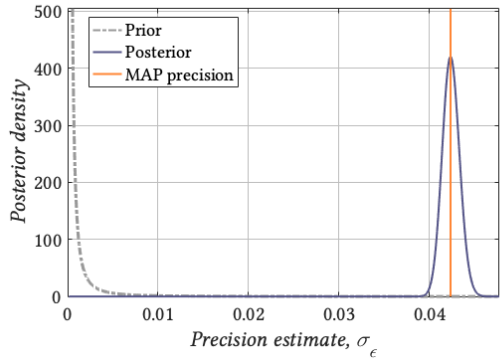
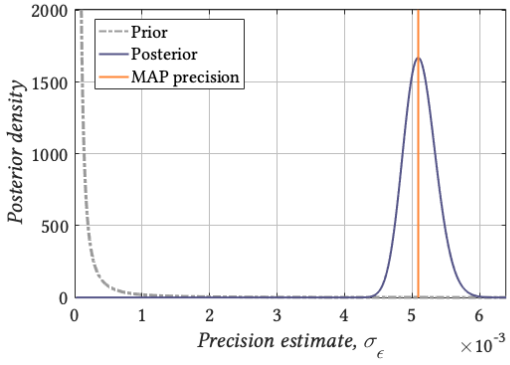


Figure 5.2. POSTERIOR DISTRIBUTIONS OF THE BIAS FOR EACH SCHEDULE. This analysis treats the mean absolute error quantity of table 5.2 as a random variable. The dashed grey line represents the uninformative prior distribution over the bias parameter μ_ϵ. The solid blue line represents the posterior distribution given, as the conditioning data, the errors in the 1000 log-evidence estimates returned by MULTINEST for the stated schedule. Note that since the prior is improper (being equal over the entire real line), the line is indicative only. The solid gold line marks out the maximum *a posteriori* parameter; the dashed gold line, the location of zero bias. Note that each subplot is centred on the MAPP estimate for its posterior, not on zero error; note also the differences in scale of the x-axis for each schedule. For numerical details of the parameters and distribution statistics, see table 5.3.

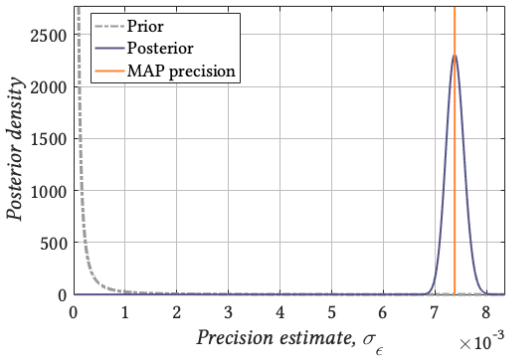
(a) Standard sampling



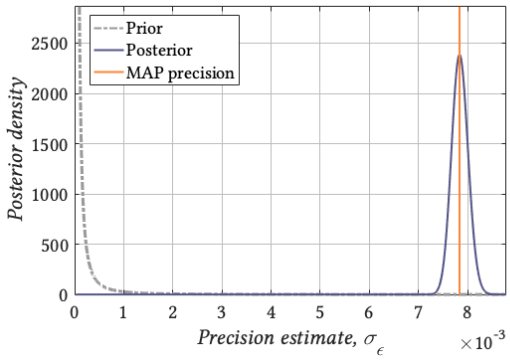
(b) High precision



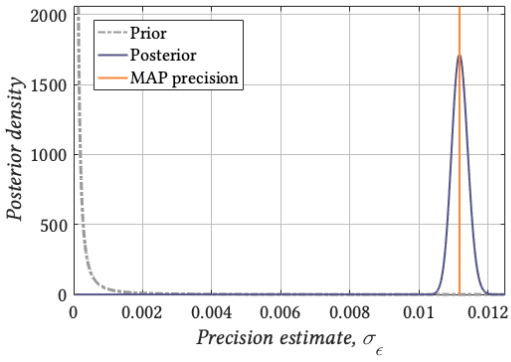
(c) Mimic $K = 100$ Beta



(d) Slack $K = 100$ Beta



(e) Mimic $K = 50$ Beta



(f) Slack $K = 50$ Beta

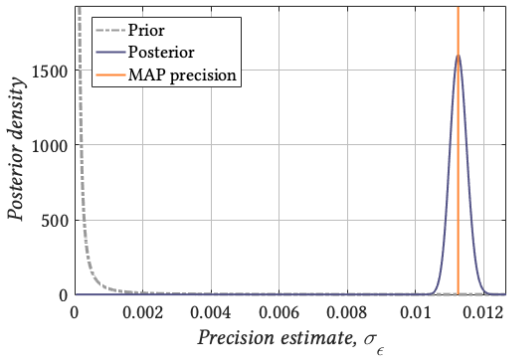


Figure 5.3. POSTERIOR DISTRIBUTIONS OF THE PRECISION FOR EACH SCHEDULE. This analysis treats the SE of table 5.2 as a random variable. The dashed grey line represents the uninformative prior distribution over the precision parameter σ_ϵ. The solid blue line represents the posterior distribution given, as the conditioning data, the absolute errors in the 1000 log-evidence estimates returned by MULTINEST for the stated schedule. Note that since the prior is improper (being equal over the log-scale of the entire real line), the line is indicative only. The solid gold line marks out the maximum *a posteriori* parameter. Note the differences in scale of the x-axis for each schedule. For numerical details of the parameters and distribution statistics, see table 5.4.

	Posterior parameters		*Distribution statistics*			
Sampling schedule	ν	$\xi_\chi/10^{-5}$	MAPP	*Mean*	SD$/10^{-3}$	*Sample* SE
Standard sampling	994	179.4	0.04231	0.04240	8.989	0.04236
High precision	227	2.616	5.092×10^{-3}	5.137×10^{-3}	1.581	5.115×10^{-3}
Mimic $K = 100$, Beta	910	5.459	7.381×10^{-3}	7.397×10^{-3}	1.603	7.389×10^{-3}
Slack $K = 100$, Beta	1098	6.154	7.838×10^{-3}	7.852×10^{-3}	1.624	7.845×10^{-3}
Mimic $K = 50$, Beta	1151	12.50	0.01117	0.01119	2.286	0.01118
Slack $K = 50$, Beta	1024	12.69	0.01125	0.01127	2.372	0.01126

Table 5.4. PROPERTIES OF THE MARGINAL POSTERIOR DISTRIBUTION OVER THE ML ESTIMATE PRECISION. The first column designates the sampling schedule. The next two columns provide the hyperparameters of the inverse chi-squared distribution which describes the marginal posterior over the bias parameter expressed as a variance ^{2}, giving the degrees-of-freedom and scale hyperparameters respectively. The fourth column states the MAPP value of the precision (i.e., the mode of the posterior), transformed to be expressed in s.d. units, ; the fifth and sixth columns give the posterior's mean and standard deviation of the posterior distribution similarly transformed. The MAPP value represents the most likely value of the uncertainty in the ML estimate; the posterior standard deviation indicates the uncertainty in the value of the uncertainty. Finally, for reference, the final column provides the SE estimate of the precision as given in table 5.2. For plots of the posteriors, see figure 5.3.

	Test		
Schedule	POST. RATIO	LIKE. RATIO	BAYESIAN T
Standard sampling	−1.75	−1.74	−0.001
High precision	−1.06	−1.05	0.010
Mimic $K = 100$, Beta	−0.29	−0.29	0.041
Slack $K = 100$, Beta	−1.70	−1.69	0.001
Mimic $K = 50$, Beta	−0.04	−0.04	0.053
Slack $K = 50$, Beta	−0.58	−0.58	0.035

Table 5.5. RESULTS OF STATISTICAL TESTS FOR EVIDENCE OF BIAS. The first column gives the sampling schedule; the rest give the results of the posterior ratio, likelihood ratio, and Bayesian t-test as described in the text. The results in each case are formed as an probability of evidence ratio of a null against an alternate hypothesis, the latter of which are described in the text. The results are given on the deviance scale, that is, as twice the natural logarithm of the evidence ratio. In each case, the null hypothesis – which takes the place of the numerator in the evidence ratio – is that the ML estimates returned by MULTINEST are unbiased. Positive values indicate that the test favours the null hypothesis; negative values, the alternate. Interpreting the results according to table 2.1, none of the tests significantly favour the hypothesis that the estimates are biased, however that hypothesis is formulated.

2:4 Assessment B: Accuracy of Bayesian evidence uncertainty estimate

Methods

PREAMBLE. With any estimator of marginal likelihood, it is necessary to establish the error in the obtained ML estimates to be able to draw meaningful conclusions about the Bayes factors between models. To establish the error, thermodynamic methods require multiple runs, the standard deviation of which estimates the sampling error (which by definition is the standard deviation of the limiting distribution of ML estimates as more and more LMLEs are obtained by the estimator). By contrast, MULTINEST returns a one-shot estimate of the error on each each run alongside the ML estimate; we will call this estimate of the SE the *autogenous log-evidence error estimate* (ALEE), since it is automatically generated as a by-product of the nested sampling process. This functionality is potentially very useful, since, if the ALEEs are reasonably accurate, it obviates the need perform multiple, potential costly, runs to calculate the quantity. The aim of this section is to assess the accuracy and usefulness of the ALEEs provided by MULTINEST.

B1. ACCURACY. The first of two parts of this assessment compares the values of the ALEEs with the posterior (over the precision parameter) and standard error of the LMLEs for each schedule, in order to establish whether the ALEEs are reasonably accurate against those more fundamental standards. Recall that the posterior distributions of the precision parameters, as found in the previous assessment, describe the likely values of the standard error for each schedule as determined by the variance in the ML estimates alone (since both the precision and the error are referring to the standard deviation of the estimator sampling). We therefore expect accurate ALEEs to fall within the bulk of the precision posterior distribution. We thus compare the ALEEs returned under each schedule to the posterior precision in figure 5.3 and to the SE reported in table 5.2. These comparisons are shown in figure 5.6 and table 5.6.

B2. UTILITY. The second part of this assessment proceeds from the fact that, in general, the principal use of stating an error alongside an estimate is (on the assumption that the estimator is unbiased) to be able to define a *credible interval*, namely, a range of values within which one can be reasonably confident the true value of the estimated quantity lies. To this end, ideally, the ALEE returned by MULTINEST on a given run could be plugged in as the standard deviation of a normal distribution centred on the associated LMLE, which normal distribution would then approximate one's posterior confidence distribution over the true value of the log-evidence, which posterior would in turn yield the credible interval.

However, inspection reveals that the LMLEs and the ALEEs returned by MULTINEST are highly negatively correlated; this raises the possibility that, even if the ALEEs returned by MULTINEST are accurate on the whole (as investigated in B1), they could not be used reliably to determine the desired credible intervals. (For a more detailed explanation of why this is the case, see appendix.)

The second part of this assessment addresses this possibility.

To determine whether the ALEEs returned by MULTINEST are useful in the sense of allowing credible intervals to be estimated, the following procedure is performed: First, the absolute error in each LMLE is re-expressed in terms of units of its associated ALEE; we refer to this scaled error value as α. Second, the proportion of runs for which the true log-evidence value falls within one, two, or three units of α calculated. Third, the distribution of α scores is tested for normality; for if the distribution of the errors of the ML estimates under transformation to α scores retains normality, then it can be concluded that the correlation between the ML estimates and the ALEEs is not sufficient to significantly render the ALEE values incapable of forming the desired credible intervals. Figure 5.6 displays the histograms of α and table 5.7 enumerates the coverage statistics.

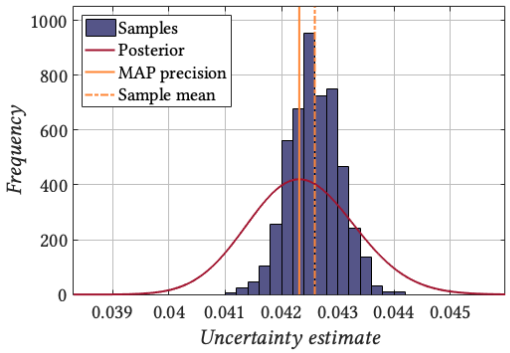
Results and discussion

B1. ACCURACY. Figure 5.4 depicts normalised histograms of the ALEEs returned by MULTINEST, set against the posterior distributions of the sampling standard deviation calculated in figure 5.3. The mean ALEE, the SE of the LMLE samples, and the absolute and relative differences of the mean ALEE against the obtained LMLE SE are shown in table 5.6. The figure shows that the error estimates returned by MULTINEST are not ideally accurate, and perhaps show a tendency towards underestimation: four of the six schedules tested show error estimates which cluster at the low end of posterior distribution over the sample standard deviation. The table, however, clearly shows that the extent of this inaccuracy is, in practical terms, relatively small: the relative error in the ALEEs compared to the LMLE sample SEs are in the worst cases on the order of 5-7%, which translates to deviations in the ALEE from its optimal value equal to a few parts in ten thousand in units of log-evidence; per the scale established in the results of A1, such a small difference is very unlikely to make meaningful differences in any inference which uses these estimates.

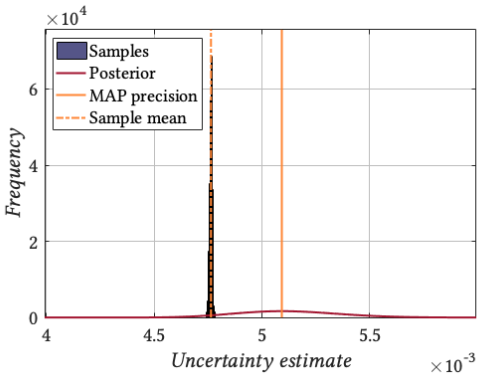
B2. UTILITY. The correlation between the ML estimates and error estimates returned by MULTINEST is portrayed in figure 5.5. Histograms of the log-evidence errors across all the runs performed for a schedule, where the error is measured in terms of α – units of the run's estimated error – are shown in 5.6. Table 5.7 shows the percentage of runs for which the true value of the log-evidence falls within one, two, and three α alongside the results of normality tests of the α-error distributions; the table also enumerates the results of normality testing for the α-error distributions against the normal distribution with zero mean and unit variance (which, ideally, the α-distributions should follow).

These results find that, for every schedule, the α-errors are approximately normally distributed about zero error with unit standard deviation; thus, the distribution of α-errors is on the whole indistinguishable from the the distribution of absolute errors in the LMLEs normalised by the SE of the LMLEs sample. We therefore surmise that, despite the correlation between the ALEE

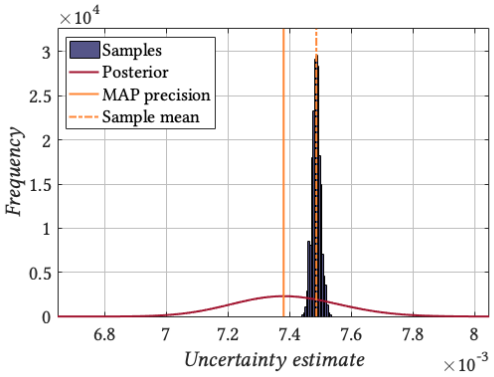
(a) Standard sampling



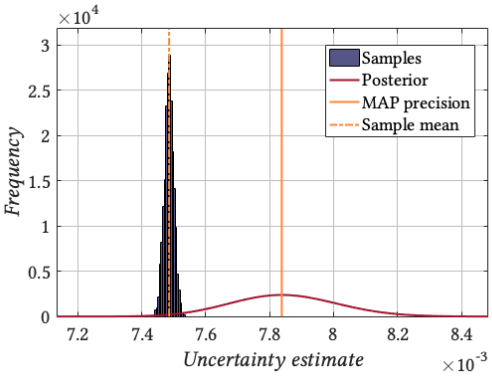
(b) High precision



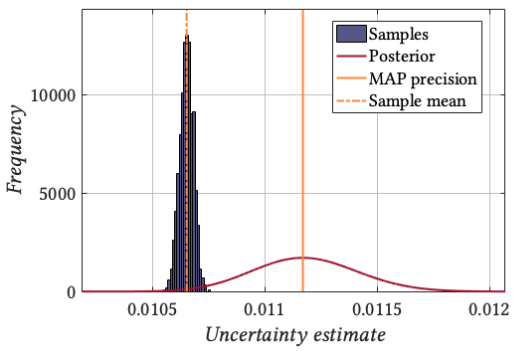
(c) Mimic $K = 100$ Beta



(d) Slack $K = 100$ Beta



(e) Mimic $K = 50$ Beta



(f) Slack $K = 50$ Beta

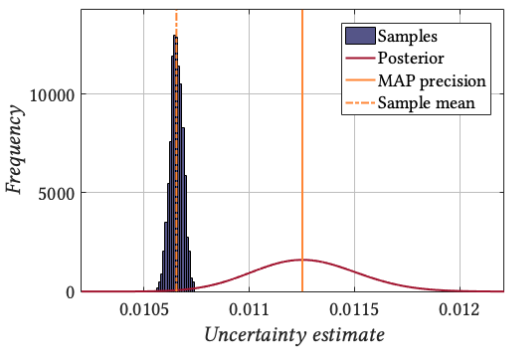


Figure 5.4. HISTOGRAMS OF THE ERROR ESTIMATES RETURNED BY MULTINEST SET AGAINST THE POSTERIOR DISTRIBUTIONS OF THE SAMPLING STANDARD DEVIATION CALCULATED IN FIGURE 5.3. The histogram in each sub-figure depicts the distribution of the error estimates returned by MULTINEST on each run within a schedule. The red line traces the posterior distribution of the standard error derived by considering solely the log-evidence estimates returned by the MULTINEST; this is identical to the posterior distribution over the precision parameter as shown in figure 5.3. The solid gold vertical bar marks the MAPP value of the SE from the posterior distribution; the dashed bar marks the . Since the MAPP precision and the observed sample SE are very close (cf. table 5.4), the difference between the two is approximately equal to the differences tabulated in table 5.6.

Schedule	*Mean est. error*	*Sample* SD	*Abs. diff.*	*Rel. diff. %*
Standard sampling	0.0426	0.0424	0.0002	0.55
High precision	0.0047	0.0051	−0.0004	−6.86
Mimic $K = 100$, Beta	0.0075	0.0074	0.0001	1.32
Slack $K = 100$, Beta	0.0075	0.0078	−0.0004	−4.58
Mimic $K = 50$, Beta	0.0107	0.0112	−0.0005	−4.70
Slack $K = 50$, Beta	0.0107	0.0113	−0.0006	−5.41

Table 5.6. DIFFERENCES OF UNCERTAINTY ESTIMATES FROM OBSERVED SAMPLING VARIANCE. The first column indicates the schedule to which a row corresponds. The second column lists the mean value of the log-evidence error estimates returned by MULTINEST across all the runs within a schedule, i.e., the mean of the SE estimates. The third column lists the standard deviation of the log-evidence estimates across runs, i.e., the observed SE as listed in table 5.2. The fourth column contains the differences between the mean SE estimate of the second column and the observed SE of the third; this value is in units of log-evidence. The final column converts the absolute errors of the fourth column into relative errors (relative to the observed SE).

and LMLEs, the the ALEEs returned by MULTINEST are useful in this particular case, *viz.*, from the results returned on a single run of MULTINEST, one can make reasonably confident statements about the probability of the true log-evidence falling within a certain number of α; put differently, the ALEE can be used to make meaningful statements about the uncertainty in the log-evidence estimate obtained without having to resort to multiple runs to calculate the SE in the ML estimates, as desired.

2:5 Assessment C: Accuracy of posterior estimate

Methods

PREAMBLE. The final assessment tests the second major function of MULTINEST (besides ML estimation), corresponding to the second major problem of statistical inference: parameter estimation, or in Bayesian terms, posterior estimation. Much as with MCMC methods, the posterior estimate in nested sampling takes the form of a set of posterior samples p^*, not as a well-specified function. To generate posterior inferences, these samples are used to derive the posterior distribution and its characteristics.

Now, with any estimator of the posterior, the samples p^* are evidently not drawn from the true posterior $\mathcal{P}_{\text{true}}$ directly.[viii] Instead, the estimator applies some mathematical scheme to first sample points from the parameter space of the model under consideration and then to assign weights those sampled points; the weighted samples so obtained are supposed to approximate samples drawn from the true posterior. With MULTINEST, the posterior samples p^* are generated by taking the full sequence of accepted points (which are points in the model's parameter space)

[viii] For if the true posterior were available so that samples could be be drawn therefrom, there would hardly be a need to estimate it.

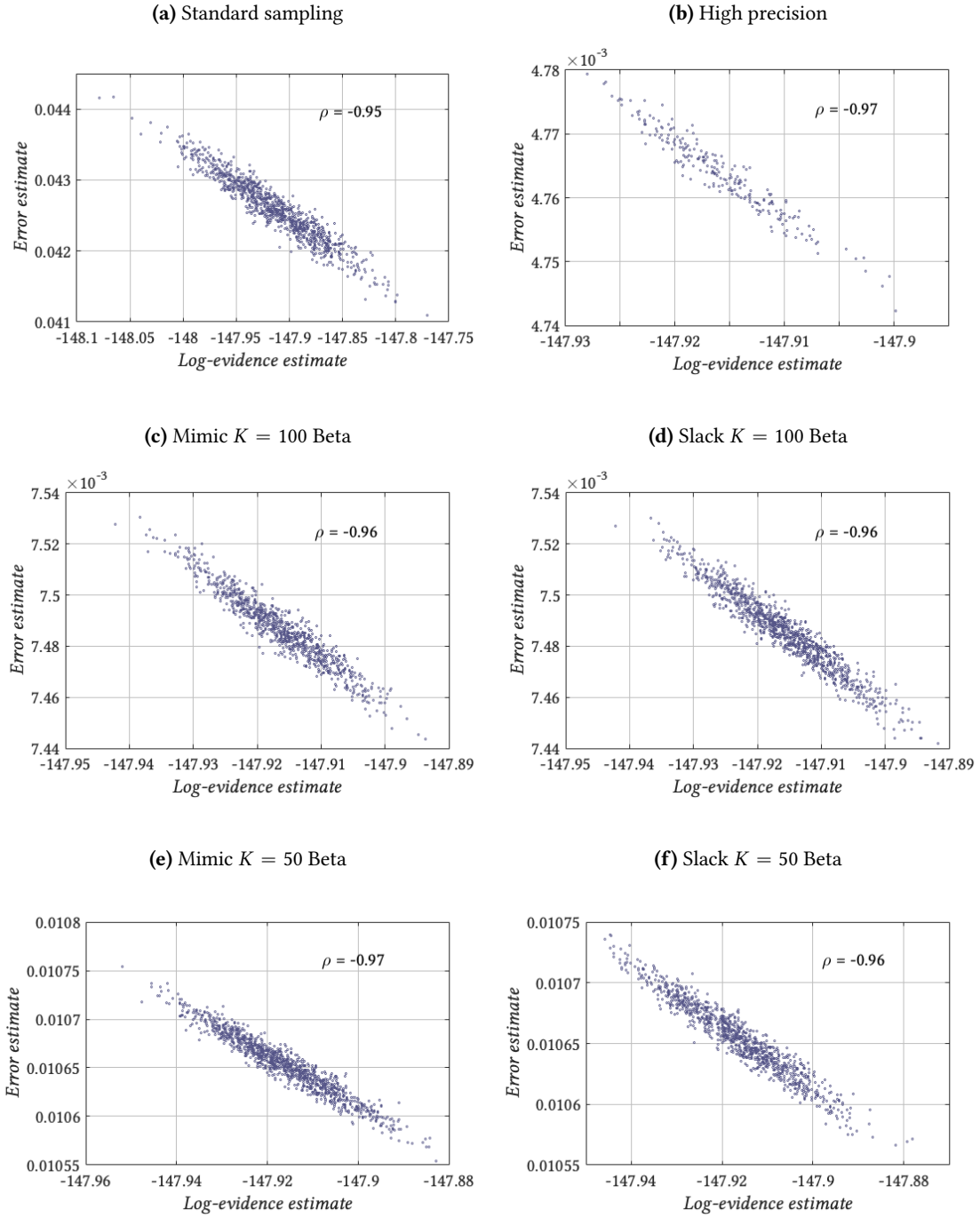
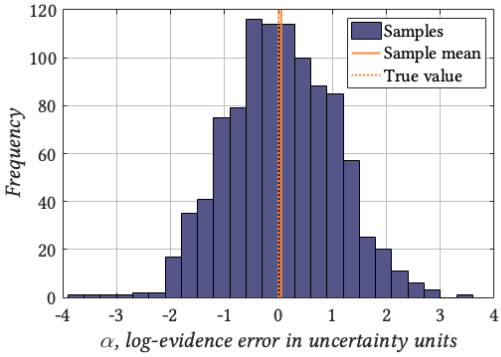
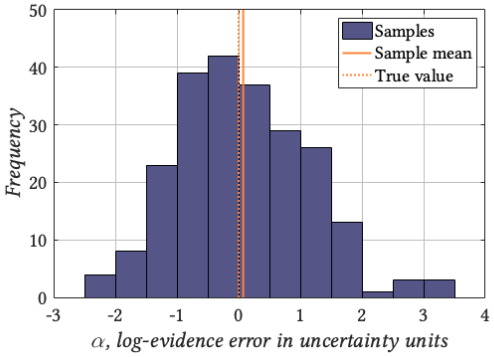


Figure 5.5. SCATTER PLOTS OF THE CORRELATION BETWEEN THE LOG-EVIDENCE AND ERROR ESTIMATES RETURNED BY MULTINEST. Each plotted point corresponds to the output of an individual run of MULTINEST under the indicated schedule. The abscissa gives the value of the log-evidence estimate obtained in the run; the ordinate, the value of the autogenous SE estimate. The Pearson's r correlation coefficient is also stated on each plot as a guide to the strength of the very obvious correlations.

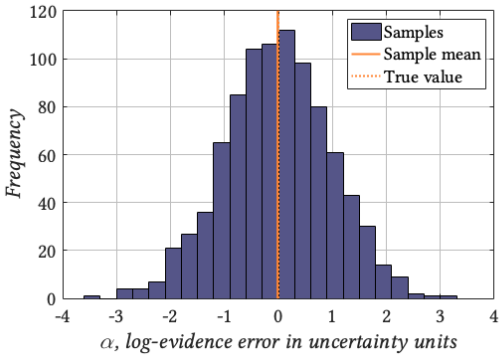
(a) Standard sampling



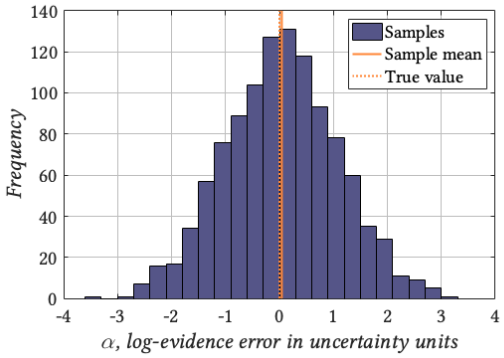
(b) High precision



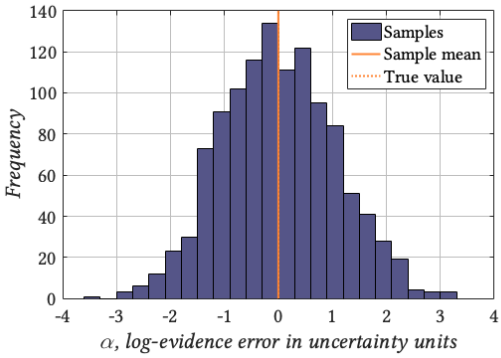
(c) Mimic $K = 100$ Beta



(d) Slack $K = 100$ Beta



(e) Mimic $K = 50$ Beta



(f) Slack $K = 50$ Beta

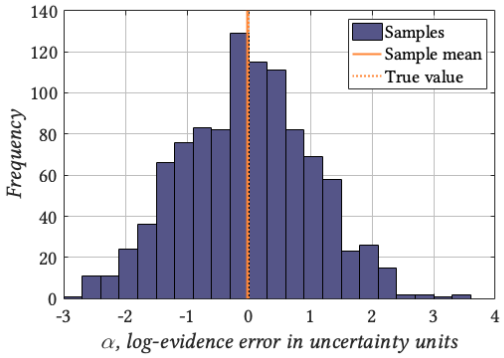


Figure 5.6. HISTOGRAMS OF THE α-ERROR IN THE LOG-EVIDENCE ESTIMATES RETURNED BY MULTINEST. For a given run, the α-error of its log-evidence estimate is the absolute error in the log-evidence estimate divided by the value of the run's autogenous SE estimate. The solid gold vertical bar picks out the mean α-error across all the runs in the stated schedule; the dashed bar, the location of zero error. Summary statistics of the α-error distributions are provided in table 5.7. Compare to figure 5.1 to see the effect of the re-expression in terms of ALEE units.

Schedule	1α %	2α %	3α %	KS p
Standard sampling	68.1	96.2	99.6	0.16
High precision	64.5	95.2	98.7	0.45
Mimic $K = 100$, Beta	69.2	96.2	99.8	0.82
Slack $K = 100$, Beta	65.4	93.9	99.8	0.36
Mimic $K = 50$, Beta	64.9	94.1	99.7	0.19
Slack $K = 50$, Beta	63.6	93.9	99.7	0.09
[Normal distribution]	68.3	95.5	99.7	-

Table 5.7. PERCENTAGE OF RUNS FOR WHICH $\mathcal{Z}_0$ FALLS WITHIN α UNITS OF MULTINEST'S AUTOGENOUS LOG-EVIDENCE ERROR ESTIMATE. Given that each run of MULTINEST returns a log-evidence estimate and an associated error estimate, for what percentage of runs does the true value of the log-evidence falls within one, two, and three units of the estimated error? If the autogenous SE estimates are to be as useful as possible, the α-errors (errors in autogenous SE units) in the log-evidence estimates should be approximately normal. The p-values of the Kolmogorov-Smirnoff test when applied to the distributions of α-error are given in the final column. Values of the percentage of estimates that would fall within one, two, and three units of standard error if an estimate was unbiased and normally distributed about the true value are shown for reference in the last row. Histograms of the α-error distributions constructed from all the runs performed under each sampling schedule are shown in figure 5.6. Scatter plots of the correlation between the log-evidence and autogenous SE estimates calculated by MULTINEST are shown in figure 5.5.

in the nested sampling process and assigning to each a posterior weight equal to that point's likelihood multiplied by the prior mass accumulated by that point's acceptance and normalised by the estimated model evidence.

In other words, we can imagine that the posterior samples are being drawn from what we shall call the *estimator parexic posterior distribution* $\mathcal{P}^*$ (which can be thought of as the distribution which results from repeatedly applying the posterior sampling process with given likelihood, prior, and data), which distribution is supposed to imitate the target distribution (the true posterior).[ix] Accurate posterior estimators will therefore have $\mathcal{P}^*$ which closely mirror the true posterior distribution $\mathcal{P}_{\text{true}}$, at least once enough samples have been collected.

Whetted against these considerations, the goal of assessing the posterior estimation capacities of MULTINEST is sharpened: the problem is to determine whether the posterior samples p^* returned are a good indicator of the true posterior distribution $\mathcal{P}_{\text{true}}$. Since, in the present situation, the true posterior distribution's functional form and parameters are known – see equation 5.4 – the problem reduces to using to using the p^* to approximate the form or the statistics of the estimator parexic posterior distribution $\mathcal{P}^*$ from which they are drawn, then comparing and evaluating the form or the statistics of $\mathcal{P}^*$ so derived to the form or the statistics of $\mathcal{P}_{\text{true}}$.

A quick note on the data used for this analysis: MULTINEST returns posterior samples in two

[ix] "Parexic" is Greek for "nearby", with roots *para*, almost, *ek*, out; its use here is borrowed from a congruent usage in Lanczos, 1956, where it is proposed to designate, roughly, things of or related to the theory of approximation by finite algorithms, in contradistinction to the exact or arbitrarily close analysis of quantities sought by algebraic methods and "pure" analysis: "[Hence] the term 'parexic analysis' can well be adopted to mean that we do not want an exact but only a 'nearby' determination of a certain quantity."

forms. The first is, as mentioned above, as a list of weighted posterior samples; the second is as a set of equally weighted posterior samples. The former are the most direct output; the latter, calculated from the former, can be conceived of as a set of draws from the estimator sampling distribution. Which of these two sets is used for a sub-assessment will be made clear below.

C1. GRAPHICAL ANALYSES OF THE ESTIMATED AND TRUE POSTERIOR DISTRIBUTIONS. Since the posterior distribution (as a whole) is one of the key outcomes of interest in posterior estimation, the first sub-assessment comprises two visual inspections of the goodness-of-fit of the estimator parexic posterior distribution function with respect to the true posterior distribution function.

The first comparison addresses the PDF representation of the distribution. The comparison is conducted for every run within every schedule, with the procedure for a single run going as follows: the set of discrete posterior samples obtained by applying MULTINEST is used to estimate the parexic posterior distribution PDF through kernel density estimation. The kernel density estimate of the parexic posterior thus constructed is then graphically compared to the true posterior function to assess the similarity of the two functions. The procedure is repeated for each schedule to obtain the PDF resulting from kernel density estimation inputting all the samples collapsed across all the runs within the schedule; these samples can be combined and considered as a single set because every individual sampled points can, in principle, be considered an IID draws from the parexic posterior distribution. Note that the samples used to form the KDE posterior PDFs are the weighted posterior samples. The kernel density estimates used an Epanechnikov kernel and a relatively broad bandwidth, set as described in appendix G. The comparison of the KDE estimates to the true posterior are shown in figure 5.7.

While KDE allows the useful PDF representation of the posterior to be estimated from the posterior samples, these estimates are subject to the distortions inherent to kernel smoothing; in particular, the choice of bandwidth can over- or under-smooth the samples. To better assess the fit of the distribution of the posterior samples to the true posterior, we inspect the probability-probability (PP) graphs of the empirical CDFs of the samples against the CDF of the analytical posterior distribution. Note that for this part of the analysis, the equally-weighted set of posterior samples returned by MULTINEST is used.

In general, given two CDFs over the same domain, with one CDF designated the "test" and the other the "reference" distribution, a PP graph consists of a set of points where, for each point, the abscissa is the value of the reference CDF at an evaluated quantile of the domain and the ordinate is the probability returned by the test CDF evaluated at that same quantile. Evidently, the abscissa and ordinate values both range from 0 to 1; further, if the test and reference distributions are identical, the plotted points on the graph fall along the straight line joining $(0, 0)$ to $(1, 1)$. Thus, when the test distribution is not equal to the reference, greater deviations of the PP function from that straight line represent greater deviations of the test from the reference distribution.

For this second visual assessment, the comparison procedure for every run in each schedule

goes as follows: we form the empirical CDF from the equally weighted posterior samples returned by MULTINEST, and take it as the test distribution. The reference distribution is, of course, the analytically correct posterior distribution; by evaluating it at those quantiles enumerated in the equally weighted posterior sample data, the desired PP graph can be formed.

Deviations of the empirical CDFs (as derived from the posterior samples) from the "ideal" straight line on the PP plot are to be expected; to judge the closeness of fit, it is necessary to know *how much* deviation ought to be expected, and the observed deviation compared to it. Therefore, to better assess whether the posterior samples returned by MULTINEST are drawn from $\mathcal{P}^*$ closely approximating $\mathcal{P}_{\text{true}}$, the PP plot analysis just described is repeated using matched Monte Carlo samples from the true posterior: if the PP plots obtained from the Monte Carlo procedure are indistinguishable from those derived from MULTINEST's posterior samples, then it may be inferred that $\mathcal{P}_*$ is similar to $\mathcal{P}_{\text{true}}$. Matched Monte Carlo posterior samples are obtained for each run i by drawing n_i random samples from the $\mathcal{P}_{\text{true}}$, where n_i is the number of posterior samples returned by MULTINEST for that run; PP graphs are then formed as they were from the MULTINEST posterior samples.

Figure 5.8 shows, on each sub-figure, the PP graphs for all the runs of a given schedule. Note that, since the number of runs, and hence the the number PP graphs, is high, the PP graph of any individual run is indiscernible; instead, the figure emphasises the *coverage*: the region that contains the PP graphs aggregated across all the runs, formed by finding the outer boundaries across all the PP graphs. Also shown is the coverage of the PP graphs obtained from the Monte Carlo procedure detailed above.

C2. STATISTICAL TESTS OF DISTRIBUTION DIVERGENCE. The graphical analyses of the previous sub-assessment provide only suggestive evidence regarding the fidelity of the estimator parexic posterior distribution to the true posterior. The final sub-assessment seeks to rigorously verify the similarity of the two distributions through the use of two statistical tests of the degree of similarity.

We wish to determine whether it reasonable to suppose that posterior samples returned by MULTINEST were drawn from the true posterior. To do this, it is necessary to establish some measure of dissimilarity between the p^* and $\mathcal{P}_{\text{true}}$; the motivating principle is that, if the posterior samples returned by MULTINEST are indeed accurate, the observed dissimilarity between them will be less than or equal to the dissimilarity observed between $\mathcal{P}_{\text{true}}$ and samples drawn directly from $\mathcal{P}_{\text{true}}$.

Let us work this idea out in more detail. We recall that MULTINEST returns discrete posterior samples. From a single of posterior samples, we can construct the empirical CDF (which we do to make the samples comparable to $\mathcal{P}_{\text{true}}$, since like can only be compared with like, i.e., distribution with distribution). The empirical CDF will, of course, diverge somewhat from the true posterior due to discretisation and random sampling effects. It is this divergence that will be precisified and

measured by the test statistics.

The first test uses an estimator of the Kullback-Leibler (KL) divergence, a well-known measure of the dissimilarity between two distributions, as the test statistic. For the parexic posterior samples returned under any and every run of MULTINEST, we calculate an estimator of the KL divergence of the empirical CDF of the posterior samples from the CDF of the true posterior (Pérez-Cruz, 2008). Thus, a distribution of test statistics is obtained.

Simultaneously, sets of Monte Carlo samples are drawn from the true posterior for each schedule, and the KL divergence of each MC sample set from the true posterior is calculated. The size of the Monte Carlo sample sets are matched to the median size of the posterior sample sets obtained from MULTINEST under a given schedule. Thereby a second distribution of KL divergences, the null distribution, is obtained.

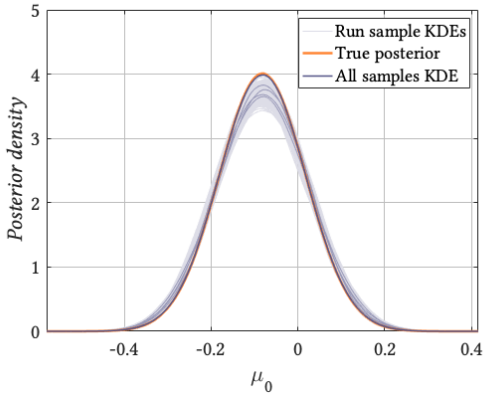
The second test will use what we shall term the D statistics, essentially the integral of the absolute difference of the empirical CDF and the true CDF. The same procedure as for the KL divergence test statistic is applied to obtain two distributions of the test statistics per sampling schedule.

Now, for both test statistics, a value of zero for the test statistic implies zero divergence – namely, identity with the true posterior. Comparing the distribution of KL divergences obtained from the MULTINEST posterior samples with the null distribution indicates the degree to which the MULTINEST posterior sample sets fall outside the expected range of dissimilarity between samples drawn directly from the true posterior. If the distribution of the test statistic as obtained from the parexic posterior samples p^* under a given schedule is on average equal to or less than that obtained from the Monte Carlo samples from the true posterior, it will be taken to mean that the p^* are adequate representations of the true posterior.

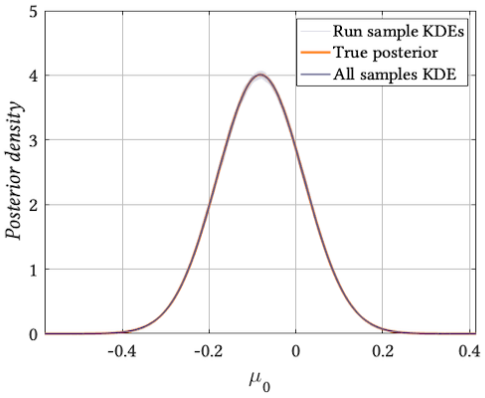
Results and discussion

C1. GRAPHICAL ANALYSES OF THE ESTIMATED AND TRUE POSTERIOR DISTRIBUTIONS. The kernel density estimates obtained from the posterior samples returned by MULTINEST are portrayed in figure 5.7. It can be seen that, in every case, the KD estimate of the parexic posterior PDF resembles the true posterior's PDF. This is especially evident in the KD estimates formed from the complete set of posterior samples obtained under a given schedule, and in the KD estimates formed from the high precision sampling schedule results, which are expected to be the most accurate. We remark that the KD estimates of the posterior samples obtained in the standard sampling schedule (when not combined) appear a little broader than the analytical posterior function. The fact that, when combined, this broadening disappears suggests that this may be an artefact of the kernel smoothing; the results of the next section *inter alia* show that this is the case.

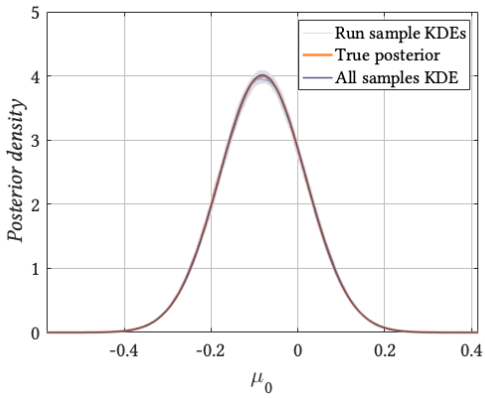
(a) Standard sampling



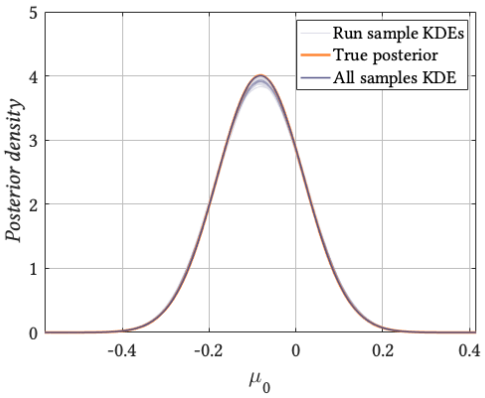
(b) High precision



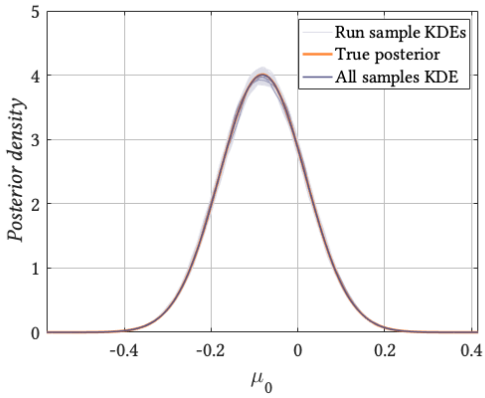
(c) Mimic $K = 100$ Beta



(d) Slack $K = 100$ Beta



(e) Mimic $K = 50$ Beta



(f) Slack $K = 50$ Beta

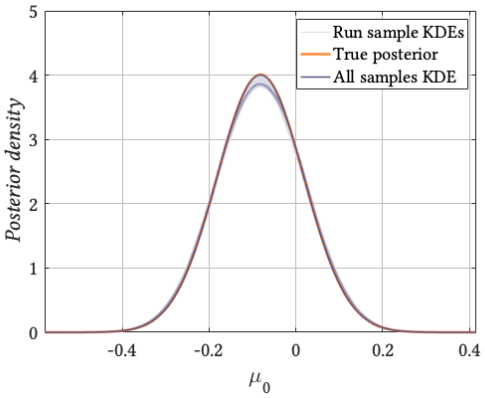
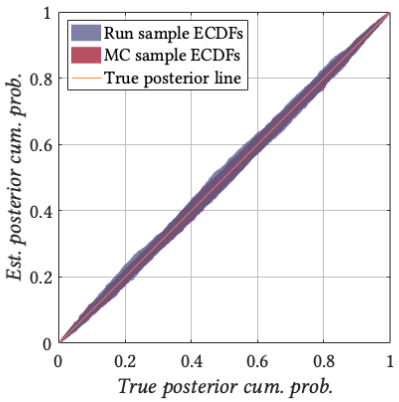


Figure 5.7. KERNEL DENSITY OF THE MODEL POSTERIOR FROM THE POSTERIOR SAMPLES RETURNED BY MULTINEST. Each figure allows for the the graphical comparison of the KDEs of the posterior function as derived from the posterior samples returned by MULTINEST. The pale blue-grey lines are the KDE posterior estimates obtained from the posterior samples from the runs within a schedule; a few, randomly chosen from amongst the schedule's run, are picked out in a slightly darker colour to better illustrate the form of the posterior estimate obtained from the samples in an individual run. The gold line marks the analytical posterior function; the dark blue line, the kernel density estimate of the posterior obtained when smoothing over all the samples from all the runs within a schedule.

(a) Standard sampling



(b) Slack $K = 50$ Beta

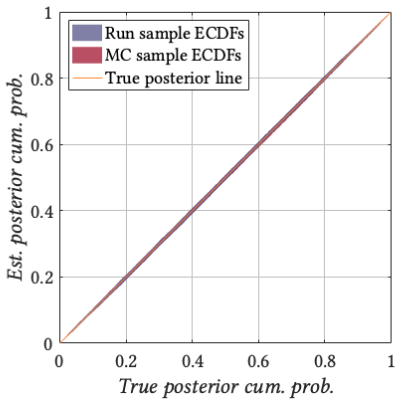


Figure 5.8. PP-PLOTS OF THE TRUE POSTERIOR DISTRIBUTION AGAINST THE ECDF FUNCTIONS FROM THE POSTERIOR SAMPLES RETURNED BY MULTINEST. The blue patch indicates the coverage of the all the PP plots (one for each run) constructed from the posterior samples returned by MULTINEST; the red patch indicates the coverage of PP plots constructed from Monte Carlo samples from the true posterior, with one such simulated plot per run matched for sample size. The PP plot obtained by plotting the true posterior against itself – a straight line from $(0, 0)$ to $(1, 1)$ – is shown for reference. Note that the PP plots are only shown for the standard and $K = 50$ slack sampling schedules; the large number of samples and their fidelity to the true posterior make the remaining PP plots graphically practically indistinguishable from the true posterior line. PP plots constructed from all the posterior samples collected across all the runs within a schedule are graphically nigh-on identical to the true posterior lines in all cases, and so are not explicitly shown.

C2. STATISTICAL TESTS OF DISTRIBUTION DIVERGENCE. The result of the statistical test of the deviance of the parexic posterior from the true posterior based on the KL divergence is shown in figure 5.9; that based on the the D statistic is shown in figure 5.10. In the latter case, simple visual inspection confirms that the observed deviances of MULTINEST's posterior samples is, on the whole, considerably less than those obtained when drawing similarly sized samples from the true posterior, as desired to establish the accuracy of MULTINEST's posterior sampling process.

In the case where the KL divergence is chosen as the test statistic, the relationship between the distributions of deviance statistics obtained from the MULTINEST samples, on the one hand, and under the null hypothesis, on the other, is not so immediately clear. Two-sample Mann-Whitney U tests of the distributions of the absolute values of the KL divergences, however, show no evidence of the average divergence of the MULTINEST posterior sample sets being greater than those obtained under the null hypothesis at the $p = 0.05$ significance level for any sampling schedule. (Indeed, except for the standard sampling schedule, where $p = 0.078$, the p values are all greater than 0.95, indicating that the posterior samples may diverge less from the true posterior than the Monte Carlo samples.)

Thus, two tests based on fundamentally conceptually different test statistics both agree that the posterior sample sets returned by MULTINEST are as close or closer to the true posterior

$\mathcal{P}_{\text{true}}$ when compared to similarly sized randomly drawn sample sets drawn directly from $\mathcal{P}_{\text{true}}$, from which we conclude that the posterior samples returned by MultiNest – at least in this simple case – are highly accurate reflections of the true posterior distribution. We note in passing that these results are unsurprising given MultiNest's ability to effectively explore challenging likelihood functions, as reported in Feroz et al. (2009).

2:6 General discussion of the first test

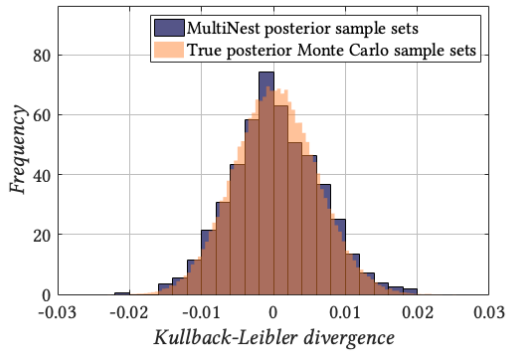
The aim of this chapter overall is to establish at least initial evidence that MultiNest could be a useful method for performing model inference in cognitive science. To that end, this section's study considered a test case involving a simple Gaussian likelihood for which the marginal likelihood (ML, or Bayesian evidence) and posterior are analytically available; in particular, this study followed and expanded on the work of Xie et al. (2011), who tested the thermodynamic integration (TI) and steppingstone (SS) methods of ML estimation, the latter of which they also introduced. The approach adopted, involving a simple and analytically tractable case, let MultiNest's efficacy as an estimator in both tasks of model inference to be set against an absolute benchmark, a basic comparison which has not previously been reported. Stringent tests found MultiNest to provide accurate and highly precise estimates of both the ML and posterior.

Aside from performing its tasks properly in the sense of producing the correct outcomes, it is important to consider how well a given tool works in comparison to alternative methods. Since TI and SS have been proposed for use in cognitive science (Annis et al., 2019; Evans & Annis, 2019), in this interim discussion, we consider what this section's results suggest regarding MultiNest's competitiveness with these methods. In particular, the focus will be on one of the most important practical considerations: the computational burden associated with the various techniques.

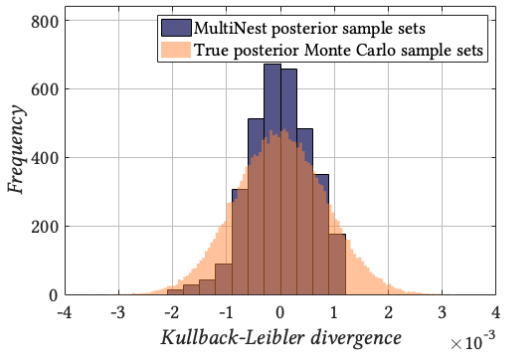
Referring to tables 5.1 and 5.2, it is clear that MultiNest is technically superior with regards to ML estimation compared to the method of thermodynamic integration, since the LMLEs obtained by MultiNest do not exhibit evidence of bias and those obtained by TI do. However, the steppingstone algorithm, like MultiNest, does not evince bias in its estimates; and it, too, achieved high precision. Furthermore, it is not immediately clear whether MultiNest offers a computational advantage over SS: while the number of posterior samples collected by MultiNest in the mimic schedules is of the same order as the number used by SS, and the number is less in the slack schedules (while still achieving similar accuracy and precision), the total number of likelihood evaluations is in all cases greater due to MultiNest's intrinsic use of rejection sampling.

Set against this, it is important to note that the implementation of SS used by Xieal2011 had the advantage of drawing samples from the power posterior directly, where it would normally rely on an MCMC sampler to provide those draws, which would introduce further uncertainty. MultiNest, by contrast, had no such information about the posterior, having to use its native

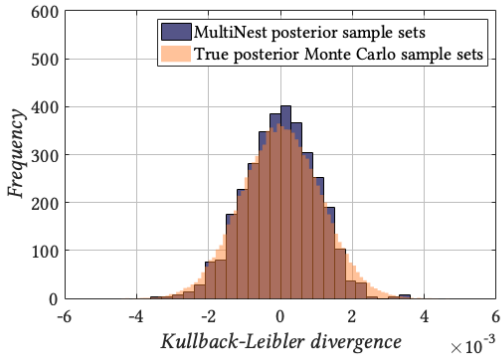
(a) Standard sampling, $p = 0.078$



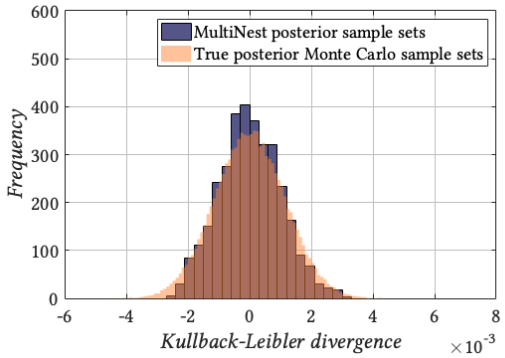
(b) High precision, $p = 1.000$



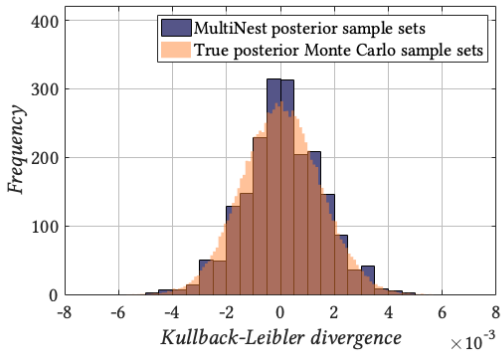
(c) Mimic $K = 100$ Beta, $p = 0.999$



(d) Slack $K = 100$ Beta, $p = 1.000$



(e) Mimic $K = 50$ Beta, $p = 0.992$



(f) Slack $K = 50$ Beta, $p = 0.952$

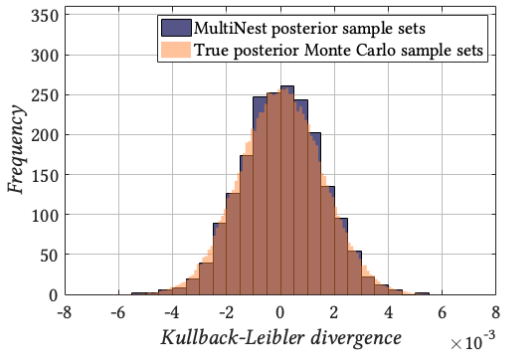


Figure 5.9. HISTOGRAMS FOR THE KULLBACK-LEIBLER DIVERGENCE TEST OF DEVIANCE. One plot for each schedule condition. The dark blue histogram on each plot represents the distribution of the estimator of the KL divergence used as the test statistic; the distribution is obtained by calculating the test statistic for every posterior sample set across all $\tilde{1}000$ runs (218 for the high precision schedule) of MULTINEST performed for the relevant schedule. The light orange histogram is of the same test statistic calculated across 10 000 independent sample sets drawn directly from the true posterior, where the size of the sample set was set to the median number of posterior samples collected per run for the schedule under consideration (cf. table 5.1).

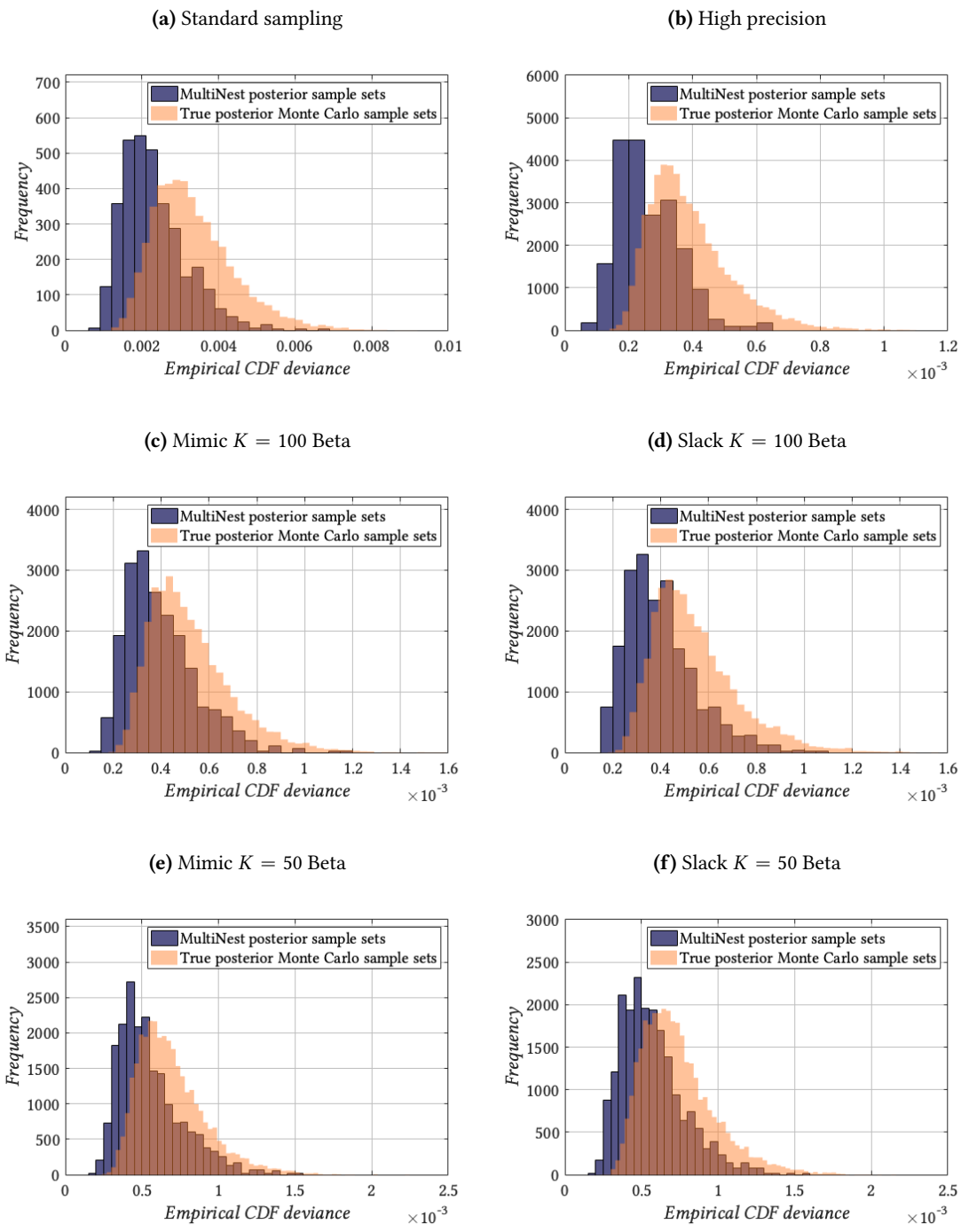


Figure 5.10. HISTOGRAMS FOR THE D-STATISTIC DIVERGENCE TEST OF DEVIANCE. One plot for each schedule condition. The dark blue histogram on each plot represents the distribution of the D stat used as the test statistic; the distribution is obtained by calculating the test statistic for every posterior sample set across all $\tilde{1}000$ runs (218 for the high precision schedule) of MULTINEST performed for the relevant schedule. The light orange histogram is of the same test statistic calculated across 10 000 independent sample sets drawn directly from the true posterior, where the size of the sample set was set to the median number of posterior samples collected per run for the schedule under consideration (cf. table 5.1).

algorithm for sampling the posterior space. A fairer test would see SS compared to MULTINEST with the former using an MCMC sampler.

All these considerations, though, are perhaps moot, for MULTINEST is capable of achieving entirely adequate results with c. 11000 total likelihood evaluations. With 50 or 100 temperature rungs, attaining an equivalent load would entail collecting little over 200 or 100 posterior samples per rung for the SS method to base its ML estimate on. At first glance, this seems like too few samples to achieve the accuracy and precision of MULTINEST. It must be stressed, however, that this is a matter for testing to establish conclusively. More broadly, it is also worth mentioning that questions of practical efficiency are in general difficult to discuss, since the choice of MCMC sampler, the particular problem at hand, and the computational equipment available all have an effect on these practical matters.

We conclude this interim discussion by saying that the results of the first test should not be taken as definitive recommendation of MULTINEST over the alternatives considered (at least when considering efficiency), especially SS, but they certainly give no reason to admonish it. It may be the case that the SS algorithm would be able to perform as or more efficiently than MULTINEST, given the opportunity. However, MULTINEST has a number of other advantages, alluded to in the Introduction and explored more fully in the Discussion, which the author believes give it further practical advantages over alternative proposed techniques. Before that, we turn our attention to another test of MULTINEST's abilities.

3. SECOND TEST · THE LINEAR BALLISTIC ACCUMULATOR

3:1 The justification for and aims of the test

The first test sought to assess MULTINEST's efficacy and competitiveness in a simple, analytically tractable scenario; such an approach allows MULTINEST's efficacy to be evaluated in a context where absolute benchmarks are available, but is arguably not very informative with respect to this chapter's ultimate goal of demonstrating MULTINEST's viability for use with cognitive models. To remedy this shortcoming, the second test applies MULTINEST to a cognitive model, the *linear ballistic accumulator* (LBA) model of choice reaction time mentioned in the introduction. The LBA, and the particular model selection problem which will be attacked here, have been used in prior research testing estimators of marginal likelihood; we therefore adopt it here to allow for a direct comparison this earlier research. In particular, the problem considered in this test case is the same as one used in Evans and Brown (2018), and later in Annis et al. (2019) and Evans and Annis (2019). In the first of these papers, the marginal likelihoods of various LBA models were calculated using the conceptually simple but computationally intensive method of marginal likelihood estimation known as *arithmetic mean* (AM) estimation; the latter two papers reported

the results of applying the steppingstone method and variants of thermodynamic integration to the same problem.

The work of this section is to apply MultiNest to the LBA selection problem mentioned above and to replicate the earlier investigations using the AM estimator. By considering these results in themselves, and comparing them to those of earlier studies, clear and direct comparisons can be made of MultiNest's performance as an estimator of the Bayesian evidence to already established methods in a cognitive setting.

3:2 Methods

Fit data and models

The fit data in this test case is the same as one of the simulated datasets used for model selection by Evans and Brown (2018), and later in Annis et al. (2019) and Evans and Annis (2019). The dataset mimics 600 trials of a two-option choice reaction time task performed by a single individual, where 300 trials are performed under one experimental condition and the remainder under another; we will denote these task conditions as 1 and 2. The data were generated using the linear ballistic accumulator model (LBA, fully described in appendix F; a brief description is in the footnote[x]) with no variation of the parameters between the conditions (a "simple" dataset). In particular, the data-generating parameters were

$$\begin{aligned} k &= 0.4, & A &= 1, & t_0 &= 0.3, \\ \mu_{\text{corr}} &= 1, & \mu_{\text{err}} &= 1, & & \\ \sigma_{\text{corr}} &= 1, & \sigma_{\text{err}} &= 1, & & \end{aligned} \tag{5.7}$$

using the same notation as the appendix, and where the subscripts "corr" and "err" refer to the accumulators for the correct and error responses. (The dataset is available for download at https://osf.io/7cbh7/.)

The model selection problem is to choose between two LBA models of this dataset, a "simple" and a "complex" version. The simple model has the same parameters across the two task conditions,

[x] The linear ballistic accumulator model for a two-choice reaction task posits that response time and choice on any given trial is the outcome of a race between two independent sensory evidence accumulators, one for each response option i (nominally, "correct" and "error"). On each trial, sensory evidence accumulates linearly in each accumulator at rate ρ_i until one reaches the (shared) response threshold b; this triggers the associated response, which will occur after a certain period of non-decision time t_0. The accumulation rate for each accumulator is drawn at random on each trial from a normal distribution with a mean and standard deviation specific to the accumulator, i.e., $\rho_i \sim \mathcal{N}(\mu_i, \sigma_i)$. Similarly, the starting distance from the threshold is drawn uniformly at random between 0 and an upper limit $A < b$. Because of the indeterminacy due to the scaling of the parameters in sensory evidence units, it is conventional to set $\sigma_{\text{corr}} = 1$. Thus, the basic two-choice LBA model is equipped with six parameters: b, A, t_0, μ_{corr}, μ_{err} and σ_{err}. The model has an analytically tractable likelihood function, covered in appendix F.

with the following priors over the LBA parameters:

$$\begin{aligned} k &\sim \mathcal{N}_+(0.4, 0.4), & A &\sim \mathcal{N}_+(1,1), & t_0 &\sim \mathcal{N}_+(0.3, 0.3) \\ \mu_{\text{corr}} &\sim \mathcal{N}_+(3,3), & \mu_{\text{err}} &\sim \mathcal{N}_+(1,1), & & \\ \sigma_{\text{corr}} &\sim \mathcal{N}_+(1,1), & \sigma_{\text{err}} &\sim \mathcal{N}_+(1,1), & & \end{aligned} \tag{5.8}$$

where $\mathcal{N}_+(\cdot,\cdot)$ indicates the truncated normal distribution with a lower bound of 0 and no upper bound. The complex model allows k, t_0, and μ_{corr} to differ between conditions. Indexing the condition in superscripts, the priors of the complex model are

$$\begin{aligned} k^1, k^2 &\sim \mathcal{N}_+(0.4, 0.4), & A &\sim \mathcal{N}_+(1,1), & t_0^1, t_0^2 &\sim \mathcal{N}_+(0.3, 0.3) \\ \mu_{\text{corr}}^1, \mu_{\text{corr}}^2 &\sim \mathcal{N}_+(3,3), & \mu_{\text{err}} &\sim \mathcal{N}_+(1,1), & & \\ \sigma_{\text{corr}} &\sim \mathcal{N}_+(1,1), & \sigma_{\text{err}} &\sim \mathcal{N}_+(1,1), & & \end{aligned} \tag{5.9}$$

Each model includes a component to simulate contaminant processes: it is taken that there is a 98% chance that the observed response is due to the LBA process described above, and 2% chance that it is instead due to a contaminant process uniformly distributed between 0 and 5 seconds and equally likely to give either the correct or error response.

Let it be marked that we may *a priori* expect the simple model to receive greater evidentiary support than the complex model, since the fit data were generated with no variation in parameters between conditions.

Analysis data and methods: Earlier studies

The first study to calculate the Bayesian evidence and Bayes factors for the data and models described above (Evans & Brown, 2018) did so using the arithmetic mean (AM) estimator of marginal likelihood, collecting 10^8 samples for each model and performing 10 replications. The arithmetic mean estimator is a Monte Carlo (though not a Markov Chain Monte Carlo) method for estimating an integral, which utilises the fact that, for sufficiently large n, the marginal likelihood for a model $\mathcal{M}$ with parameters Θ, given data $\mathcal{D}$, is

$$\begin{aligned} p(\mathcal{M}) &= \int_\Theta p(\mathcal{D} \mid \theta, \mathcal{M})\ p(\theta \mid \mathcal{M})\, d\Theta \\ &\approx \sum_{i=1}^{n} p(\mathcal{D} \mid \theta_i, \mathcal{M})\ p(\theta_i) \end{aligned} \tag{5.10}$$

when the θ_i are randomly sampled from Θ. The AM estimate is guaranteed to converge to the true marginal likelihood as $n \to \infty$; however, if the high-likelihood regions occupy only a small proportion of the prior parameter space (as is usually the case), the Monte Carlo sampling is unlikely to sufficiently sample this region of the space unless a great number of samples are

collected, relative to the dimensionality of the parameter space. The AM estimator is therefore prone to underestimation of the true Bayesian evidence when not enough samples have been collected.

Annis et al. (2019) applied the TI and SS estimators to the problem using four sampling schedules: 10, 20, 35 and 50 temperature rungs spaced at equal quantiles of the Beta(0.3) distribution. For (power) posterior sampling, they used the popular differential evolution Markov Chain Monte Carlo method (DE-MCMC, Turner et al., 2013) with a burn-in of 300 iterations and 700 subsequent samples for each rung.[xi] For the individual subject LBA case, they do not state how many chains were collected per rung during the MCMC sampling, though it is standard with DE-MCMC sampling to use two to three times the number of model free parameters for the number of MCMC chains (so 12-18 chains for the simple LBA model and 18-27 chains for the complex model). Further, in another test in the same paper, where the same techniques were applied to a hierarchical LBA model, Annis et al. used three times the number of model parameters; and Evans and Annis (2019), in applying TI to the same problem, used twice the number of free parameters. It is therefore assumed herein that, in the application reported in Annis et al. (2019), similar considerations were applied in choosing the number of chains collected per power posterior. Finally, we remark that ten replications were performed for each method-schedule condition to estimate the variance in the LMLEs.

Evans and Annis (2019), already mentioned, again applied TI to the model selection problem, this time with 2300 samples per chain with the first 300 discarded as burn-in. (The temperature schedules and the choice of MCMC sampler were the same as above.) In their research, they propose a further method called *thermodynamic integration via differential evolution* (TIDE) which reduces the computational burden associated with standard TI by requiring only a single chain per power posterior.[xii] TIDE was applied using up to 5000 samples per chain with 1500 samples of burn-in, and tested at 10, 20, 35 and 50 rungs.

The results and performance data (*qua* computational burden as measured by the number of posterior samples required) of these experiments form the reference against which MULTINEST's competitiveness and efficacy will be assessed.

Analysis data and methods: Present study

MULTINEST was run using the three different sampling schedules for each model, with ten independent replications performed per schedule, per model. The sampling schedules had a fixed tolerance parameter ($c_{\text{tol}} = 0.5$) and varied in the number of live points ($N_{\text{live}} = 200, 400, 1000$). As

[xi] It is common, though arguably not essential, in many applications of MCMC techniques to "thin" the chains – by, for example, using only every tenth sample – due to the high correlation of adjacent samples in a chain Gelman et al., 2013. No such thinning is reported in the applications reviewed here.

[xii] For the technically minded, this is achieved by allowing the MCMC chains to mix *between* power posteriors, i.e. temperature rungs, rather than just within them.

before, each run yielded a log-marginal likelihood estimate (LMLE), an autogenous log-marginal likelihood error estimate (ALEE), and a set of parexic posterior samples p^*. To allow for comparisons in performance efficiency, the number of posterior samples collected and the number of likelihood evaluations performed were recorded with each run. Furthermore, replicating the work done in previous studies, LMLEs were obtained using the AM estimator for both the simple and complex models, with 10^8 samples collected per run. Sixteen and seventeen AM runs were performed for the simple and complex models respectively.

To assess MULTINEST's efficacy and competitiveness in this second test case, the following is done: first, the LMLEs obtained through MULTINEST are compared to those obtained via AM estimation and in earlier studies. Second, the number of posterior samples and likelihood values required to obtain results to comparable performance figures for the rival techniques, as estimated from the report earlier studies. Finally, marginal plots of the posterior estimates are shown, for interest.

3:3 Results and discussion

Log-marginal likelihoods and Bayes factors

The LMLEs obtained in the present study are shown in figure 5.11 alongside the results of earlier studies. Focusing our attention, for now, only on the results of the investigations performed by the author, a few remarks can be made. First, as expected, the simple model receives greater support than the complex model. Second, the LMLEs returned by MULTINEST are slightly greater than those given by AM estimation in the simple model (though the AM's LMLE obtained across all 16×10^8 samples is not statistically significantly different to the LMLEs returned by MULTINEST under any sampling schedule), and significantly higher for the complex model. This discrepancy is not too concerning *vis-à-vis* MULTINEST's performance. We recall that the AM method is prone to underestimation of the marginal likelihood if too few samples are collected; the high variance in the AM LMLEs for the complex case suggests that the AM estimator has not converged, and therefore that there is probable underestimation in those LMLEs. This is consistent with the fact that the complex model has three more (1.5 times as many) parameters as the simple model, which extra dimensions require more samples for convergence to be achieved. That the complex case has not converged is further supported by the fact that the results obtained by the author for the AM method when applied to the complex case are noticeably different (by a few units of log-evidence) to those obtained in earlier studies.

Inaccuracies in the AM method notwithstanding, when broadening our considerations to include the results of earlier studies, we see that the results obtained in the present study are generally consistent with earlier results, with similar values obtained by MULTINEST for the LMLEs for both the simple and the complex models. If there is a discrepancy in the results between studies, it may be that the MULTINEST estimate appears slightly higher for complex model. Unfortunately,

quantitative comparisons cannot be made since exact figures for the LMLEs obtained are not stated in earlier studies. Without further data, it is impossible to state the exact extent or significance of this potential discrepancy; nevertheless, we will address it now as if it does exist.

We start by noting that this (presumed) discrepancy may represent either slight overestimation on the part of MULTINEST or slight underestimation on the part of the thermodynamic methods, assuming one of the two classes of technique is accurate. From the theoretical perspective, there is no particular reason to believe that MULTINEST is overestimating the Bayesian evidence; there are, however, reasons to think that the thermodynamic methods might underestimate them.

First, TI and SS rely on taking averages of the MCMC samples from the power posteriors; if the MCMC sampler fails to sufficiently converge to the posterior function, in particular under-sampling the regions of very highest density, or if too few samples are discarded as burn-in, then the resulting ML estimate will also be an underestimate. Since establishing the convergence of the samples in an MCMC chain to the posterior is not always straightforward (Gelman et al., 2013), this factor may result in slight underestimates of the ML.

Second, the division of the temperature scale into rungs introduces discretisation error, and there is no known way of setting the number of rungs in the temperature schedule optimally.[xiii] It may be that, were the number of rungs greatly increased, the slight discrepancy would disappear.

Furthermore, it is worth bearing in mind that, in the first test case, which was much simpler than the current case and where the exact marginal likelihood was known, MULTINEST displayed no bias in its estimates under the most stringent tests, whereas TI has been shown to exhibit – albeit slight – underestimation of the true marginal likelihood (despite using fewer parameter, a greater number of rungs, a greater number of samples per rung, and no MCMC sampling error; cf table 5.2).

The author wishes to point out that this reasoning is speculative, and that the proper resolution of this putative discrepancy would require further study. The discrepancy – if real – may instead result from, say, slight differences in model implementation. The important point is that, from the data presented above, there is no reason to believe that MULTINEST has failed to do its job properly; furthermore, it is also worth remembering that, in the data presented, the various methods are in general in good agreement regarding the problem at hand. With these considerations in mind, we take these results to establish the efficacy of MULTINEST in solving the model comparison problem addressed here.

Performance metrics

The performance data for the various methods across the various studies are shown in table 5.9. Using the data in this table and the results presented just above, it can be seen that MULTINEST

[xiii] Note that the thermodynamic integration method with correction is known to produce over-corrections when the changes between rungs are too sharp, e.g., when there are too few rungs (Annis et al., 2019; Friel et al., 2014).

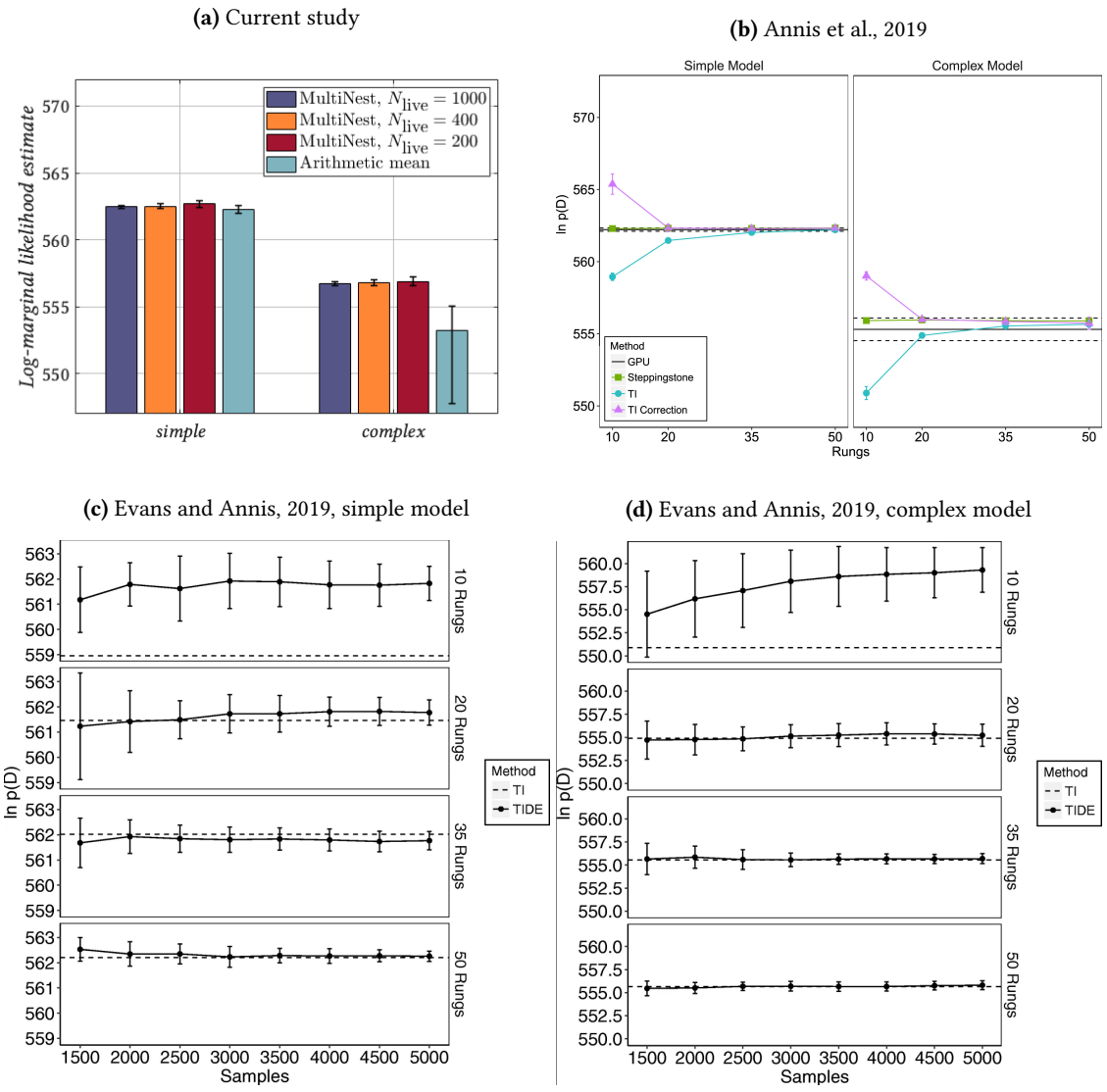


Figure 5.11. ESTIMATES OF THE LOG-MARGINAL LIKELIHOOD OBTAINED FOR BOTH THE SIMPLE AND COMPLEX MODELS. **(a)** The LMLE results obtained in the author's study using MULTINEST and the arithmetic mean estimator. For the MULTINEST results, the height of the bar is the mean of the LMLEs across the ten runs performed; the error bars represent the mean ALEE across runs. For the AM method, the bar height is the log-likelihood when combining the samples across all 16 or 17 runs performed, where 10^8 samples were collected per run (simple 562.3, complex 553.3). The error bars in this case represent the maximum and minimum LMLEs obtained per run. See table 5.8 for numerical details. **(b)** The results from the application of various ML estimation methods to the same problem from Annis et al. (2019) for comparison; note the difference in scale on the ordinate axis. The black line indicates the mean estimate across ten runs of the arithmetic mean estimator, with 10^8 (10^9) estimates performed for the simple (complex) model; the green line, the results from the steppingstone technique; the cyan line, for thermodynamic integration; and the magenta line, for thermodynamic integration with an additional correction term. The error bars (or the dashed lines, in the case of the arithmetic mean method) in all cases represent the standard deviation in the LMLEs obtained across runs. **(c)** Results for the simple model from the applications of methods in Evans and Annis (2019): TI (dashed line, representing LMLE obtained) and TIDE (solid line; points represent mean LMLEs across ten repeats with the stated number of samples per chain, and error bars are associated standard deviations). **(d)** Results for the complex model from Evans and Annis (2019), analogous to (c). In all cases, refer to §5.3:2 for more details of the sampling parameters used.

Method	*Simple model*	*Complex model*
MULTINEST ($N_{\text{live}} = 200$)	562.68 ± 0.25	556.91 ± 0.31
MULTINEST ($N_{\text{live}} = 400$)	562.49 ± 0.18	556.82 ± 0.22
MULTINEST ($N_{\text{live}} = 1000$)	562.47 ± 0.12	556.70 ± 0.14
Arithmetic mean estimator	562.26 ± 0.28, 0.25	553.26 ± 1.80, 5.48

Table 5.8. BAYESIAN EVIDENCE ESTIMATES AND ERRORS IN THE LBA. For each model and each method-schedule condition undertaken by the author. The errors are those plotted and described in figure 5.11; for the arithmetic mean estimator, the first and second error figures in each column are the ranges respectively above and below the all-samples LMLE. Note that the ALEE's returned by MULTINEST and listed here are, in this case, all greater than the standard error as estimated from the LMLE samples.

is at least as efficient, and likely much more efficient, than the rival methods: for both models, the number of likelihood evaluations performed by MULTINEST to reach its ML estimates was less than the number required by the rival methods; this is especially true when using fewer live points in sampling, a choice which – according to results presented above – still appears to result in reasonable ML estimates.

A few more comments are in order on the question of efficiency. First, TIDE may at first glance appear to do better for the complex model, but it should be borne in mind that the variance of the LMLEs is considerably greater: at the lower end of TIDE the standard errors are on the order of one unit of log-evidence for TIDE, and on the order of 0.1 units for MULTINEST.

It is also worth noting that the efficiency distinction is more sharply in favour for the simple model, where the ML estimate results agree more closely across models. This agreement perhaps make the simple model a better case for comparison. Also, building on the ideas begun earlier, the smaller difference in the complex case may indicate under-sampling in the thermodynamic methods, though this is far from certain.

Furthermore, insofar as the MCMC methods require multiple runs to establish the variance in the ML estimates, and insofar as one trusts MULTINEST's autogenous log-evidence error estimates, the efficiency of MULTINEST is greater again by a factor of the number of runs performed to establish the variance.

Finally, it should be noted that the analysis presented here is not complete, for two reasons: the uncertainty regarding the accuracy of the estimate of the complex model (which may ultimately speak against MULTINEST), and the need to fully compare the precision in the estimates achieved, once accuracy is established. Further research to establish the relative efficiency of these methods should investigate these aspects more closely.

Posterior plots

Plots of the marginal posteriors for both the simple and complex models as calculated by MULTINEST are shown in figures 5.12 and 5.13 for reference. Note how MULTINEST successfully deals

Method	*Simple model*	*Complex model*
MULTINEST ($N_{\text{live}} = 200$)	14 510 ± 1 966	35 077 ± 2 879
MULTINEST ($N_{\text{live}} = 400$)	27 000 ± 1 224	67 435 ± 5 596
MULTINEST ($N_{\text{live}} = 1000$)	66 916 ± 2 251	159 409 ± 11 516
TI, Annis et al. (2019)	252 000 to 378 000	378 000 to 576 000
SS	120 000 to 180 000	180 000 to 270 000
TIDE	108 000 to 255 000	108 000 to 255 000

Table 5.9. BAYESIAN EVIDENCE ESTIMATION PERFORMANCE DATA. The first four rows contain the mean and standard deviations of the number of likelihood evaluations performed by MULTINEST to reach convergence across the ten runs performed for each model and for each of the sampling schedules undertaken. For reference, running on an eight-core CPU, the complex model took around three minutes to fit with $N_{\text{live}} = 200$ and around seven with N_{live}=400. The fifth row contains an estimate of the number of MCMC samples collected to obtain reasonably accurate estimates using the TI method in Annis et al. (2019); this was calculated using the 1000 samples per rung figure stated in that paper, with accuracy taken as occurring at 20 rungs (so 21 temperatures; with the TI correction method being preferred over vanilla TI, cf. figure 5.11(b)) and 12 to 18 (18 to 27) chains for the simple (complex) models – i.e., two to three times as many chains as model free parameters, the standard range for the DE-MCMC sampler. The sixth row contains the same estimates for the SS method as applied in the same study, with convergence taken to occur at 10 rungs. The final row contains estimates for the number of MCMC samples required to obtain the results for the TIDE method in Evans and Annis (2019), with the lower bound set at the level where Evans and Annis state that "TIDE produces log-marginal likelihoods that are close to those obtained by standard TI and have low variance estimates (SD < 1)" (given as 3000 to 3500 samples at 35 rungs and 2000 at 50 rungs, cf. figure 5.11(c)-(d)). The upper bound is set for the TIDE results which have the lowest variance, and which seem to most closely match the results reported in studies in terms of accuracy and precision, at 50 rungs with 5000 samples per temperature.

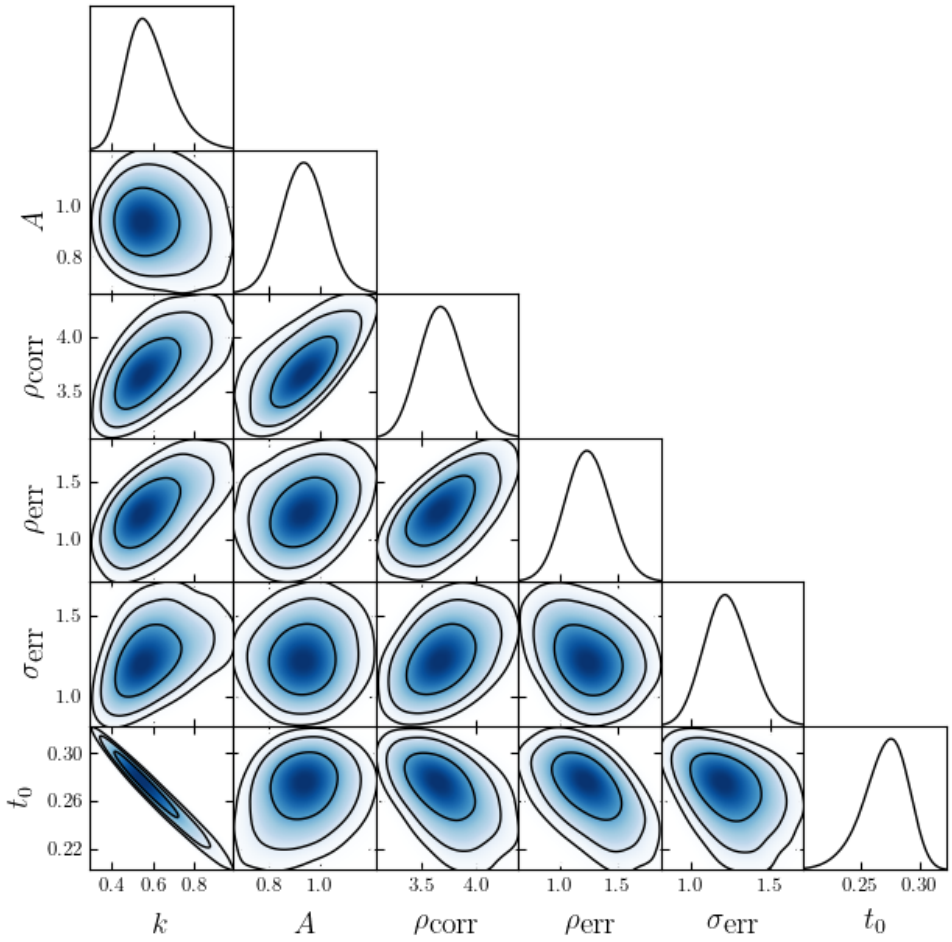


Figure 5.12. MARGINAL POSTERIOR PLOTS FOR THE SIMPLE MODEL AS ESTIMATED BY MULTINEST. The plots along the top (diagonal) edge are the marginal posteriors plots for each of the six parameters in the simple LBA model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

with the strong degeneracy between certain pairs of parameters.

4. DISCUSSION

The principal aims of this study were to establish the basic viability of MULTINEST as a method of Bayesian model inference for cognitive science for both of the two main problems of model inference, the estimation of marginal likelihoods (ML, or Bayesian evidence), and the estimation of posterior parameter distributions. Establishing viability here was taken to mean verifying MULTINEST's efficacy (its ability to accurately and precisely solve the problems of model inference) and competitiveness (its performance in comparison to other techniques with respect to

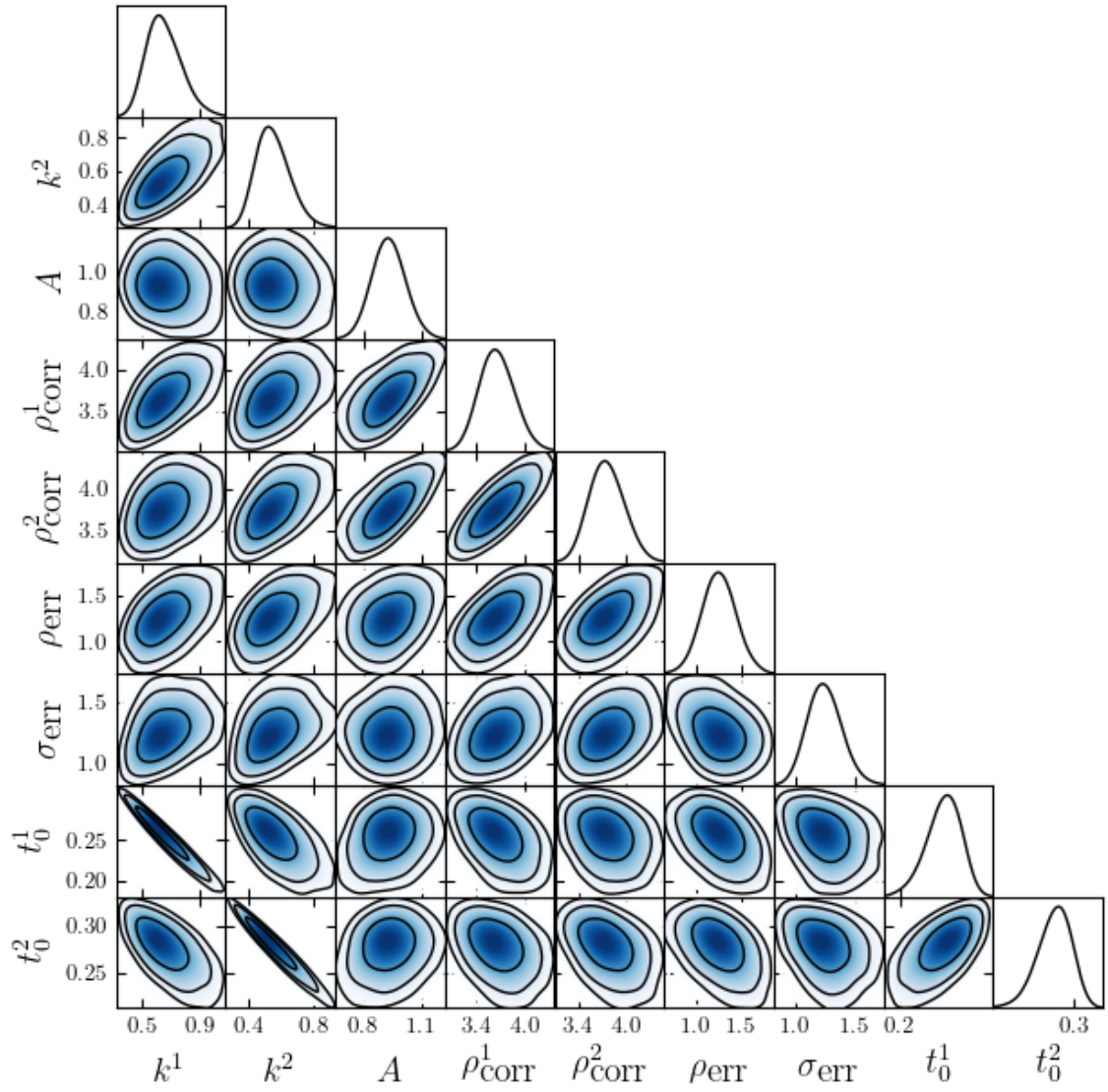


Figure 5.13. MARGINAL POSTERIOR PLOTS FOR THE COMPLEX MODEL AS ESTIMATED BY MULTINEST. The plots along the top (diagonal) edge are the marginal posteriors plots for each of the nine parameters in the complex LBA model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

practical desiderata, especially computational efficiency). Two studies were conducted to gather complementary data on MultiNest's performance to these ends.

The first, based on the work of Xie et al. (2011), took as its test case a simple Gaussian likelihood function for which the Bayesian evidence and posterior are analytically available, allowing for MultiNest's efficacy to be measured against a fundamental benchmark. The second, responding to the work of Annis et al. (2019) and Evans and Annis (2019), applied MultiNest to a previously studied toy problem of model selection involving a prototypical cognitive model, the linear ballistic accumulator (LBA) model of decision-making. In both cases, and to address competitiveness, MultiNest's performance as an estimator of marginal likelihood was compared to those of the method of thermodynamic integration (TI) and the steppingstone (SS) method. These methods, both of which rely on Markov Chain Monte Carlo (MCMC) sampling to collect posterior samples, have been prominently advocated as appropriate tools for the task of Bayesian evidence estimation in cognitive science.

The first study found that MultiNest is capable of producing highly accurate and precise estimates both of the posterior and the ML; it further found that the error estimates which MultiNest returns with each run are accurate enough to be useful. This is the first report (to the author's knowledge) which verifies MultiNest's abilities against an absolute benchmark. Furthermore, the data collected strongly suggests that MultiNest is capable of producing adequate estimates while performing far fewer likelihood evaluations than the rival methods considered.

In the second study, MultiNest was found to give results generally consistent with those of earlier studies, further demonstrating its efficacy and extending the demonstration into the cognitive domain. Once again, it was found that MultiNest was able to produce useful ML estimates while performing considerably fewer likelihood evaluations than either TI or SS. Furthermore, the problem illustrated MultiNest's ability already well-documented to easily deal with posterior distributions involving highly correlated, degenerate parameters.

Before we proceed, let us address some of the necessary shortcomings of the method and this study. First, there is, of course, no guarantee that MultiNest's ability to provide accurate and precise estimates of both the Bayesian evidence and posterior in this study's test cases extends to the general case. In the first case, both the fit data and the model being considered were exceedingly simple in nature; perhaps more complicated data, models, or combinations thereof would produce less favourable results (although it is worth bearing mind that MultiNest has been shown to be able to solve canonically extremely difficult problems of posterior estimation and *ipso facto* ML estimation). In the second case, there is no way to know for certain whether the results obtained are correct, though the comparisons to those obtained from conceptually different methods are promising. In general, there is no way to know, definitely and in advance, whether MultiNest will fully and accurately find the Bayesian evidence and posterior distribution.

Further, no matter how efficient MultiNest is, if a model is sufficiently complex and its

likelihood costly to evaluate, the level of efficiency demonstrated here may not be sufficient to render the model susceptible to analysis.

Still, insofar as these shortcomings are true of MultiNest, it is also true of all other estimator methods that would be put to the same end. Further, there are good theoretical reasons to believe that MultiNest's estimates will be accurate under a range of challenging conditions, as well as the just-mentioned practical demonstrations which serve as proof. It therefore stands to reason that, for the kinds of problem studied in cognitive science, unless and until cases of comparative failure can be established between MultiNest and alternative methods, MultiNest should be preferred.[xiv] Nevertheless, closer comparisons of MultiNest with other techniques across a broad range of conditions is warranted in future research, especially since the development of techniques of MCMC sampling and ML estimation are areas of active research.

The rest of this section will be given over to a summary of the competitive advantages and attractive features of MultiNest, especially those alluded to but not explicated in the main text. These can be summarised by saying that MultiNest is an *accurate*, *efficient*, *general purpose*, *all-in-one*, and *easy-to-use* method for model inference. The accuracy and efficiency have already been well-discussed, so we turn now to the remaining elements of the foregoing statement.

By the phrase *all-in-one*, we wish to highlight that MultiNest, like all nested sampling techniques, does both major aspects of model inference – model selection and parameter estimation – simultaneously, since in the process of calculating the Bayesian evidence, posterior samples are automatically obtained. If nothing else, it could be said that it is more straightforward to deal with a single tool than multiple interacting ones in performing model inference. However, in this all-in-one property, there *is* also an "else": other techniques of Bayesian evidence estimation intrinsically rely on the use of MCMC sampler for posterior sampling, and therefore inherit the weaknesses of the sampler used. As will be argued below, this division of the tasks of posterior sampling and evidence estimation in other techniques renders them less general-purpose and less easy to use than nested sampling techniques.

By saying that MultiNest is a general purpose method, it is meant that the technique can effectively deal with difficult posterior morphologies, such as those exhibiting multi-modality and degeneracy. Put differently, MultiNest – and advanced methods of nested sampling in general – are capable of dealing with a wide range of model likelihoods. No MCMC sampler (of which the author is aware) can necessarily do this, at least not without a great deal of technical difficulty, meaning no MCMC sampler can be relied up in advance for an arbitrary model. While DE-MCMC sampling works well in the cognitive case presented here, it is not certain that cognitive models, as they advance, will not exhibit posterior morphologies which are not amenable to this form of sampling. The use of MCMC-based methods for model inference might therefore require the

[xiv] It is worth mentioning that another nested sampling technique, PolyChord, has been developed explicitly to resolve shortcomings with MultiNest when dealing with high-dimensional problems, which may be of interest to those who wish to use large hierarchical models (Handley et al., 2015).

careful selection of an appropriate sampler with each new class of models developed. With a general purpose method, fortunately, this need not be a concern.

The previous two advantages are amplified by MultiNest's ease of use. This is best seen in contrast to operational requirements of the rival methods presented here. Those techniques of Bayesian evidence estimation which rely on the use of an MCMC sampler inherit all the technical difficulties attendant with their use. In particular, when sampling using MCMC, one must deal with assessing whether the chains have properly converged and mixed. In practice, this requires testing various sampling parameters at the MCMC level, performing multiple runs, and using various heuristic tests to try to determine convergence. This is in addition to the operational requirements of whatever ML estimation method is chosen to supervene on the MCMC sampling, which then requires the setting of a sampling schedule using abstract notions such as temperature schedules.

By contrast, using nested sampling methods, one need only plug in a few simple sampling parameters (usually, nothing more than the number of live points and the tolerance), and to provide the likelihood function, the prior, and the fit data – which are just the minimum logical requirements for performing model inference. Furthermore, these methods returns one-shot error estimates, obviating the need for multiple runs to establish accuracy in the ML estimates obtained.

Taken together, the results of this study, the results of previous studies involving MultiNest, and the practical aspects discussed herein suggest that MultiNest is a viable method – and should perhaps be the method of choice – for Bayesian model inference in cognitive science. As such, this chapter provides the elementary justification for the use of MultiNest in later chapters, where it will be used to address current problems in the theory of simple responding using models of unprecedented complexity.

—ω—

CHAPTER 6

MultiNest and models of response time: A simulation study

Abstract. One of the central tenets of my thesis is that advanced methods of Bayesian inference, especially those using nested sampling, are practical and effective for use with cognitive models and data. In the previous chapter, I took one such technique, MultiNest, and validated its efficacy in some simple but important test cases and demonstrated its competitiveness when set against rival techniques. Before applying these techniques to models of the SRT and empirical data, I aim to further ground the method's reliability for this specific application by testing it on simulated data for which the underlying generative model and parameters are exactly known.

I have chosen models that, in light of current understanding, create plausible response time distributions to imitate those obtained from the simple reaction time task. From each model, I have generated simulated data for numerous virtual subjects, with realistic inter-subject differences in parameters. Using the nested sampling approach, I show that the Bayesian evidence of each model can be estimated for every dataset in a feasible timescale, and to high accuracy and precision. The evidence is the critical quantity that enables us to determine which model actually produced the data; here, it allows us to do so with high certainty. Moreover, that certainty is greater than that obtained through other inference methods (information criterion methods) that trade precision for speed through simplifying assumptions. I therefore conclude that MultiNest is *prima facie* suitable for application to response time models for the purposes of both model selection and parameter estimation, and further bolster the evidence of its advantages over other, more broadly used methods.

1. INTRODUCTION

The paradigm of Bayesian inference (BI) has, in recent years, been increasingly adopted in the cognitive sciences for the purposes of model selection and parameter inference. The particular techniques of BI in general use in the field, however, do not represent the most advanced thus far developed. In particular, nested sampling methods should in principle (and, demonstrably so in other contexts, do in practice) provide far more accurate estimates of Bayesian evidence and richer posterior estimates, at less computational expense, than more commonly applied techniques. Two facts are of particular note: nested sampling aims, first and foremost, to accurately calculate

the Bayesian evidence for a model, with the posterior distribution calculated as a by-product; and it does this without making any simplifying assumptions about the form of the posterior. This stands in stark contrast to most BI techniques, which provide only a gross estimate of the Bayesian evidence, if any at all.

While the Bayesian evidence is the most principled and rigorous way of assessing a statistical model's ability to account for data, the computational expense of calculating the evidence, combined with the complexity of cognitive models, has long stymied the use of Bayesian inference for model selection in cognitive science. The promise of nested sampling is to feasibly allow for the proper comparison of cognitive models; properly exploited, these techniques have the potential to open up new avenues of scientific exploration.

Later chapters will begin to leverage this potential; however, it is first necessary to demonstrate that the chosen technique does, in fact, do what is claimed elsewhere when applied to the type of and scope of data typically analysed in cognitive scientific tasks and assessments. If nested sampling techniques fail to perform their job, or to outperform other techniques, then the conclusions that can be drawn from their use are either void or uninteresting. In this chapter, I therefore aim to establish *prima facie* the efficacy, feasibility, and superiority of nested sampling techniques – in particular, the MULTINEST algorithm – for cognitive models and data.

To achieve these ends, I conduct a simulation study in which MULTINEST is applied to virtual, but plausible, data of response times from a simulated simple reaction time task (response times being the bread and butter of cognitive data). These data are generated from commonly used cognitive models of the task. In particular, four models which produce plausible RT distributions are chosen: the shifted Wald distribution model, the ex-Gaussian distribution model, the ex-Wald distribution model, and Ratcliff one-choice drift-diffusion model. For each model, datasets of a size typical for a cognitive study are generated for five virtual subjects, with the model parameters chosen to reflect the empirically observed inter-subject variability in underlying neural and psychophysical parameters. Bayesian inference – for both model selection and parameter estimation – is performed on each model-and-dataset combination, and post-inference analyses conducted to determine the efficacy of the inference procedure. Specifically, I examine whether, for all the underlying generative models considered, MULTINEST can

1. correctly identify the generating model in all cases and with high certainty;
2. achieve the above inferences with the typical scale of data collected from a subject during assessment; and
3. do the above across a range of plausible observed distributions.

Having thus established the efficacy and feasibility of MULTINEST, I then examine its comparative superiority to other inference techniques. In general, comparisons must be made to two broad classes of technique: posterior-based and posterior-free techniques. The algorithmic

advantages of MultiNest over other posterior-based techniques – *viz.* those which attempt to calculate the posterior distribution, the most popular of which are based on the MCMC approach – were established in the previous chapter, so will not be further explored here. The more pressing comparison is with the posterior-free techniques; I compare the outcomes of MultiNest to the results obtained through the application of the Akaike, Bayesian, and Hannan-Quinn Information Criteria (AIC/BIC/HQIC).

2. GENERAL THEORY AND METHODS

2:1 Choice of models

Four models of response time in the SRT were chosen to form the basis of the study; these were the shifted Wald, the ex-Gaussian, the ex-Wald, the Ratcliff one-choice drift-diffusion model, and a Wald-mixture model. These models were chosen first and foremost because they have all previously been proposed and used to account for the distribution of response latencies in the SRT. All these models produce – perhaps only superficially – similar predictions in accord with the broad empirical findings regarding simple response time distribution, namely, a unimodal, positively-skewed distribution with relatively heavy tails and a peaked or rising hazard function. On the basis of their prior use and their (at least) rough accordance with empirical findings, we posit that, for the purposes of testing model inference techniques, simulated data derived from these models can be used as an adequate surrogate for real data.

Three further considerations were taken into account when choosing these models for this purpose. First, one of the models (the Ratcliff one-choice drift-diffusion model, henceforth the ROCDD) was included because it has five parameters, more than the three each of the remaining models. Its inclusion provides a test of MultiNest's ability to successfully detect overfitting and account for model complexity. Secondly, the Wald-mixture model was specifically designed to provide a further, more refined test of the same ability; the manner by which it does so will become clear later. Finally, many models in cognitive science have no explicit, closed-form expression for their PDFs; this has resulted in much recent work centred on so-called "likelihood-free" methods of posterior inference. One such model is, again, the ROCDD. Since these methods rely on only approximately calculating the likelihood, they potentially pose difficulties for various methods of posterior inference, MultiNest included. We therefore also test MultiNest's ability to cope with the type of model that has become increasingly common in cognitive science.

Finally, it is worth remarking that a natural advantage of this choice of models, from the perspective of evaluating the efficacy of the inference procedure, is that they produce – at least superficially – very similar predictions; this increases the difficulty of discriminating between the models and thus provides a more stringent test of the nested sampling approach.

For a review of these models, the reader is directed to §4.4.

2:2 Data generation procedure

Simulated data from five virtual subjects were generated for each model, using the process described presently. First, five real subjects were chosen at random from the SRT experiment described in chapter 8. Then, for each model-by-subject combination, the maximum likelihood parameters of the model for the subject's data were found using standard function minimisation methods (via MATLAB's `mle` function). Finally, synthetic data was generated for each model-by-subject data by drawing random samples from the model at the MLE parameters so found. Each dataset (one per model-by-subject combination) contained 450 samples (i.e., virtual trials), in keeping with a typical number of trials collected during a real experiment.

The fact that data were generated using accepted models of RT, and since the data-generating parameters were derived from real SRT data, it is taken that the data so generated is a relatively plausible imitation of RT data for our purposes. The plausibility of the data was further validated by checking that the first two moments of the generated distributions were in line with those expected, and also by visual inspection. Histograms of the simulated datasets are shown in figure 6.1.

Aside from producing plausible data, this procedure ensureS that there will be a balanced test of MultiNest's ability to discriminate across models: the five "trials" of MultiNest for a given model will involve a similar diversity of forms in the underlying data as in the other models.

Dataset	*Parameters*				
	μ	$\eta/10^{-3}$	τ	δ_τ	σ
RDD1	117.76	1.6306	149.51	28.75	0.011690
RDD2	186.22	0.8772	26.35	18.48	0.013046
RDD3	86.73	2.6609	161.19	1.27	0.047138
RDD4	92.40	4.0116	125.77	9.27	0.012789
RDD5	161.85	1.2596	68.79	17.72	0.011978
	μ	σ	τ	-	-
WSH1	149.42	0.020095	123.55	-	-
WSH2	91.49	0.055647	127.37	-	-
WSH3	91.00	0.055041	160.73	-	-
WSH4	87.06	0.075945	151.83	-	-
WSH5	101.24	0.046252	136.20	-	-
	μ	σ	τ	-	-
EXG1	243.76	22.704	29.203	-	-
EXG2	169.76	10.581	49.093	-	-
EXG3	202.90	12.817	48.822	-	-
EXG4	181.69	9.804	57.202	-	-
EXG5	190.96	13.697	46.481	-	-
	μ	σ / 10^{-3}	τ	-	-
EXW1	244.94	6.2987	28.032	-	-
EXW2	170.00	5.0753	48.851	-	-
EXW3	203.76	4.7487	47.963	-	-
EXW4	181.69	4.1088	57.195	-	-
EXW5	191.42	5.5088	46.022	-	-

Table 6.1. Parameters used to generate data. Values are given to five significant figures, except for a few which have been truncated to four. Note that shared symbols across different models *do not* represent the same parameter; rather, the interpretation of the parameters for each model is the same as that given in the preceding section.

Dataset	*Data sample moments*		
	Mean	*Std. dev.*	*Skewness*
EMP1	272.97	37.334	1.0066
WSH1	273.37	37.332	0.8975
EXG1	273.49	35.036	0.9462
EXW1	271.02	32.898	0.6305
RDD1	270.54	35.470	1.0433
EMP2	217.02	46.665	1.4017
WSH2	217.63	53.020	1.8996
EXG2	223.11	51.666	1.6751
EXW2	221.02	49.806	2.0738
RDD2	216.36	48.242	1.1075
EMP3	251.73	48.134	1.7055
WSH3	250.38	46.851	1.4780
EXG3	248.62	51.591	2.1462
EXW3	252.20	47.217	1.2670
RDD3	251.12	40.919	1.2487
EMP4	234.60	62.066	2.7890
WSH4	241.83	65.016	2.2624
EXG4	235.78	53.431	1.8088
EXW4	237.84	54.184	1.5714
RDD4	240.66	76.890	4.8485
EMP5	243.52	44.139	1.7743
WSH5	242.71	45.821	1.0356
EXG5	239.67	60.732	3.2774
EXW5	236.13	48.842	1.4677
RDD5	239.02	53.325	2.1162

Table 6.2. MOMENTS OF SOURCE AND GENERATED DATA. The similarity of the moments within the generated datasets corresponding to a given base subject are evident even by quick inspection (with, as expected, closer correspondence in the lower moments), as is the indistinguishably, on the basis of sample moments, of the various generated samples. These intuitions can be backed up mathematically: First, every has a significant correlation to the base empirical data, as . Second, no statistically significant difference exists between any two samples of corresponding moments, where a sample is the list of a particular moment (mean, standard deviation, or skew) drawn from a single corpus (EMP,RDD, etc.), as established by two-way Kolmogorov-Smirnoff tests. Standard deviation and skewness use unbiased estimators.

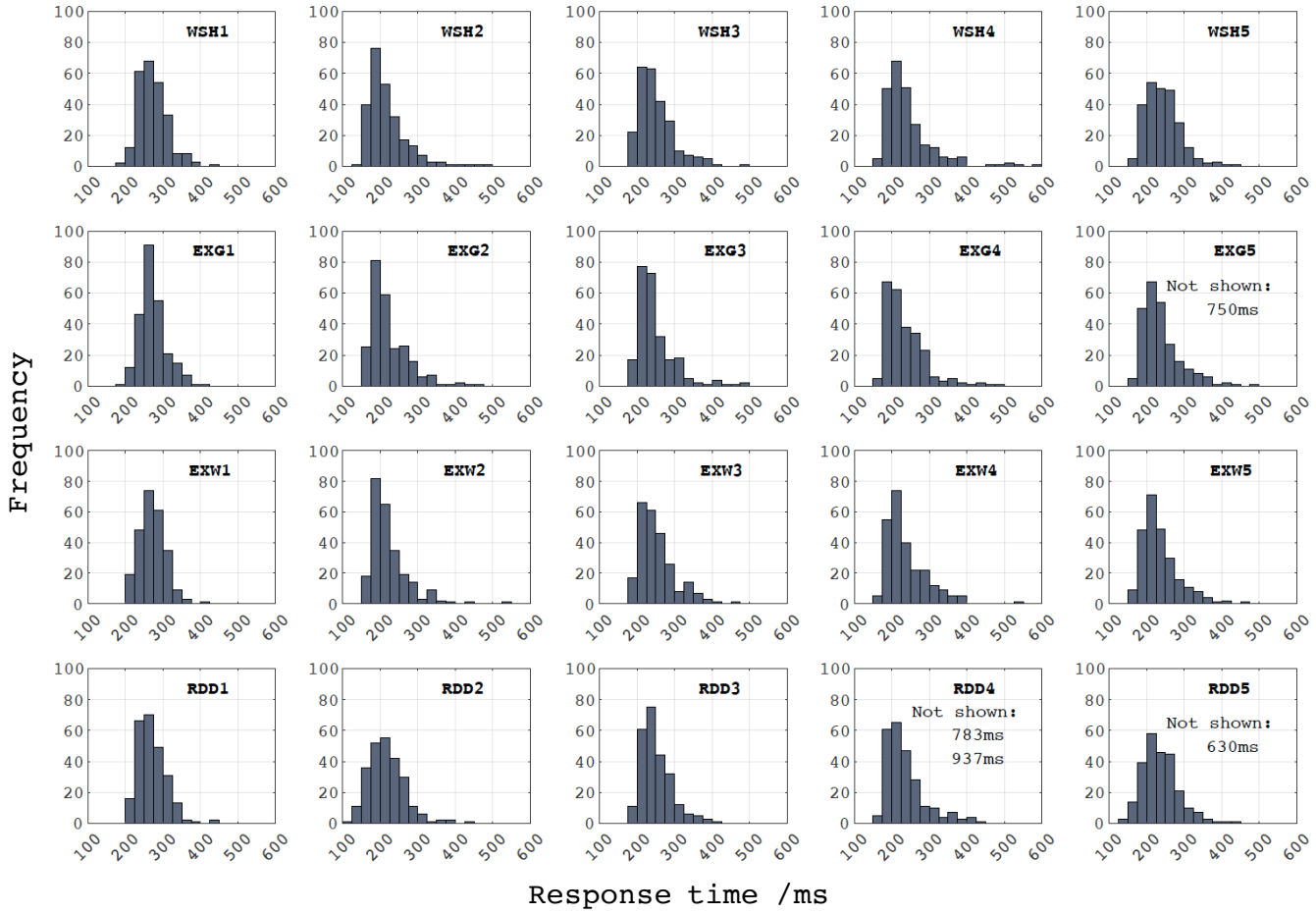


Figure 6.1. HISTOGRAMS OF SIMULATED SRT DATA. Each row contains the data from one corpus, each column, the equivalent virtual subjects across the datasets; the particular dataset is indicated on each histogram by a label, as explained in the main text. Response times are grouped into bins of width 25ms, and are shown here in the range from 100 to 600ms. Where response times were simulated that fall outside of this range, it has been noted in the individual histogram (no generated response time was below 100ms). For a comparison with real response time histograms, refer to the discussion on the empirical findings in the SRT in the previous and following chapters.

2:3 Bayesian model selection and parameter estimation; MultiNest

This section provides a brief recap of the basics of Bayesian inference. The two main branches of Bayesian inference are parameter estimation and model selection. In both cases, the notion of making an inference means calculating the weight of evidence in favour of one hypothesis over another, as determined by data. This is achieved through the application of Bayes' theorem,

$$p(H \mid D) \;=\; \frac{p(D \mid H)\; p(H)}{p(D)}\,. \tag{6.1}$$

Here H stands for some hypothesis, D stands for the data by which the hypotheses are to be judged, and $p(X)$ is the probability of (understood to be the rational degree of confidence to be placed in) the proposition X. The quantity $p(H \mid D)$, known as the *posterior*, gives the relative weight of evidence for the hypothesis H given by D, so that inferences between hypotheses can be made from ratios of the form $p(H_1 \mid D) \,/\, p(H_2 \mid D)$. Note that Bayes' theorem specifies only how to *change* one's degree of confidence in a hypothesis given data; the inference is not an absolute measure of model validity. In particular, the exact inference depends crucially on the quantity $p(H)$, known as the prior, which encodes the initial probability (or degree of credibility) of H. Unlike the remaining quantities on the right-hand side of the equation, $p(D \mid H)$ (the *likelihood*) and $p(D)$ (the *marginal likelihood*), the prior is not uniquely determined by the problem to hand. (The interpretation of these quantities will be covered in the next paragraph.) In this sense, Bayes' theorem should be considered a tool applied by a particular user, whom we shall refer to as the "Bayesian analyst."

For parameter estimation, the hypothesis H is a statement of the form, "the true values of the parameters, assuming the model M, are θ," from which we can render (6.1) as

$$p(\theta \mid D, M) \;=\; \frac{p(D \mid \theta, M)\; p(\theta \mid M)}{p(D \mid M)} \tag{6.2}$$

as applied to data D and a statistical model M with parameters θ. The prior encodes the Bayesian analyst's degree of confidence that the model's true parameters assume particular values, before the data has been accounted for; the likelihood, $\mathcal{L}(\theta) \equiv p(D \mid \theta, M)$ gives the probability of observing the data D when referred to the model's predictions at parameters θ; the posterior $\mathcal{P}(\theta)$ is the counterpart to the prior after the data has been accounted for. The quantity $\mathcal{Z} \equiv p(D \mid M)$, the marginal evidence or *Bayesian evidence* will be discussed later.

The posterior constitutes the complete Bayesian inference for the problem of parameter estimation; once calculated, it can be marginalised to directly obtain parameter estimates and intervals for each parameter. The Bayesian evidence, having no dependence on the parameters, is often ignored for this purpose. It acts only as a normalising factor, obtained by marginalising

over the likelihood

$$\begin{aligned} \mathcal{Z} &= p(D \mid M) \\ &= \int_{\theta \in \Theta} p(D \mid \Theta, M)\ p(\Theta \mid M)\ \mathrm{d}^n\theta \\ &= \int_{\theta \in \Theta} \mathcal{L}(\theta)\, \pi(\theta)\ \mathrm{d}^n\theta\,. \end{aligned} \tag{6.3}$$

By contrast, Bayesian evidence is the crucial quantity for model selection. Here, the hypothesis H is of the form, "the model which accounts for the data is M." Thus, for model selection, equation (6.1) becomes

$$p(M \mid D) = \frac{p(D \mid M)\ p(M)}{p(D)}\,. \tag{6.4}$$

As the posterior ratio $p(\theta_1 \mid D, M) \,/\, p(\theta_2 \mid D, M)$ grounds the inference from (6.2) regarding parameter values within a model, so inferences from (6.4) comparing two models are based on the model posterior ratio $p(M_1 \mid D) \,/\, p(M_2 \mid D)$.[i] By simple laws of probability, it can be seen that

$$\frac{p(M_1 \mid D)}{p(M_0 \mid D)} = \frac{p(D \mid M_1)}{p(D \mid M_0)} \frac{p(M_1)}{p(M_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0} \frac{p(M_1)}{p(M_0)} = B_{10} \frac{p(M_1)}{p(M_0)} \tag{6.5}$$

where $p(M_i)$ is the Bayesian analyst's prior belief in model i. That is to say, within a prior-weighting factor, the degree of support the data provides to one model over another is given by the ratio of their Bayesian evidences. This critical ratio is known as the *Bayes factor*.

(A quick note on jargon: While prior and posterior distributions over both models and parameters exist technically, the terms are usually used to refer to those distributions over a model's parameters, in the sense of equation (6.2). Unless otherwise specified, the terms when referred to henceforth are meant in this sense.)

In general, the calculation of the evidence integral (6.3) is analytically intractable and numerically expensive. As such, standard Markov Chain Monte Carlo (MCMC) techniques often focus on making posterior inferences alone; these are obtained by by taking unnormalised samples from the posterior, the chain of samples being, at equilibrium, distributed across parameter space according to the posterior, from which the relative posterior values can be calculated and the inference made. This, of course, means foregoing the valuable inferences that we might wish to make *between* models. When they *are* applied for the calculation of Bayesian evidence, MCMC methods are at least an order of magnitude more costly than when applied for parameter estimation; the already computationally-intensive process therefore makes evidence calculation practically unfeasible in many cases.

However, methods in a newer class, falling under the rubric of *nested sampling*, provide

[i] The Bayesian evidence is also known as the *model likelihood*, which nomenclature can be understood by referring to and comparing equations (6.2) and (6.4).

accurate estimates of the Bayesian evidence (and posterior inferences as a direct by-product), while being orders of magnitude more computationally efficient than their MCMC counterparts. MULTINEST is amongst the most advanced, well-studied, and widely employed of these nested sampling techniques. In summary, MULTINEST works by efficiently sampling parameter space to find successively smaller regions which contain ever-higher values of the product $\mathcal{L}(\theta)\,\pi(\theta)$, since this quantity is the determinant of both the posterior and the Bayesian evidence. The regions are estimated to conform to isolikelihood contours in parameter space; combining the sampled likelihoods and information from the estimated contours allows for the transformation of the multi-dimensional evidence integral (6.3) into a simple – and numerically speaking, thoroughly tractable – one-dimensional form. The sampling procedure makes no simplifying assumptions about the form of the posterior, and is therefore capable of dealing with posteriors exhibiting multiple modes and degeneracies. This new plane of performance affords the possibility of bringing the powerful statistical tools of Bayesian inference fully to bear on the often difficult-to-assess models and data used in cognitive science. Further details of the technique can be found in chapter 2.

From the above, it is clear that, to perform Bayesian inference – to find the posterior and Bayesian evidence of a model – we must provide three things: the data against which the models will be judged; a means of calculating the model's likelihood for that data and at specified parameters; and a prior over the model's parameters. The data was described in §6.2:2, and the general background of the models has also already been covered in chapter 4. We now turn our attention to the calculation of likelihoods and the choice of priors.

2:4 Prior specifications

The prior expresses the Bayesian analyst's confidence in each of the hypotheses under consideration. The scheme of Bayesian inference presented above requires that a prior be provided, but does not offer any formal guidance on how it should be set, except to say that it must be expressed as a probability distribution over the hypotheses. Justifications for the initial choice of prior must therefore be made independently, taking into account – as far as possible – any extant relevant information. In our case, for any given prior, the hypotheses under consideration are the "true" parameter values of the model. It is therefore required that, for each model, a choice be made regarding a reasonable range of parameter values to consider, and a distribution over those ranges be decided and justified.

Under variation of their parameters, each of our models is capable of producing a range of outcome distributions, some obviously wildly different to a response time distribution. (There is no mathematical impediment to producing negative values from a shifted Wald or ex-Gaussian distribution, or RT values in the billions of years.) Moreover, many of the parameters are capable of taking on any real (or real positive) value. Since searching over an infinite parameter space is

both difficult and unnecessary, we start by considering the crucial, known quantitative features of response times distributions that we wish our models to capture, and their relationship to model parameters.

The general approach we shall take springs from the observations stated earlier: that simple response time distributions of the kind we are simulating tend to have a mean in the range of 190-350ms, a standard deviation in the range 20-100ms, and moderately positive skew. We take these to be the relevant facts to incorporate into the prior belief. For each model, the model parameters will be related to these facts about the resultant distributions in order to decide on reasonable ranges and distributions over the model parameters.

Before we state the choice of priors, two further general issues are worth mentioning. The first is that the choice of prior is intrinsically entwined with the parameterisation of the model; careful consideration must therefore be given to the formal aspects of the model and their relation to prior information. The meaning of this statement will be illustrated through its application below. Second, there is the issue of prior sensitivity: while posteriors tend to converge quite quickly under a range of priors, the Bayesian evidence calculated in each case can differ quite markedly, resulting in divergent inferences on the basis of the Bayes factor. (See §2.4:2 for an extended discussion.) The inference regarding model selection is therefore buttressed by showing that the same conclusions are reached under a range of initial priors. To address this point, we will mention alternative priors for each case below.

Shifted Wald model

In terms of the parameterisation given earlier, equation (4.6), it is known that the first three moments of the shifted Wald distribution (the mean, variance, and skewness) are

$$m = \mu + \delta_t\,, \qquad v = \mu^3\sigma^2\,, \qquad \text{and} \quad k = 3\mu^{\frac{1}{2}}\sigma\,. \tag{6.6}$$

That is, the mean of the distribution is simply the sum of the mean decision time μ and the non-decision time δ_t. The standard deviation, as the square root of the variance, is proportional to noise level and the root-cube of the mean decision time; slower responses therefore produce greater spreads in response time. The skewness scales directly with noise level and with the square root of the mean, so that slower responses also produce more skewed response times.

We note that the system of equations above can be solved for the parameters, if given the moments, since there are equations for three unknowns. It is therefore possible to directly calculate the parameters from the moments. Since the prior information which informs our prior belief is directly related to the moments of the resultant distributions, we can specify a prior distribution over the moments and convert that into a prior over the model parameters as required. This will be the basis of the priors that are used.

The priors which are used here are the same as those used in chapter 9, to which the reader is referred for the precise mathematical details.

2:5 Log-likelihood calculations

Given a data-set D comprising n independent data d_i, $D = \{d_i, i = 1, 2, \ldots, n\}$, the likelihood of a model M at parameters θ is calculated as

$$\mathcal{L}(\theta) \;=\; p(D \mid \theta) \;=\; \prod_{i=1}^{n} f_M(d_i; \theta) \;, \tag{6.7}$$

that is, as the product of the likelihood of each datum. The likelihood of each datum is simply the probability of observing the value recorded by d_i in the probability distribution f_M predicted by the model under consideration at θ. Since each $f_M(d_i; \theta)$ is, by necessity, less than or equal to one, the product in equation (6.7) quickly becomes unmanageably small. It is therefore more usual to work with the log-likelihood, calculated as the sum of the individual datum log-likelihoods:

$$\log \mathcal{L}(\theta) \;=\; \log\left[\prod_{i=1}^{n} f_M(d_i; \theta)\right] \;=\; \sum_{i=1}^{n} \log f_M(d_i; \theta) \;. \tag{6.8}$$

For models which have closed forms for their PDFs, then, the log-likelihood calculation is a straightforward affair: the data and parameters under consideration are passed to the relevant PDF, f_M, and calculated directly. We use this method to calculate log-likelihoods for the shifted Wald and ex-Gaussian, which are of this form, using the PDFs described in §4.4. The Ratcliff one-choice diffusion model, however, does not have a closed-form solution. We must therefore construct a method for calculating the log-likelihood in the absence of an explicit function. A similar method is also used for the ex-Wald model, on account of technical difficulties encountered with implementing the analytical solution.

To construct this method, we proceed by noting that we must approximate as closely as possible the PDF f_{ROCDD}. Recall from §4.4 that the ROCDD views each trial as being drawn from a shifted Wald distribution with randomly sampled parameters, where the drift rate parameter v is sampled from a normal distribution with mean $\bar{v}$ and variance η and the non-decision time is sampled uniformly from the range $\left[\bar{\delta}_t - s, \bar{\delta}t + s\right]$ (and the noise level σ is kept fixed from trial-to-trial. Since the distribution parameters are drawn randomly on each trial, the overall PDF must simply be the continuous mixture of all possible individual trial PDFs, weighted by the probability of each possible trial PDF's occurrence, as determined by the probability of drawing

its parameters:

$$f_{\text{RDD}}\left(t; \bar{v}, \eta, \bar{\delta}_t, s, \sigma\right) = \int_{\bar{\delta}_t - s}^{\bar{\delta}_t + s} \int_{-\infty}^{\infty} p(v)\ p(\delta_t)\, f_{\text{Wald}}(t; 1/v, \sigma, \delta t)\ \mathrm{d}v\ \mathrm{d}(\delta_t) \tag{6.9}$$

where $p(v) = f_{\text{Gauss}}(v\,;\ \bar{v}, \eta)$ and $p(\delta_t) = f_{\text{Uni}}\left(\delta_t\,;\ \bar{\delta}_t, s\right)$.

Two methods present themselves naturally for approximating the resultant PDF. The first, which we shall call the *integral method*, is to approximate the integral in equation (6.9) by taking the weighted average of numerous f_{Wald}. The second method, which we call the *sampling method*, is to sample across many trials, on each trial drawing the parameters as described above, and fitting the resultant distribution. With a sufficiently large number of trials, the distribution so fitted will converge to the true distribution.

We opt to use the second method, since, in the cognitive sciences more generally, there are many models whose predictions can only be approximated by a similar sampling procedure. (It also has the advantage of being computationally more efficient.) We will refer to log-likelihoods calculated in this manner – that is to say, approximately – as *parexic* log-likelihoods, and the estimated function $\widetilde{f}$ as the parexic likelihood function.

The log-likelihood calculation for the ROCDD comprises the following procedure:

1. A set of parameters for evaluation, θ^*, are drawn or chosen according to the prior.
2. A number $N = 5 \times 10^6$ trial samples are generated from the ROCDD at θ^*; each trial sample is computed as the sum of a Wald-distributed random variable – with each sample having an independently drawn drift rate, according the strictures of the model – the base non-decision time, and a randomly drawn decision time perturbation.
3. The set of samples is used to find an approximation $\widetilde{f}_{\text{RDD}}$ of the true likelihood function f_M via kernel density estimation (described below).
4. The log-likelihood is calculated using the approximated function $\widetilde{f}_{\text{RDD}}$.

The parexic approximation for the ex-Wald log-likelihood was arrived at using the following similar procedure:

1. A set of parameters for evaluation, θ^*, was drawn or chosen from the prior.
2. $N = 5 \times 10^6$ trial samples were generated by summing a like number of instances of Wald-distributed and exponential random variables, sampled using the appropriate parameters.
3. The parexic likelihood functions $\widetilde{f}_{\text{ExW}}$ was obtained via the same kernel density estimation used for the ROCDD.
4. The log-likelihood was calculated using $\widetilde{f}_{\text{ExW}}$.

2:6 Overview of subsequent analyses

The posterior inferences and Bayesian evidence scores obtained through the application MultiNest as described above form the basic data which will be used to determine MultiNest's efficacy and suitability in the context of cognitive modelling. The determination proceeds from multiple assessments sensitive to the principal concerns outlined in the introduction. The following sections present the results of these various assessments which feed into the overall determination, with each section devoted to a single assessment. Every section in turn consists of subsections detailing its assessment's specific aims, methods, and results, and interim discussions of those results where useful or necessary. The final section consolidates the findings of the various assessments to form the summative determination and draw general conclusions about the uses and usefulness of MultiNest.

The first two analyses seek to provide an intrinsic test of MultiNest's abilities at model selection, conducted by ascertaining the degree to which MultiNest is capable of identifying the model which generated a particular dataset. The two analyses will be labelled the "intra-subject" and "trans-subject" analyses. The intra-subject analysis concerns itself with quantifying MultiNest's basic powers of model selection at the level of the individual datasets of the virtual subjects; for each one, the inference process is assessed for its ability to discern the underlying generating model (the "generator") from a set of candidates (the "pool"). In the trans-subject analysis, the results of the first are extended to reflect the usual situation in cognitive scientific studies, in which data is collected from multiple units and considered together; this part comprises an analysis of MultiNest's capacity to adequately make a selection from a pool of model ensembles over data corpora compiled from datasets sharing a generating model. The third and final analysis compares the performance of MultiNest to another popular approach to model selection, the use of *information criteria* such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC).

3. INTRA-SUBJECT MODEL SELECTION CAPABILITIES

3:1 Aims and methods

Using log-Bayesian evidences (henceforth simply "log-evidences") obtained through the nested sampling process, it is possible to compare the members of any set of two or more distinct models which have been evaluated against the same data, *qua* how well they account for that data. Within a single such comparison, we will call the models under consideration the *candidates*; the set of candidates, the *pool*; and the model selected from the pool by the inference process as best accounting for the data, the *selectee*. MultiNest's ability to identify the best-performing model

can be estimated from a sample of comparisons by the proportion of the sample for which the selectee is the same as the generating model, when the generator is both known and in the pool for each comparison in the sample. We shall call this quantity the *accordance rate* (for the selectee and the generator are in accord); it will serve as the primary metric for assessing MULTINEST's powers of model selection. The sample of comparisons from which an accordance rate is calculated will be referred to as the *trial set*.

Accordance rates were initially calculated for trial sets consisting of comparisons performed on individual datasets; for each dataset, trial sets were formed under each of a series of increasingly difficult selection conditions, starting with what we shall call *two-way selection*, then moving on to three- and four-way selection. Here, an "n-way selection condition" (or "problem") refers to a case in which the pool from which the inference process must pick a selectee contains n candidate models. Clearly, the greater number of candidates, and the more similar they are, the more difficult the selection problem posed. For each dataset and under the restrictions placed by each n-way selection condition, the accordance rate was calculated by forming a trial set consisting of every possible pool containing the generator. The accordance rates of the various trial sets were then combined to calculate the same measure at higher levels: first, the trial set formed from all n-way comparisons against all the datasets within a corpus, and then from all n-way comparisons against all the datasets available.[ii]

Having established accordance rate as the metric of MULTINEST's powers of model selection, the problem remains of how to interpret it to arrive at a conclusion regarding MULTINEST's practical usefulness. Let us consider the problem in the abstract. We have a testee process which has performed m comparisons, of which k have the selectee and generator in accord; the principal question is, do these data provide strong evidence of the testee process's ability to correctly pick out the generator, and thus of its suitability for model comparison? To answer this question, we begin by observing that the situation may be framed as a series of Bernoulli trials, where each trial corresponds to a single selection attempt (for a particular dataset and from a particular pool) and the Bernoulli success rate corresponds to the accordance rate q (across all the datasets and pools considered). Therefore, the likelihood function for observing the data (k in accord across m attempts) for a selection process with accordance rate q ($0 \leqslant q \leqslant 1$) is given by

$$p(m, k \mid H_0) = \binom{m}{k} q^k (1 - q)^{m-k} \tag{6.10}$$

from the standard formula for the binomial distribution. In this frame, we can re-state the principal question as one of parameter estimation: To what extent do the data provide support for a model in which the accordance rate suggests the ability to accurately identify the generator, against the model that it does not (the null hypothesis)? The determination of whether MULTINEST (or

[ii] Trial sets containing a mixture of different n-way comparisons in the same datasets were not considered to avoid the complications arising from the correlations of the two-way to the three-way comparisons, etc.

indeed, any other selection process) is effective at model selection, then, can be made via the following process:

1. Select a prior distribution over the accordance rate q of the selection process.
2. Calculate the resultant posterior distribution over q given the observed accordance across a sample of selection attempts.
3. Determine the evidence in favour of the hypothesis that the accordance rate is q_1 against that that it is q_0 using the posterior ratio

$$R = \frac{p(m, k \mid q_1)}{p(m, k \mid q_0)} \tag{6.11}$$

   using the likelihood function above.

The null hypothesis that allows us to answer our principal question clearly corresponds to chance selection, which in turn is clearly formalised by fixing q_0 to

$$q_0 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \tag{6.12}$$

where there are N comparisons in the trial set and the i'th comparison was n_i-ways. The alternative hypothesis q_1 will be taken to be $q_{\max}$, the accordance rate which maximises the posterior probability.

The posterior ratio establishes the evidence in favour of the *existence* (or not) of the model selection abilities of the testee process from the evidence in the trial set. The *extent* of the superiority of the testee process can be quantified in numerous ways; this study calculates *absolute advantage*, $v_{\text{abs}} = q_{\max} - q_0$ (the expected difference in accordance rate between the testee selection process and chance selection) and *relative advantage*, $v_{\text{rel}} = (q_{\max}/q_0) - 1$ (the expected gain in accordance rate, as a proportion of the accordance rate of the chance process).

To calculate the posterior, we make use of the result that the so-called *conjugate prior* for the binomial distribution is the beta distribution; this result states that, if the prior distribution specified over q takes the form of a beta distribution with shape (hyper)parameters α and β, the posterior distribution after observing k successes in m trials will also be a beta distribution, now with parameters $\alpha + k$ and $\beta + m - k$; that is to say, α and β are updated by the number of observed successes and failures respectively. The beta distribution is a highly flexible distribution capable of taking on many different qualitative forms under variation of its parameters, thereby admitting the expression of a wide range of different prior dispositions despite being mathematically relatively straightforward.

In order to guard against false or limited conclusions arrived at through the influence of prior sensitivity, I performed the posterior comparison described above under a range of reasonable

priors, and also under a challenging prior. In particular, the analysis was repeated for three priors which – it can be argued – are uninformative:

[iii] the Bayes-Laplace, or uniform, prior ($\alpha = 1, \beta = 1$); the Haldane prior (obtained in the limit $\alpha \to 0, \beta \to 0$, approximately computed using small values $\alpha = \beta = 0.01$); and the Jeffreys prior ($\alpha = 0.5, \beta = 0.5$). The "challenging" prior expresses a bias towards assuming that the selection process is ineffective, and is given by $\alpha = \left\lfloor \frac{1}{2} q_0 m \right\rfloor, \beta = \left\lfloor \frac{1}{2} m \right\rfloor - \alpha$.[iv]

The final step in the intra-subject analysis was to compare the discrepancies in log-evidence for those cases in which an alternative model was selected over the generator.

3:2 Results

The basic results of the intra-subject branch of the analysis are found in tables 6.3-6.6, with one table per corpus. These tables contain the basic inference results obtained from MultiNest for each dataset, organised by corpus. The tables share a common format: the first column identifies the particular dataset at the base of the inference results presented in the sub-tables, with naming according to the scheme introduced in 6.2:2; the second column ranks the candidate models from highest to lowest by Bayesian evidence (within each dataset's sub-table); the third column reports the models' Bayes factors on a twice log-scale, referenced against the selectee; the fourth contains the errors in the Bayes factors; and the final column summarises the strength of evidence according to the interpretative scheme listed in 2.1. More detailed descriptions of the results for each corpus and dataset are given in the captions to the tables; the remainder of this section provides further descriptions of the results summarising and comparing across the various datasets and corpora.

Table 6.7 shows the accordance rates across two-, three- and four-way comparisons for each dataset derived from the results presented in the earlier tables; it also displays the summative accordance rates for across the three levels of comparison for each corpus and across all available datasets. We highlight here that, across all available two-way comparisons, MultiNest selects the

---

[iii] As will be recalled from §2.4:2, all attempts to craft uninformative priors arise from the "the principle of indifference," which states that every possible outcome should be assigned equal weight. The Bayes-Laplace prior arises from a perhaps naïve application of this principle, and assigns equal *probability* to every possible outcome; the Haldane prior is uniform over the outcomes on the log-odds scale (essentially from the claim that the weight that should be equalised is log-odds, not probability); and the Jeffreys prior is a more sophisticated application of the principle which demands invariance under any re-parameterisation of the problem (that is, the notion of "outcome" should not depend on the particular mathematical form of the problem).

[iv] The common parameterisation of the beta distribution is used:

$$\mathrm{Beta}(x; \alpha, \beta) = \begin{cases} [B(\alpha, \beta)]^{-1} x^{\alpha-1} (1-x)^{\beta-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $B(\cdot, \cdot)$ is the beta function. The parameters α and β may be roughly interpreted as the number of successes and failures in a series of Bernoulli trials, which justifies the setting of their values in the case of the challenging prior: it is as if half as many observations as exist in the trial set had already been made and were distributed according to chance.

generator in 45 cases out of 60 (compared to an expected 30 out of 60 for chance); across three-way comparisons, the rate is 37 in 60 (compared to 20 in 60); and across four-way comparisons, 11 in 20 (against 5 in 20).

The results of the central analysis of this section, which quantify MultiNest's powers of model selection, are presented in table 6.8. From the results therein, it can be seen that this analysis finds strong to very strong evidence that MultiNest is capable of properly performing model selection, even when it is undertaken with a prior biased towards viewing the process as ineffective. Furthermore, not only is there significant evidence that MultiNest identifies the generating model significantly better than chance, the analysis finds that .

Across the nine datasets in which a model other than the generator is selected, the average Bayes factor against the generating model with respect to the selectee is (on the twice-log scale) 1.1, which is read as "barely any" evidence against in the heuristic scheme offered by Kass and Raftery Kass and Raftery, 1995. The maximum evidence found against the generating model in any one case is for dataset RDD3; there, the Ex-Wald model is preferred to the generator by a factor of 2.76 ("positive" evidence). For all corpora, the average evidence against the generator in the cases of discrepancy are weak: 1.88 for the Wald shift corpus, 0.34 for the ex-Gaussian corpus, 1.00 for the ex-Wald corpus, and 1.46 for the RocDD corpus.

Furthermore, in those cases where the generator is not selected, there are prior mathematical reasons to expect that this could be the case, beyond the general similarity of the models. The

Dataset	*Model*	$2\ln(B_{10})$	*Error*	*Evidence against*
WSH1	Ex-Wald	0	0	Best candidate
	Ex-Gaussian	-0.88	±0.33	Barely any
	Shifted Wald	-1.88	±0.32	Barely any
	ROCDD	-2.33	±0.33	Positive
WSH2	Shifted Wald	0	0	Best candidate
	Ex-Wald	-1.86	±0.28	Barely any
	Ex-Gaussian	-2.66	±0.28	Positive
	ROCDD	-3.24	±0.30	Positive
WSH3	Shifted Wald	0	0	Best candidate
	Ex-Wald	-3.87	±0.30	Positive
	Ex-Gaussian	-4.81	±0.30	Positive
	ROCDD	-5.33	±0.32	Positive
WSH4	Shifted Wald	0	0	Best candidate
	Ex-Wald	-4.78	±0.30	Positive
	ROCDD	-4.83	±0.32	Positive
	Ex-Gaussian	-5.43	±0.20	Positive
WSH5	Shifted Wald	0	0	Best candidate
	ROCDD	-5.04	±0.29	Positive
	Ex-Wald	-5.07	±0.27	Positive
	Ex-Gaussian	-8.53	±0.27	Strong

Table 6.3. MultiNest model selection results in the datasets of the shifted Wald corpus. For a description of the table format, see the main text. MultiNest successfully identifies the shifted Wald model as the generator in four of the five datasets in this corpus.

Dataset	*Model*	$2\ln(B_{10})$	*Error*	*Evidence against*
EXG1	Ex-Gaussian	0	0	Best candidate
	Ex-Wald	-1.11	±0.35	Barely any
	ROCDD	-4.32	±0.35	Positive
	Shifted Wald	-4.67	±0.35	Positive
EXG2	Ex-Wald	0	0	Best candidate
	Ex-Gaussian	-0.33	±0.29	Barely any
	Shifted Wald	-2.29	±0.29	Positive
	ROCDD	-5.34	±0.31	Positive
EXG3	Ex-Gaussian	0	0	Best candidate
	Ex-Wald	-0.16	±0.30	Barely any
	Shifted Wald	-1.08	±0.31	Barely any
	ROCDD	-1.85	±0.32	Barely any
EXG4	Ex-Gaussian	0	0	Best candidate
	Ex-Wald	-0.13	±0.31	Barely any
	Shifted Wald	-2.02	±0.30	Positive
	ROCDD	-6.67	±0.33	Strong
EXG5	Ex-Wald	0	0	Best candidate
	Ex-Gaussian	-0.34	±0.29	Barely any
	ROCDD	-2.88	±0.31	Positive
	Shifted Wald	-6.49	±0.29	Strong

Table 6.4. MULTINEST MODEL SELECTION RESULTS IN THE DATASETS OF THE EX-GAUSSIAN CORPUS. For a description of the table format, see the main text. MULTINEST successfully identifies the ex-Gaussian model as the generator in three of the five datasets in this corpus. Note the consistent pattern of results across the datasets: the ex-Gaussian and ex-Wald models compete for the top two spots, with in every case only weak evidence suggest one model or the other; the ROCDD and shifted Wald models always come in below, often with a considerably larger gap in evidence separating from the exponential models. The closeness between the ex-Wald and ex-Gaussian models is to be expected, mathematically: as covered in the general methods section, the Wald distribution, as its shape parameter increases, tends towards the Gaussian. An ex-Wald model can therefore be made to fit an ex-Gaussian dataset by finding Wald parameters which result in a close approximation to the dataset's underlying mean and standard deviation parameters; and since such Wald parameters can *always* be found, an ex-Wald distribution can *always* be made to closely fit an ex-Gaussian dataset.

Dataset	*Model*	$2\ln(B_{10})$	*Error*	*Evidence against*
EXW1	Shifted Wald	0	0	Best candidate
	rocdd	-0.51	±0.34	Barely any
	Ex-Wald	-0.93	±0.34	Barely any
	Ex-Gaussian	-3.05	±0.34	Positive
EXW2	Ex-Gaussian	0	0	Best candidate
	Ex-Wald	-0.27	±0.29	Barely any
	Shifted Wald	-4.27	±0.29	Positive
	rocdd	-4.31	±0.31	Positive
EXW3	Ex-Wald	0	0	Best candidate
	Shifted Wald	-0.51	±0.29	Barely any
	Ex-Gaussian	-0.74	±0.29	Barely any
	rocdd	-3.95	±0.31	Positive
EXW4	Ex-Wald	0	0	Best candidate
	Ex-Gaussian	-0.15	±0.29	Barely any
	Shifted Wald	-1.57	±0.29	Barely any
	rocdd	-6.00	±0.31	Positive
EXW5	Shifted Wald	0	0	Best candidate
	Ex-Wald	-1.81	±0.29	Barely any
	Ex-Gaussian	-2.75	±0.29	Positive
	rocdd	-4.34	±0.31	Positive

Table 6.5. MultiNest model selection results in the datasets of the ex-Wald corpus. For a description of the table format, see the main text. The ex-Wald model is the selectee in two cases out of five. Continuing the commentary from the previous table, while the ex-Wald distribution can always be found which closely resembles a given ex-Gaussian, the converse is not true; since the two distributions are alike in the convolving exponential component, and unlike only in the convolved Wald or Gaussian component, for an ex-Gaussian distribution to be found which closely resembles that of a given ex-Wald, the latter's parameters must be such that the Wald component is approximately Gaussian. Where the Wald component deviates significantly from normality – say, by having high skew – there is no warrant to assume that a closely matching ex-Gaussian can be found.

Dataset	*Model*	$2\log(B_{10})$	*Error*	*Evidence against*
RDD1	ROCDD	0	0	Best candidate
	Shifted Wald	-0.23	±0.34	Barely any
	Ex-Wald	-1.61	±0.34	Barely any
	Ex-Gaussian	-3.31	±0.34	Positive
RDD2	Shifted Wald	0	0	Best candidate
	Ex-Wald	-0.87	±0.24	Barely any
	ROCDD	-1.27	±0.26	Barely any
	Ex-Gaussian	-4.50	±0.25	Positive
RDD3	Ex-Wald	0	0	Best candidate
	Ex-Gaussian	-0.25	±0.31	Barely any
	Shifted Wald	-0.80	±0.31	Barely any
	ROCDD	-2.76	±0.32	Positive
RDD4	ROCDD	0	0	Best candidate
	Ex-Wald	-6.05	±0.33	Strong
	Ex-Gaussian	-7.46	±0.32	Strong
	Shifted Wald	-8.97	±0.32	Strong
RDD5	Ex-Wald	0	0	Best candidate
	Ex-Gaussian	-0.20	±0.27	Barely any
	ROCDD	-0.36	±0.29	Barely any
	Shifted Wald	-2.98	±0.25	Positive

Table 6.6. MULTINEST MODEL SELECTION RESULTS IN THE DATASETS OF THE ROCDD CORPUS. For a description of the table format, see the main text.

Dataset	*Two-way (/3)*	*Three-way (/3)*	*Four-way (/1)*
WSH1	1	0	0
WSH2	3	3	1
WSH3	3	3	1
WSH4	3	3	1
WSH5	3	3	1
All in corpus	13 /15	12 /15	4 /5
EXG1	3	3	1
EXG2	2	1	0
EXG3	3	3	1
EXG4	3	3	1
EXG5	2	1	0
All in corpus	13 /15	11 /15	3 /5
EXW1	1	0	0
EXW2	2	1	0
EXW3	3	3	1
EXW4	3	3	1
EXW5	2	1	0
All in corpus	11 /15	8 /15	2 /5
RDD1	3	3	1
RDD2	1	0	0
RDD3	0	0	0
RDD4	3	3	1
RDD5	1	0	0
All in corpus	8 /15	6 /15	2 /5
All datasets	45 /60	37 /60	11 /20

Table 6.7. Accordance rates between MultiNest's selectee and dataset generator model. Each row represents the indicated data-set; each column represents a particular *n*-way class of comparisons. The datum in each cell of the table is number of model comparisons in which the MultiNest's selectee model from amongst the candidates was the same as the generator. We have a total of four models, and the generator must always be in the pool for the accordance rate to be meaningful; therefore, for each of the twenty data-sets, there are three possible two-way comparisons (between the generator and any one of the remaining three models), three possible three-way comparisons (between the generator and any two of the remaining three models), and one possible four-way comparison (between the generator and all the remaining models). The table also summarises the accordance rate across each corpus (i.e., each set of datasets with a shared generator) and across all the datasets.

Trial set	*Prior*	q_{obs}	q_0	q_{max}	$2 \ln R$	*Interpretation*	v_{abs}	v_{rel}
Two-way	Bayes-Laplace	0.7500	0.5000	0.7500	15.7	Very strong	0.250	0.50
	Haldane	0.7500	0.5000	0.7585	16.3	Very strong	0.259	0.52
	Jeffreys	0.7500	0.5000	0.7542	16.0	Very strong	0.254	0.51
	Challenge	0.7500	0.5000	0.6271	7.7	Strong	0.127	0.25
Three-way	Bayes-Laplace	0.6167	0.3333	0.6167	20.1	Very strong	0.283	0.85
	Haldane	0.6167	0.3333	0.6206	19.9	Very strong	0.287	0.86
	Jeffreys	0.6167	0.3333	0.6186	20.0	Very strong	0.285	0.86
	Challenge	0.6167	0.3333	0.4746	10.0	Very strong	0.141	0.42
Four-way	Bayes-Laplace	0.5500	0.2500	0.5500	8.1	Strong	0.300	1.20
	Haldane	0.5500	0.2500	0.5555	7.6	Strong	0.305	1.22
	Jeffreys	0.5500	0.2500	0.5526	7.9	Strong	0.303	1.21
	Challenge	0.5500	0.2500	0.3947	3.8	Positive	0.145	0.58

Table 6.8. RESULTS OF PARAMETER ESTIMATION OF MULTINEST'S ACCORDANCE RATE. The first column identifies the trial set over which the accordance rate was estimated; for each n-ways label, the trial set contained every possible n-way comparison which included the generator across all twenty virtual subjects' datasets (cf. the previous table). The second column identifies the prior over the accordance rate q used in the analysis. All the priors were expressed as beta distributions; for full details of their parameterisations, see the main text. The third column lists the empirically observed accordance rates, as calculated in the previous table, and the fourth, the accordance rate q_0 expected under the null hypothesis of chance selection. The fifth column details the maximum *a posteriori* accordance rate q_{max}, i.e., MULTINEST's most probable "true" accordance rate over the trial set under posterior beliefs. The sixth column lists twice the natural logarithm of the posterior ratio of q_{max} to q_0, which measures the evidence in favour of the hypothesis that MULTINEST's true accordance rate is q_{max} against that that it is q_0, while the seventh interprets the strength of evidence per the commonly utilised scale for likelihood-ratio test statistics. The final two columns list the absolute and relative gain in efficacy, as described in the main text.

3:3 Interim discussion

The principal feature of this analysis is that strong evidence is found in favour of MultiNest's capacity for model selection.

A few other features of these results are worthy of remark. First, it will be noted that the evidence in favour of MultiNest's efficacy is strongest at the level of three-way comparisons. Coupled with the conclusions of MultiNest's efficacy . Compared to the two-way analysis, there are as many samples (i.e., comparisons) from which to draw inferences regarding the process's efficacy; however, three-way selection is a more difficult selection problem than two-way selection, therefore affording greater opportunity for the method to showcase its capabilities. Four-way analysis provides an even more stringent test, and therefore would allow for even greater discrimination, but is here limited by a considerably smaller sample (20 against 60). It is suggestive, nevertheless, that while the evidence of efficacy is less certain (though still strong) in the case of four-way comparison, the estimated gain in efficacy grows stronger with each step up the n-ways ladder, in absolute as well as in relative terms: put simply, the more difficult the comparison problem, the greater the degree to which MultiNest's performance is evinced.

Second, across a range of prior distributions, we see convergence of the posteriors, with the Bayes-Laplace, Haldane, and Jeffreys priors all resulting in extremely similar posterior estimates, with the posterior estimate of the accordance rate q converging on the observed rate; even under a biased "challenge" prior that presumes efficacy no better than chance, there is a significant shift in the most probable accordance rate towards the observed rate, closing the distance by about one half. (To see this most clearly, compare the values in the v_{rel} column of table 6.8.) The convergence across a range of posteriors can be interpreted to mean that, for the conclusions encoded in the posterior, the information contained in the evidence significantly outweighs the information contained in the prior – namely, the the conclusion that MultiNest is effective, as measured by its accordance rate, should be reached from a very wide range of prior beliefs regarding the method's efficacy.

Third, as noted in the captions to various tables (6.4,6.5), the separation in evidence between the models often reflects the differences (or lack thereof) between the models mathematically. To take the example of table 6.5, the ex-Wald model can always be made to closely fit an ex-Gaussian dataset; with limited data, purely mathematical considerations demand that it ought to be difficult to distinguish an ex-Wald from an ex-Gaussian model in an ex-Gaussian dataset. Similarly, The fact that MultiNest consistently finds little evidence between the two, sometimes confusing the two in its pick of selectee, should therefore be interpreted, not as a failure of the model selection process, but as a healthy sign that the method does indeed work properly.

4. TRANS-SUBJECT MODEL SELECTION CAPABILITIES

4:1 Aims and methods

It is commonplace in cognitive science for data of a certain type to be collected across multiple units (e.g., subjects) in order to study the workings of some cognitive process. It is usually presumed that, on account of their similarity, the various units can be adequately described by a single family of statistical models, albeit with varying parameters from unit to unit. When performing model inference over such data corpora, composed of multiple independent and similar datasets, the critical question becomes, not which model from a pool of candidates best describes a particular dataset, but which *ensemble* of models best accounts for the data across the whole corpus. (An ensemble is formed by assigning a model to each dataset in corpus.) In particular, if we wish aggregate the evidence for particular models across multiple units, this can be achieved by finding the evidence in favour of the "pure" ensembles formed by assigning the same model to each unit within a corpus.

Performing inferences over ensembles has a number of advantages compared to performing inferences over individual datasets. First, the technique allows empirical evidence to be combined from a variety of sources, allowing stronger evidence for a hypothesis to be gathered in the aggregate. Furthermore, in cognitive contexts, it is not always clear that collecting more data, to increase the power of an experiment; in particular, tasks often become well-trained in individuals over time, so that early and late stages in the datasets cannot be considered to be samples drawn from the same model, as it were.

The methodology for performing ensemble comparisons over data corpora is developed in appendix C; there, it is shown that the likelihood of an ensemble could be simply calculated as the product of individual model likelihoods (that is, Bayesian evidences) for the assigned model in the ensemble for each dataset in the corpus.

The trans-subject analysis starts by gathering the datasets of those virtual subjects which share common generating models into data corpora; this results in four corpora of five units (virtual subjects) each. Inference is performed from each of the four corpora for four ensemble models, each a pure ensemble formed by assigning the same candidate model to each virtual subject. The total log-evidence for each ensemble is found from the log-evidences of the constituent units according to the method developed in appendix C.

4:2 Results and discussion

The results of the trans-subject branch of the analysis are in 6.9. The table shares the same format as those for the intra-subject analysis of the previous section. There, it can be seen that for every corpus, MultiNest selects the generating model as the best explanation of the data in aggregate

Data corpus	*Model*	$2\log(B_{10})$	*Error*	*Evidence against*
Shifted Wald	Shifted Wald	0	0	Best candidate
	Ex-Wald	-13.69	±1.47	Very strong
	ROCDD	-18.90	±1.56	Very strong
	Ex-Gaussian	-20.43	±1.47	Very strong
Ex-Gaussian	Ex-Gaussian	0	0	Best candidate
	Ex-Wald	-0.73	±1.54	Barely any
	Shifted Wald	-15.87	±1.53	Very strong
	ROCDD	-20.39	±1.63	Very strong
Ex-Wald	Ex-Wald	0	0	Best candidate
	Shifted Wald	-3.34	±1.50	Positive
	Ex-Gaussian	-3.69	±1.51	Positive
	ROCDD	-16.10	±1.58	Very strong
ROCDD	ROCDD	0	0	Best candidate
	Ex-Wald	-4.13	±1.54	Positive
	Shifted Wald	-8.58	±1.52	Strong
	Ex-Gaussian	-1.13	±1.54	Very strong

Table 6.9. MULTINEST INFERENCE RESULTS FOR PURE ENSEMBLES ON THE COMPELETE DATA CORPORA.

across the corpus. The only exception is the ex-Gaussian model, where the evidence in favour of the ex-Wald model is within the error limits of the evidence estimate; however, as was argued in the previous section, the ability of ex-Wald models to functionally emulate ex-Gaussian models means that this result is to be expected. Thus, we have further proof that MULTINEST is capable of properly selecting the correct model from cognitive-like data.

5. COMPARISON TO INFORMATION CRITERIA MODEL SELECTION

5:1 Aims and methods

Thus far, MULTINEST's ability to perform reasonable model assessments has been proven against only the lowest possible standard – that of chance. A fuller picture of the technique's capacities is provided by comparing it to alternative model selection procedures; if MULTINEST is not capable of performing better than alternative model selection methods, it can be questioned why it should be used at all. Therefore, the third analysis compares the model comparison as performed using the Bayes factor analysis to more common information criterion approaches, the Aikake information criterion (AIC), the Bayesian information criterion (BIC), and the Hannan-Quinn information criterion (HQIC). These approaches are The information criteria were calculated using the maximum likelihood estimates obtained from MULTINEST.

Dataset	*Model*	$2\ln(B_{10})$	$2\ln(\mathcal{L}_{10})$	-ΔAIC	-ΔBIC	-ΔHQIC
WSH1	Ex-Wald	0	0	0	0	0
	Ex-Gaussian	-0.88	-0.65	-0.65	-0.65	-0.65
	Shifted Wald	-1.88	-1.79	-1.79	-1.79	-1.79
	ROCDD	-2.33	-0.62	-4.62	-11.67	-7.46
WSH2	Shifted Wald	0	-1.74	0	0	0
	Ex-Wald	-1.86	-2.82	-1.08	-1.08	-1.08
	Ex-Gaussian	-2.66	-3.90	-2.16	-2.16	-2.16
	ROCDD	-3.24	0	-2.26	-9.30	-5.10
WSH3	Shifted Wald	0	0	0	0	0
	Ex-Wald	-3.87	-4.25	-4.25	-4.25	-4.25
	Ex-Gaussian	-4.81	-5.55	-5.55	-5.55	-5.55
	ROCDD	-5.33	-0.31	-4.31	-11.35	-7.14
WSH4	Shifted Wald	0	0	0	0	0
	Ex-Wald	-4.78	-5.13	-5.13	-5.13	-5.13
	ROCDD	-4.83	-0.46	-4.46	-11.50	-7.29
	Ex-Gaussian	-5.43	-6.14	-6.14	-6.14	-6.14
WSH5	Shifted Wald	0	0	0	0	0
	ROCDD	-5.04	-0.35	-4.35	-11.39	-7.18
	Ex-Wald	-5.07	-4.79	-4.79	-4.79	-4.79
	Ex-Gaussian	-8.53	-8.76	-8.76	-8.76	-8.76

Table 6.10. COMPARISON OF RESULTS FROM VARIOUS MODEL SELECTION METHODS IN THE DATASETS OF THE SHIFTED WALD CORPUS. For a description of the table format, see the main text.

5:2 Results

The results from the comparison to the information criterion approaches are shown in tables 6.10-6.13. For certain individual subjects, it appears that an information criterion approach (especially the AIC) might be doing better; however, when multiple datasets are combined into corpora, MULTINEST is consistently capable of selecting the generator from the pool of candidate models, a feat no information-criterion approach tested was able to perform. Since the aggregate data is more informative than the individual data, this suggests that MULTINEST is performing a better job of model comparison than any of these more basic approaches.

All in all, then, the results of this chapter – especially when considered in light of the results of the previous chapter, and the reviews in chapter 3 – give *prima facie* good reasons to believe that MULTINEST is capable of performing model selection with cognitive-like data of the kind it will be applied to later.

—ω—

Dataset	*Model*	$2\ln(B_{10})$	$2\ln(\mathcal{L}_{10})$	-ΔAIC	-ΔBIC	-ΔHQIC
EXG1	Ex-Gaussian	0	0	0	0	0
	Ex-Wald	-1.11	-1.01	-1.01	-1.01	-1.01
	ROCDD	-4.32	-2.67	-6.67	-13.71	-9.50
	Shifted Wald	-4.67	-5.27	-5.27	-5.27	-5.27
EXG2	Ex-Wald	0	0	0	0	0
	Ex-Gaussian	-0.33	-0.45	-0.45	-0.45	-0.45
	Shifted Wald	-2.29	-3.21	-3.21	-3.21	-3.21
	ROCDD	-5.34	-1.35	-5.35	-12.39	-8.18
EXG3	Ex-Gaussian	0	-2.55	-0.60	-0.60	-0.60
	Ex-Wald	-0.16	-1.96	0	0	0
	Shifted Wald	-1.08	-3.14	-1.19	-1.19	-1.19
	ROCDD	-1.85	0	-2.04	-9.09	-4.88
EXG4	Ex-Gaussian	0	-0.18	-0.18	-0.18	-0.18
	Ex-Wald	-0.13	0	0	0	0
	Shifted Wald	-2.02	-3.27	-3.27	-3.27	-3.27
	ROCDD	-6.67	-2.22	-6.22	-13.27	-9.06
EXG5	Ex-Wald	0	-0.53	0	0	0
	Ex-Gaussian	-0.34	-2.00	-1.47	-1.47	-1.47
	ROCDD	-2.88	0	-3.47	-10.51	-6.30
	Shifted Wald	-6.49	-8.02	-7.49	-7.49	-7.49

Table 6.11. Comparison of results from various model selection methods in the datasets of the ex-Gaussian corpus. For a description of the table format, see the main text.

Dataset	*Model*	$2\ln(B_{10})$	$2\ln(\mathcal{L}_{10})$	-ΔAIC	-ΔBIC	-ΔHQIC
EXW1	Shifted Wald	0	-0.50	0	0	0
	ROCDD	-0.51	0	-3.50	-10.55	-6.34
	Ex-Wald	-0.93	-0.86	-0.36	-0.36	-0.36
	Ex-Gaussian	-3.05	-2.99	-2.50	-2.50	-2.50
EXW2	Ex-Gaussian	0	-0.44	-0.44	-0.44	-0.44
	Ex-Wald	-0.27	0	0	0	0
	Shifted Wald	-4.27	-5.02	-5.02	-5.02	-5.02
	ROCDD	-4.31	-0.66	-4.66	-11.70	-7.50
EXW3	Ex-Wald	0	-0.72	-0.14	-0.14	-0.14
	Shifted Wald	-0.51	-0.58	0	0	0
	Ex-Gaussian	-0.74	-1.39	-0.81	-0.81	-0.81
	ROCDD	-3.95	0	-3.42	-10.47	-6.26
EXW4	Ex-Wald	0	0	0	0	0
	Ex-Gaussian	-0.15	-0.47	-0.47	-0.47	-0.47
	Shifted Wald	-1.57	-2.43	-2.43	-2.43	-2.43
	ROCDD	-6.00	-1.61	-5.61	-12.65	-8.45
EXW5	Shifted Wald	0	0	0	0	0
	Ex-Wald	-1.81	-2.00	-2.00	-2.00	-2.00
	Ex-Gaussian	-2.75	-2.92	-2.92	-2.92	-2.92
	ROCDD	-4.34	-0.50	-4.50	-11.54	-7.33

Table 6.12. COMPARISON OF RESULTS FROM VARIOUS MODEL SELECTION METHODS IN THE DATASETS OF THE EX-WALD CORPUS. For a description of the table format, see the main text.

Dataset	*Model*	$2\ln(B_{10})$	$2\ln(\mathcal{L}_{10})$	-ΔAIC	-ΔBIC	-ΔHQIC
RDD1	ROCDD	0	0	-3.90	-10.94	-6.73
	Shifted Wald	-0.23	-0.10	0	0	0
	Ex-Wald	-1.61	-1.76	-1.66	-1.66	-1.66
	Ex-Gaussian	-3.32	-3.95	-3.85	-3.85	-3.85
RDD2	Shifted Wald	0	-4.63	-0.63	-0.17	-0.17
	Ex-Wald	-0.87	-4.46	-0.46	0	0
	ROCDD	-1.27	0	0	-6.58	-2.37
	Ex-Gaussian	-4.50	-7.31	-3.31	-2.85	-2.85
RDD3	Ex-Wald	0	-0.30	0	0	0
	Ex-Gaussian	-0.25	-0.99	-0.69	-0.69	-0.69
	Shifted Wald	-0.80	-0.87	-0.57	-0.57	-0.57
	ROCDD	-2.76	0	-3.70	-10.74	-6.53
RDD4	ROCDD	0	0	0	-5.00	-0.80
	Ex-Wald	-6.05	-6.04	-2.04	0	0
	Ex-Gaussian	-7.46	-11.92	-7.92	-5.88	-5.88
	Shifted Wald	-8.97	-13.46	-9.46	-7.42	-7.42
RDD5	Ex-Wald	0	-3.49	0	0	0
	Ex-Gaussian	-0.20	-4.54	-1.05	-1.05	-1.05
	ROCDD	-0.36	0	-0.51	-7.56	-3.35
	Shifted Wald	-2.98	-8.95	-5.46	-5.46	-5.46

Table 6.13. Comparison of results from various model selection methods in the datasets of the ROCDD corpus. For a description of the table format, see the main text.

CHAPTER 7

An empirical study of the simple reaction time task

Abstract. The previous two chapters established the efficacy and reliability of MultiNest when compared to alternative methods of Bayesian inference, and its *prima facie* capacity to deal with problems of inference involving the kinds of models and data typical of cognitive science experiments that use reaction time as their key independent variable. The next three chapters work towards the application of MultiNest to new models of simple responding that account for phenomena not incorporated into extant models. Within that context, the primary purpose of this chapter is to detail the empirical data which will be used in the model inference; to this end, I report the results of a simple reaction time experiment performed by 18 subjects and show that the findings are consistent with those reported in the prior literature when considered in light of the SRT's major phenomena: response time distributions, hazard functions, and the various foreperiod effects. Certain difficulties in the interpretation and analysis of the data are noted – especially where they affect the Bayesian model analysis reported in later chapters – and the methods for their resolution explicated and justified.

In doing so, I use the opportunity this chapter affords to fulfil a secondary purpose: to report and comment on an uncommon but distinct effect, mentioned little in the literature and mostly set aside for the remainder of this work, that I will refer to as *corrected responding*. Corrected responding occurs when a subject responds (in error) before the appearance of a response signal and subsequently responds correctly upon signal onset. The characteristics of trials with corrected responses, when compared to those with uncorrected responses, supports the notion of a distinction between anticipatory and reactive response processes, and may contain insights into subjects' self-monitoring of performance.

Finally, we note that the methods to describe and analyse the data used in this chapter will be instrumental in later chapters, when the output of various models of simple responding are compared to the results reported here.

1. INTRODUCTION · A VERIFICATION STUDY

The previous two chapters demonstrated the efficacy and reliability of MultiNest as a tool for performing Bayesian inference in general, and showed that it can select between competing models with simple-response-time-like (SRT-like) data. While the data used in those studies was synthetic, on the strength of those results, we will henceforth suppose that MultiNest can be fruitfully applied to make inferences from real SRT data. For the remainder of this work, then, our

attention now turns from the subject of inference to its object, namely the models and data that are to be subjected to analysis by MultiNest.

As seen in chapter 4, real response time data exhibits complex inner structure. Later chapters will develop new models which attempt to account for at least part of this complexity, before fitting and testing them against data exhibiting that complexity. In order for this purpose to be fulfilled, it is, of course, necessary to acquire real data and confirm that it does indeed possess the complex properties reported elsewhere. The description and confirmation of the complexity in the data to be used is the primary aim of this chapter. This will be achieved by reporting the results obtained from eighteen subjects performing an SRT task with a uniform FPD designed to elicit the complex foreperiod effects detailed in §4.5. We will that show that the findings are consistent with those in the prior literature, in particular considered in light of the SRT's major phenomena: RT distributions, hazard functions, the foreperiod distribution (FPD) effect, and the asymmetric sequential foreperiod (ASF) effect. This analysis of the data will also be used to answer an issue flagged in §4.5, of whether the FP effects described in that section are present at the level of individual subjects, or are artefacts of aggregating data across multiple subjects.

A secondary theme of this chapter deals with the difficulties of how within-trial behaviour is to be classified when response actions are made which do not fall neatly into the usual pattern of occurring once a short time after stimulus onset. (This problem was mentioned briefly in §4.2.) Dealing with these difficulties of data interpretation brings a miscellaneous issue to light, an uncommon but distinct effect we shall dub *corrected responding*. Corrected responding occurs when a subject responds (in error) before the appearance of a response signal and subsequently responds correctly upon signal onset. While this phenomenon will play little role in the overall argument of this thesis, it is mentioned here for interest and completeness. Specifically, it will be argued that the presence of corrected responses supports the notion of a distinction between anticipatory and reactive response processes, and may contain insights into subject's self-monitoring of performance.

2. GENERAL METHODS

2:1 Data collection

PARTICIPANTS. Data was collected from 18 subjects, of whom eleven were male and seven female; four were left-handed, with all left-handers being male. The mean age of the group was 27.0 years old, with a standard deviation of 2.7 years; the youngest participant was 23 and the eldest 35 years old. All participants were healthy and had normal or corrected-to-normal vision.

EXPERIMENTAL PROCEDURE. The subjects performed a visual simple response time task as the first in a battery of tasks. The task was designed to elicit short response latencies and to exhibit clear

foreperiod effects. The task consisted of five blocks of fifty trials, with the blocks being separated by short breaks; each trial consisted of a foreperiod (FP) between 500 and 2000 msec long, followed by a response signal (RS). Subjects were instructed to respond as quickly as possible to any onset of the RS; the task was administered on a laptop, with the FP and rs stimuli displayed on-screen and responses given by pressing the space key.

Foreperiod durations were uniformly, continuously, and randomly distributed between 500 and 2000 msecs, except for the foreperiod on the first trial of each block, which was fixed at 3000 msecs. The foreperiod stimulus was a centrally located fixation cross. The foreperiod sequence was generated only once, so that each subject did the task under precisely the same temporal structure – that is, exposed to exactly the same sequence of foreperiods.

The RS was presented for a fixed period of 800 msecs on every trial; the RS stimuli were all large, simple, centrally placed shapes set at high contrast to the field, so that they were highly visually salient. The onset and offset of the RS stimulus were discrete events, seen as the sudden appearance and disappearance of the RS stimulus replacing and being replaced by the foreperiod fixation cross. Each block within a complete run of the task used a different stimulus for the RS, and the same five stimuli were presented to each subject over the course of the task; different subjects received those stimuli in differing orders.

Note that no explicit warning signal (WS) was presented to indicate the start of trials (cf. §4.2) – rather, the start of each trial was implicitly marked by the disappearance of the RS (and the reappearance of the FP fixation cross), which would occur predictably after the 800 msec of RS presentation. As such, there was also no inter-trial interval (ITI) distinct from the FP. The salience of the RS stimuli, and the sudden nature of their onset and offset on each trial, made it unlikely that the temporal structure of the task would be made plainer to the subjects through the additional apparatus of warning signals and ITIs.

Each block was introduced with a display, lasting eight seconds, reminding the subject of the task instructions and presenting the RS stimulus specific to that block. The end of each block (except, evidently, the final one) was followed by a ten-second break, during which the subjects were displayed an instruction to rest.

Responses were given by pressing the space key on the laptop on which the task was performed. Participants were instructed to keep their finger on the key throughout each block and to make responses (when desired) by pressing the key once in the manner most comfortable and natural to them. Once informed of the presentation of the task, they were further instructed that only goal of the task would be to respond as quickly as possible to the appearance of the RS on each trial, but that preemptive responses (for example, made during the FP) would be discounted. Participants were also told that only one response was needed per trial. No feedback was given to the subjects concerning their performance on the task, and no behavioural data except for the responses was recorded.

DATA ACQUIRED. The data directly recorded during the course of the task were computer-generated timestamps of events corresponding to the stimulus onsets, stimulus offsets, and keypresses. Taken together, these form a complete record of the timecourse of the task and the relevant behaviour of the partipant.

TASK CONDITIONS, TASK CODE, AND APPARATUS. Subjects performed the task alone in a dimly lit, distractionless room seated in front of a MacBook Pro 15-inch (mid-2015 model) placed a comfortable distance from the participant. The task was coded and run in MATLAB (version 2017a) using the tools afforded by PsychToolBox (version 2.0); care was taken to ensure the most accurate timing possible given hardware and software constraints; task code will be made available for inspection.

2:2 Data pre-processing: Classifying responses and extracting response times

Once the source data has been collected in the manner described above, it is necessary to extract from it response times, the main independent variable of interest. On most trials, a single key press is made a short while after the appearance of the RS; at first glance, then, calculating the RT on a given trial is as simple as taking the difference between the time of stimulus onset on the trial under consideration and the time at which the key press was made.

However, three complications can supervene on the simple pattern (of a single press per trial shortly after the RS) outlined above, requiring a more nuanced approach to properly extract response time data from the source recordings. The cruxes of difficulty are first, assigning each key press to a stimulus; second, dealing with the possibility of no or multiple key presses being made per stimulus; and third, deciding whether to label certain key presses as likely not being the result of the response process. This section examines these causes of difficulty in the interpretation of the source data and the calculation of response time, then sets out and justifies the system used to classify the responses given on each trial and to derive response times from the source data.

Difficulties in extracting response times from source data

TRIAL ASSIGNMENT. We tackle the question trial of assignment first; to do so, let us begin by stating the problem more fully. The ultimate objective is to derive response times from the directly observed events, namely, key press times and stimulus onset/offset times. Self-evidently, to make the very notion of a *response* time sensible, we must first declare, for each key press, that *this* key press was a response to *that* stimulus onset; so, we require an *a priori* scheme by which to assign each behavioural event (each key press) to a stimulus onset (or a trial) for an RT to even be calculable. This will be true no matter the mechanism by which we construe the response to occur.

Now, we note that each key press must fall during either a foreperiod or a presentation of

the RS. The latter sort of event poses few issues in the question of trial assignment: it is adequate to suppose that a key press made during the presentation of a given RS is just a response to that RS. Key presses which occur during the foreperiod, however, cannot be so easily dealt with. There are three reasonable interpretations of a key press which falls in an FP: as a slow response to the preceding trial; as a premature response to the current trial (construed as voluntary, or at least cognitively driven, but mistaken according the rules of the task); or as accidents, due to, e.g., twitchy fingers.

We may eschew accidents for our present purpose. This may seem strange at first: accidental key presses, if they could be identified, could not be pegged to a stimulus *by definition*, since such key presses would not properly speaking be *responses* resulting from a cognitive process acting on stimuli; the concept of a response time associated to an accidental key press is intrinsically incoherent. It would therefore be desirable to exclude them from any analysis focusing on the response process. However, given the nature of the data, there is in principle no conceivable way to identify accidental key presses from the data *except by* their relative timings to other task events, including stimulus onsets – which *de facto* returns us to the requirement that each press be assigned to a particular trial.

The problem of distinguishing between slow and premature responses is solved here by setting a cut-off point during the foreperiod that marks key presses prior to the cut-off as (probable) late responses to the preceding trial, and those following the cut-off as (probable) premature responses to the trial currently underway. Here, we choose 350 msecs into the foreperiod as the threshold, which accommodates two facts. First, it is well-established that there is often a long tail of slow responses in reaction time tasks, so such responses are to be expected; on the timescale of the SRT RT distribution, this cut-off gives ample time for the long tail to make itself shown. Second, no matter their cause (about which more will be said below), premature responses in the first 350 msecs of the foreperiod seem highly unlikely, since no foreperiod was shorter than 500 msecs.

MULTIPLE RESPONSES. Once each key press is matched to a trial, it can and does happen that certain trials have multiple key presses associated to them. The number of responses given on a trial can, in practice, be classed three ways as either none, one, or more than one – the vast majority of trials having one associated response, and exceedingly few trials having more than two. Insofar as an analysis requires that each trial have a unique RT assigned to it, a decision must be made on multi-response trials about how to select a single RT from amongst those available; the choices made here in this respect will be covered below.

POTENTIALLY PREMATURE RESPONSES. Unusually quick responses also warrant special attention. Key presses occurring very soon after the start of the RS seem unlikely to be the result of a cognitive process involving the actual detection of the RS and the issuing of a key press in response, for there

is presumably a minimum amount of time required for the detection and motor response processes to occur. (And those key press which occur before RS onset, but where it is not reasonable to take the key press as a belated response to the previous RS, are certainly not "responses" in the normal sense at all.) But in most work on RT, the focus is on putatively "normal" responses, that is, those made nearly definitely in reaction to – as opposed to in anticipation of – the presentation of the stimulus. Unusually quick responses are then, at worst, mere accidents (*ipso facto* noise) and, at best, imprecise indicators of some failure or quirk in the response process being studied (illustrative hypotheses being, e.g., that they result from over-strong anticipation of stimulus onset rather than detection thereof; or from the failure of some inhibitory process).

On these very general theoretical grounds, and without committing to any particular hypothesis about the nature of unusually quick responses, it makes sense to demarcate these unusual events prior to any analysis of the data. The question that follows is how to separate the slowest events of this premature type from the quickest "normal" responses. (See the discussion of premature responses in §4.2:2.) Here, key presses made before 110 msecs of the RS have elapsed are labelled as premature responses. This threshold is based on conservative estimates of minimum signal detection time as established in finger-tapping studies (Woods et al., 2015). As such, those responses which occur before this threshold can reasonably be assumed to be either accidental, anticipatory, or otherwise exceptional in nature.

Classification scheme

Taking these considerations into account, the behaviour on each trial was classed into one of six categories: *hit*, *late*, *multiple hit*, *miss*, *premature*, and *corrected*. The meanings of these classes are as follows:

1. *Hit*: On a given trial, a single key press was made during the presentation of the RS, which key press occurred after the 110 msec threshold for premature responses but before RS cessation.

2. *Late*: A key press was made after the RS had ceased on trial k, but within the first 350 msecs of the foreperiod of trial $k + 1$. Response time was taken against stimulus onset for trial k.

3. *Multiple hit*: Multiple key presses were recorded, all within the window demarcating the "hit" or "late" classes.

4. *Miss*: No key press was given or recorded at any point in the trial after the first 350 msecs of the foreperiod had elapsed.

5. *Premature*: On a given trial, a key press (or presses) was (or were) made only during the foreperiod after the first 350 msecs of it had elapsed, or during the first 110 msecs of stimulus

presentation. Response time was taken against stimulus onset; if the key press was during the valid part of the foreperiod, the recorded RT was negative.

6. *Corrected*: A key press or presses were made which satisfied the "premature" class criteria, followed by a response which met the criteria of the "hit" or "late" classes.

Note that these classifications are mutually exclusive and completely exhaustive; note also that these classes are meant to impose as little psychological interpretation on the results as possible. To that that end, they are – as far as possible operational – in nature, that is, defined solely by the observational features of the responses.

In the analyses that follow, premature responses will be discarded; on multiple hit trials, the first response time will be taken as representative of the trial; and on corrected trials, the RT of key press which satisfies the "hit" condition will be used, for reasons that will become clear in the discussion of corrected responding in §7.6. In addition, the data from the first trial in each block is disregarded, since it is uniquely not preceded by another trial, and has a FP that is outside the range of the other trials.

We now turn our attention to inspecting the data in light of the three major empirical phenomena covered in chapter 4: the response time distribution, the foreperiod distribution effect, and the asymmetric sequential foreperiod effect.

3. RESPONSE TIME DISTRIBUTIONS AND HAZARD FUNCTIONS

3:1 Aims and methods

We begin by examining the distribution of response times, which is done in three steps. The first is simply drawing up the response time histograms for each subject, after classifying the responses and calculating the response times as described above. These are shown for in figure 7.1.

In the second step, each subject's RT distribution is also represented in hazard function form. Hazard functions were estimated using two methods of the author's own invention. The first, "kernel" method involved estimating the subject's RT PDF density and survivor function density using kernel density estimation (appendix G); the hazard function is then evaluated from the kernel density estimates using the relationship in equation (4.1). The second method is based on the empirical CDF. To begin, the complement of the empirical CDF gives the empirical survivor function $\hat{S}(t)$, defined at various discrete $t = (t_1, t_2, \ldots)$. Using the vectors $\hat{S}$ and t in the discretised version of the relationship in equation (4.2) gives extremely jagged estimates with the data here, probably due to the small number of samples per subject and the uneven spacing of the independent variable t (test distributions with known hazard functions used to generate similar sample sizes resulted in similar behaviour). The empirical survivor function data is therefore

first smoothed using a Gaussian weighted moving average filter, and new values of the survivor function interpolated at evenly spaced values of t from the smoothed data, before being input into the formula in (4.2). The hazard functions so obtained are displayed in figure 7.2.

Finally, since the individual RT data sets are small, and to attempt to get a better picture of any underlying form in the hazard functions, an aggregate hazard function is calculated. This is done using both the kernel and methods for each of two ways of grouping the subject data. Both aggregating methods start by normalising each subject's RT data by subtracting their (individual) mean RT and dividing through by their individual RT standard deviation. The first "mean individual hazard function" (MIHF) aggregating method starts by calculating the individual subject's hazard functions on the normalised data, before averaging the resulting curves to obtain the aggregate curve. The second "hazard function of group data" (HFGD) aggregating method first groups all the normalised response time data into a single sample set, and calculates the aggregate hazard function directly from this grouped sample. These aggregate hazard function are shown in figure 7.3.

Note that the estimation of hazard functions requires sufficient data density in the underlying distribution to be accurate; where data is sparse, hazard function estimates become extremely ragged. To allay this and make the general form of the hazard function clearer, the top 0.015 percentile of data – in the long tail of slow responses – is removed for each subject (or for the group data) in both hazard function analyses above.

3:2 Results and discussion

The findings regarding the response time data can be summarised by saying that, in terms of the PDF and hazard function, the RT data corresponds to what is typically found in tasks of this nature. The justification for this claim is given below.

First, the response time histograms in figure 7.1 show that stereotypical results of the kind reported in previous literature and described in §4.3 were obtained across all subjects: a large peak of early, fast responses, followed by a long tail of slow responses. As is evident even from inspection of the graphs, and as confirmed by the analysis of the distribution moments, every subject exhibits positive skew in their RT data. Furthermore, subject RT means and standard deviations are similar to those reported elsewhere (Woods et al., 2015).

We also see that the individual subject's hazard functions agree well with those obtained in previous studies for highly salient, sudden-onset stimuli: early peaks in the hazard rate followed by a drop to a lower level (cf. §4.3; compare the plots here to those in figures 4.2 and 4.3). While the individual subjects' hazard functions are a little ragged in form (figure 7.2) – very likely due to the small sample sizes of around 250 RTs per subject – the aggregate hazard function(s) (figure 7.3) shows a clear peak in hazard rate at the point of mean RT, followed by a drop to steady level, in line with the expected form. (Note again the increasing jaggedness of the estimate, especially

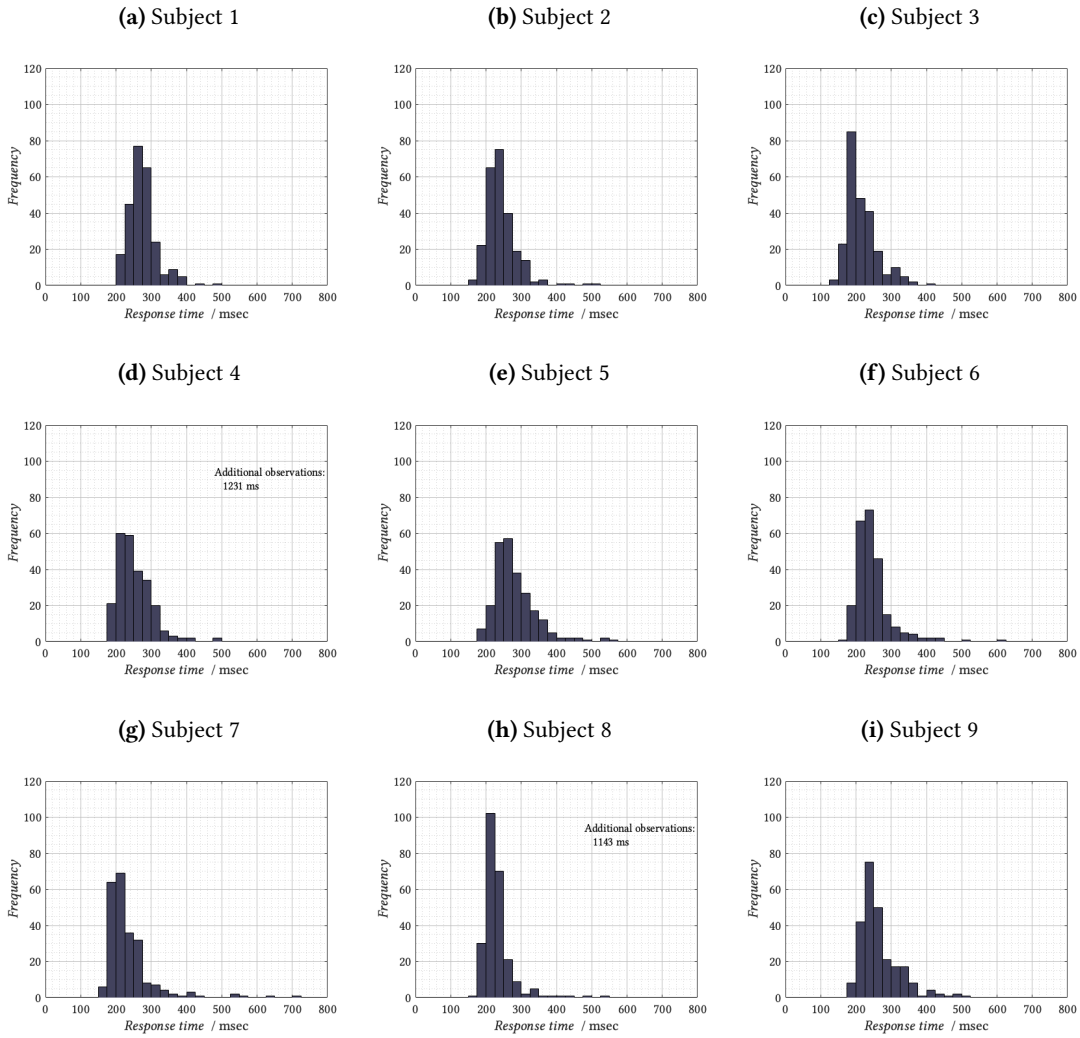


Figure 7.1. Histograms of individual subjects' reaction times (part 1 of 2).

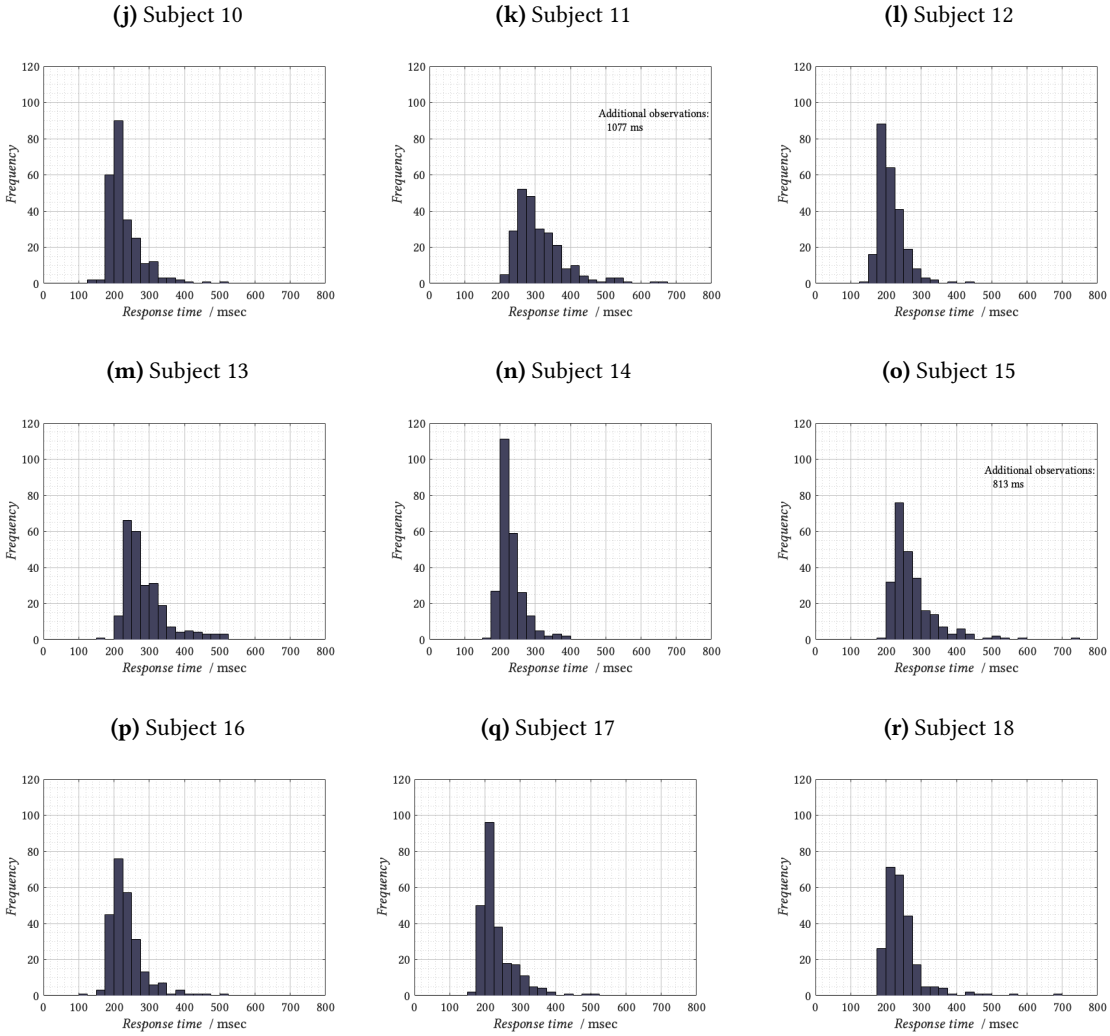


Figure 7.1. HISTOGRAMS OF INDIVIDUAL SUBJECTS' REACTION TIMES (PART 2 OF 2).

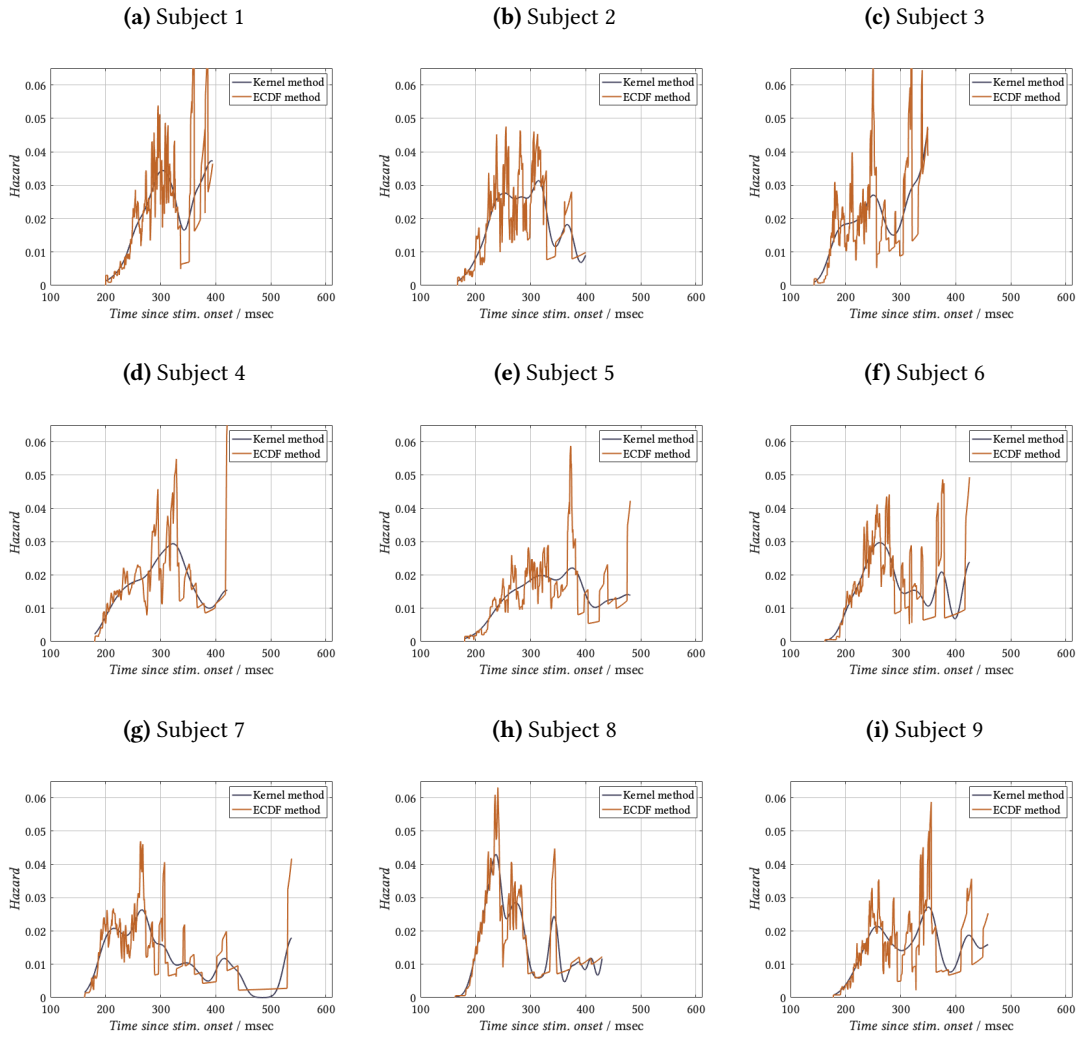
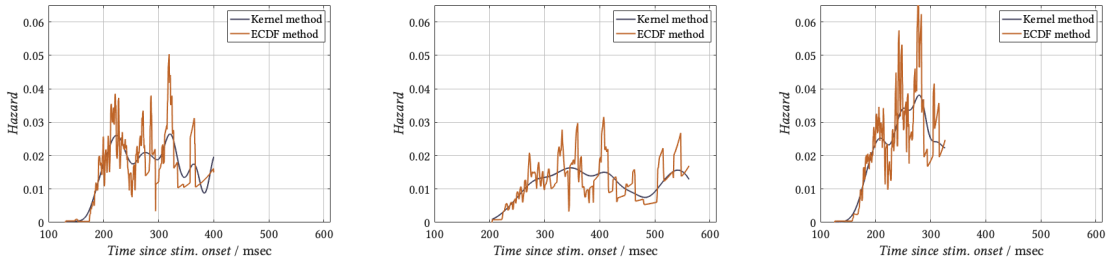
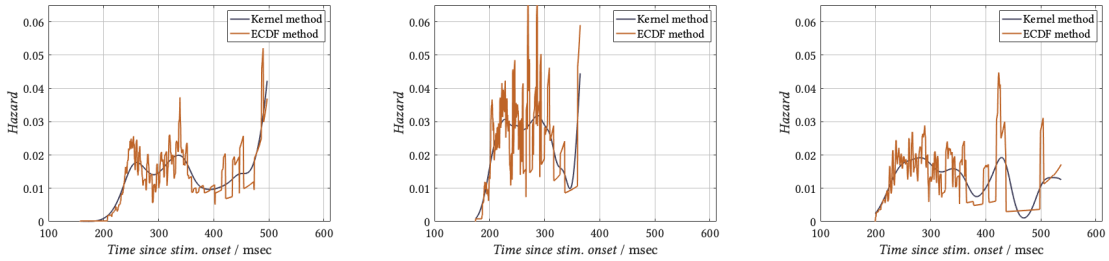


Figure 7.2. INDIVIDUAL SUBJECTS' HAZARD FUNCTIONS (PART 1 OF 2).

(j) Subject 10 **(k)** Subject 11 **(l)** Subject 12



(m) Subject 13 **(n)** Subject 14 **(o)** Subject 15



(p) Subject 16 **(q)** Subject 17 **(r)** Subject 18

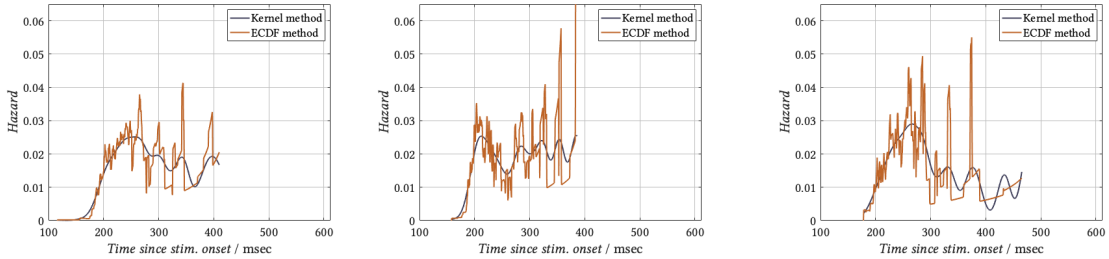


Figure 7.2. INDIVIDUAL SUBJECTS' HAZARD FUNCTIONS (PART 2 OF 2).

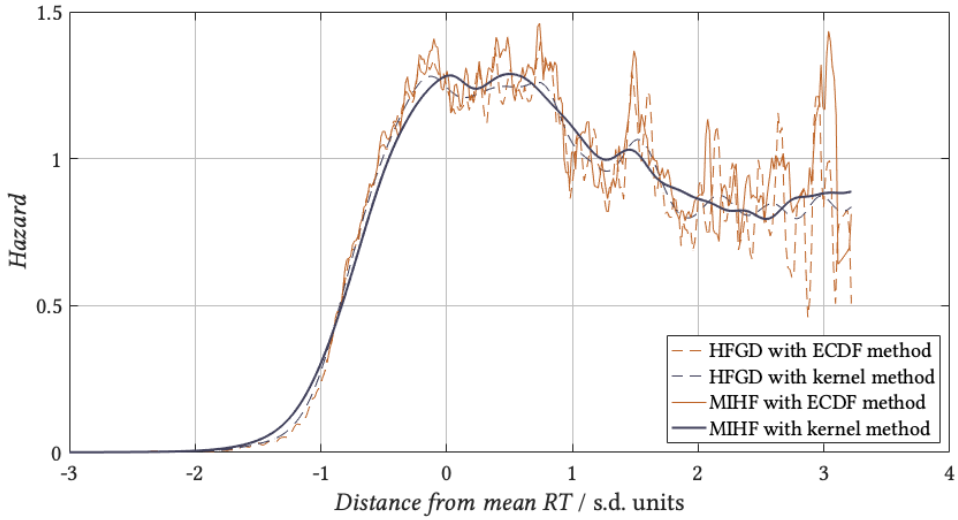


Figure 7.3. AGGREGATE HAZARD FUNCTION.

with the ECDF method, as the data become sparser – that is, towards the right hand side of the plot.)

We remark that, despite the sparse data, a number of subjects (6, 7, 8, 16 and 18 in particular) appear to have relatively clear stereotyped forms for their individual hazard functions. To have this appear in so many subjects from such few data strongly suggests that the aggregate pattern is real, rather than due to chance; given the multitude of forms a hazard function could take in principle, it is unlikely that this particular form would be repeated in such a high proportion of the subjects. Further, the relatively distinct methods of calculating the hazard function agree well with one another, suggesting that they are each working properly. Taking these observations together, we conclude that the hazard functions presented here are satisfactory representations of the hazard functions underlying the data, though the author by no means claims the methods used here are optimal.

4. FOREPERIOD DISTRIBUTION EFFECTS

4:1 Aims and methods

To examine the foreperiod distribution effect, we plot the FP curve (mean RT against foreperiod duration), for each subject and on aggregate, using two different methods. The first, "segment" method divides the foreperiod range into five equal bins or segments (500–800, 800–1100, 1100–1400, 1400–1700, and 1700–2000 msecs). For each segment, the mean RT and associated error are calculated from all the trials with a foreperiod that falls into the segment's range. The

second, "kernel smoothing" method, developed to handle this study's unusual continuous-FP design, is a little more involved, involving both kernel regression and data smoothing.

In the absence of data smoothing, the kernel regression component is performed as follows: a weighted mean of RTs is calculated at every point in the foreperiod range; the weighting of the RTs at a given point t is calculated against the FP associated with the RT, with the weight governed by an Epanechnikov kernel with a bandwidth equal to one-tenth the foreperiod range. Thus, trials with foreperiod durations close to the considered duration t have the RTs given higher weight in the averaging process, and those further away have little influence on the average.

Data smoothing is applied because RT data varies widely from trial to trial, and because our data is sparse on account of the continuous variation of the FP (meaning we only have one RT per FP value; contrast this to a discrete-FP design, where numerous RTs are accrued per possible FP). Thus, a straightforward plot of RT against foreperiod across all 250 trials performed by a subject would show a very jagged line, with any underlying shift in mean RT obscured by trial-to-trial "noise"; this noise gets propagated into a straightforward kernel regression of the raw RT data.

To overcome this, rather than performing the kernel smoothing on the raw RT data, each subject's RTs are first smoothed using a Savitzky–Golay filter, with smoothed values being interpolated on each observed FP value; thus, the set of raw RTs is replaced by a set of smoothed *pseudo*-RTs. The kernel regression described above is then applied to these pseudo-RTs.

The purpose of this method is to provide a continuous estimate of the presumably continuous FP curve when the FP can vary continuously, that is, when the critical moments (CMs) are contiguous. The presumption that the FP curve is relatively smooth, and that "local" trials around CM – those with FPs *close* to the point – therefore provide information about the value of the curve at that point. Hence, the narrow bandwidth of the regression kernel, and the form of the Epanechnikov kernel, work to ensure that the mean RT estimate is not diluted by the less relevant information from further away.

(This procedure may seen complicated; however, preliminary testing supports this approach. In this testing, synthetic RT data was produced by a variety of skewed unimodal distributions whose means and standard deviations varied over a foreperiod parameter. The synthetic RT data was matched in size to the kind used here, with FP curve relationships controlled and known. Using the above method to recover the FP curves from the data resulted in close fits to the known relationships.)

The individual subject FP curves are shown in figure 7.4. The aggregate FP curves are shown for both absolute and normalised values of the RT in figures 7.5(a) and 7.5(b); they are formed by taking either mean or the median at each point of the FP curve, averaging across the values calculated for each individual subject. To perform a quick investigation of the general form of the FP curve, quadratic and exponential functions are fitted to the aggregate FP curve formed using the kernel smoothing method from the normalised RT data and by taking the mean across the

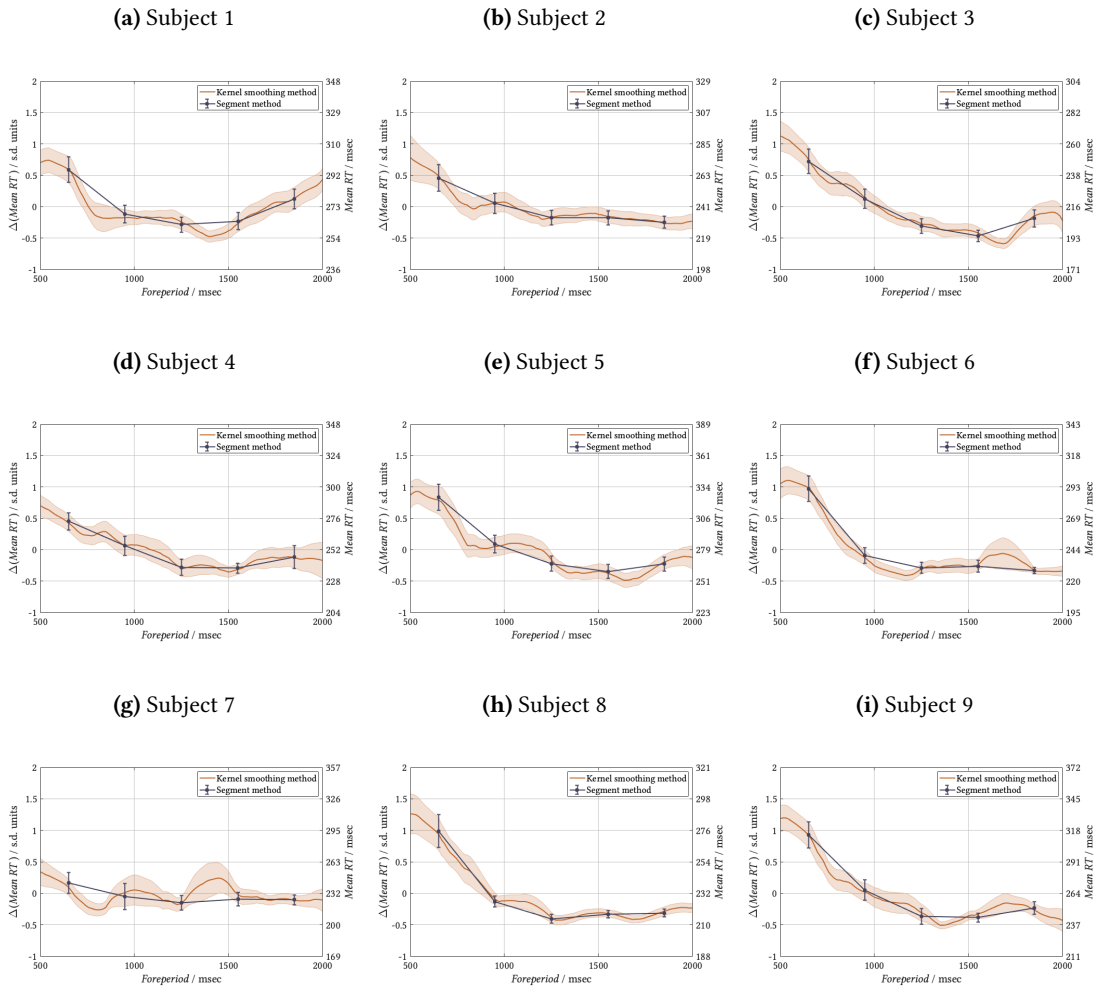


Figure 7.4. INDIVIDUAL SUBJECTS' FOREPERIOD CURVES (PART 1 OF 2). Note the presence of two scales on the ordinate axis, one for normalised and the other for absolute response time.

subjects. The results of the fit is shown in figure 7.5(c).

4:2 Results and discussion

As with the RT distributions and hazard functions, the data collected here show good agreement with past results. The aggregate curve in figure 7.5(a) shows a pronounced FPD effect, similar in form and magnitude of mean RT to the one portrayed in figure 4.5(a) for a similar experimental design (albeit with discrete, not continuous CMS). The companion curve in figure 7.5(b), calculated against the normalised RT and taking the *median* across subjects, is used to show that the FPD effect arises from a shift in the overall RT distribution at lower foreperiods, not from the addition of a few very slow foreperiods. The median averaging also ensures that the result is not due to a few subjects with particularly pronounced FP effects. Further speaking to this conclusions, the

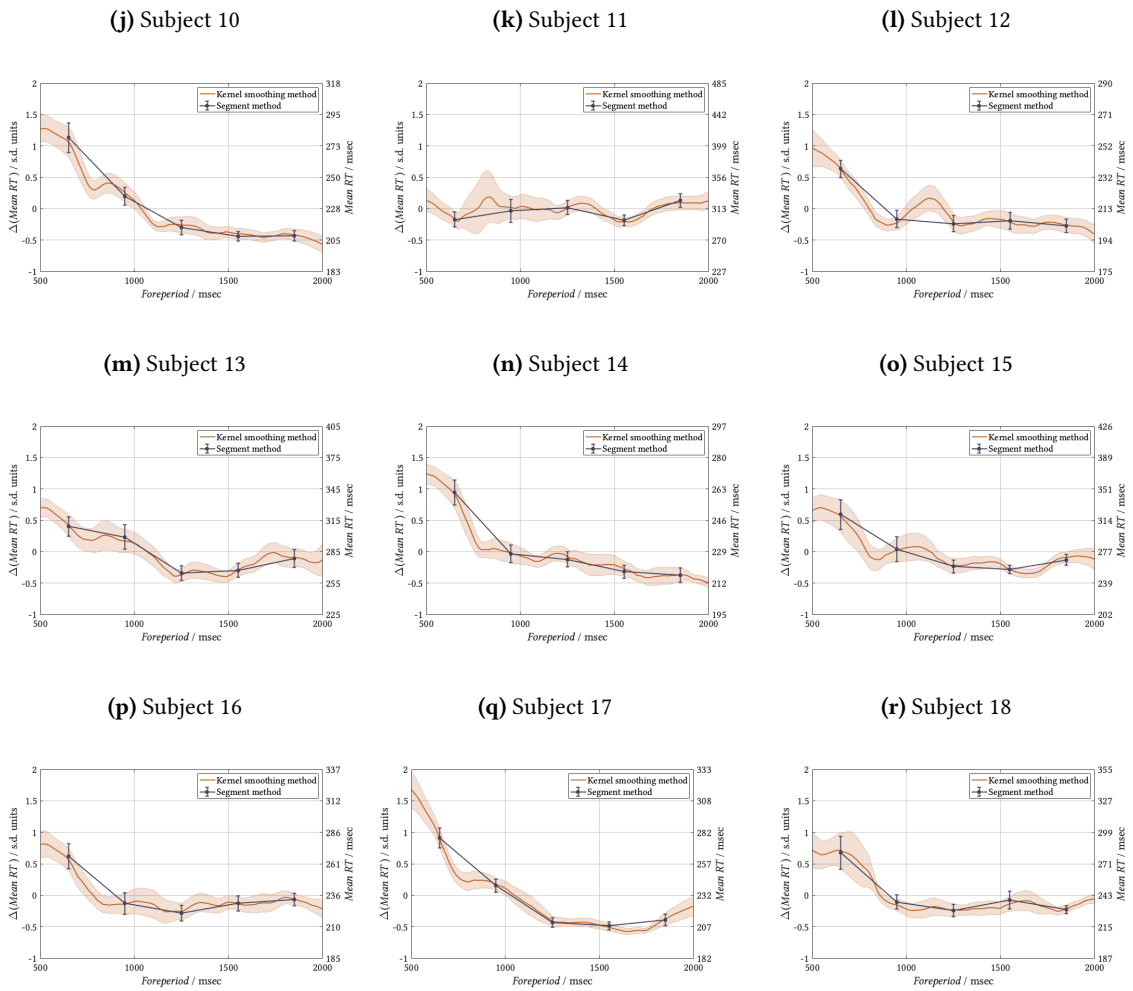


Figure 7.4. INDIVIDUAL SUBJECTS' FOREPERIOD CURVES (PART 2 OF 2). Note the presence of two scales on the ordinate axis, one for normalised and the other for absolute response time.

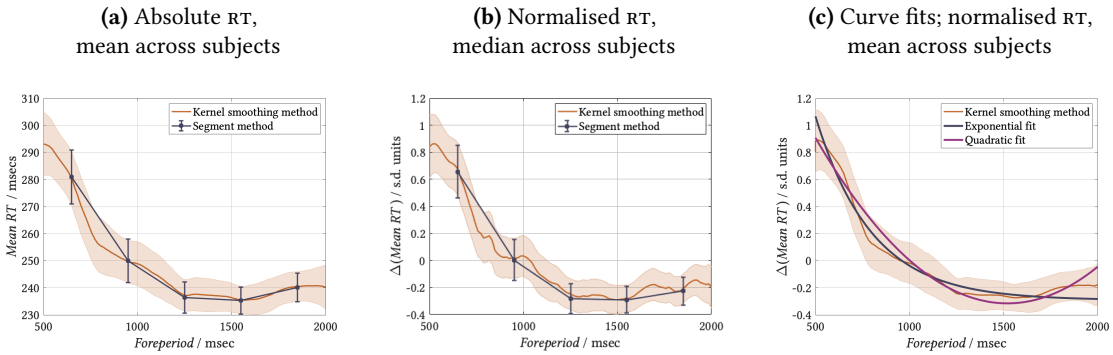


Figure 7.5. AGGREGATE FOREPERIOD CURVES.

aggregate curves are similar to many of the individual FP curves, strongly suggesting that the FP effect observed in the aggregate is due to the presence of real FPD effects in individuals, and is not an artefact of the averaging process.

Interestingly, a few subjects (7 and 11) do not show any particular foreperiod distribution effect, suggesting that they were insensitive to the temporal structure and statistics of the task. This anomaly will be mirrored in the results for the asymmetric sequential foreperiod effect, further supporting the supposition of insensitivity. Speculatively, this data may indicate that a small subset of subjects perform the SRT using a different cognitive strategy to most, or have unusually weak systems for learning the FP distribution; or they may simply have been suffering the effects of sleep deprivation. Owing to the dearth of relevant data here, understanding these anomalies will not be pursued any further.

Figure 7.5(c) shows the fits of the quadratic and exponential functions to an aggregate FP curve intended to best approximate the most typical individual-subject FP curve. Inspection of the figure shows that both functions provide reasonable fits to the FP curve. The sum of squared error (SSE) provides a rough measure of the quality of the fits; by this measure, the exponential curve fits better than that quadratic curve (SSE $= 1.05$ *vs* SSE $= 1.26$). We will therefore take the exponential curve as a reasonable description of the FP curve here, without making any claims as to its general superiority. (This result will come into use in later chapters.)

5. ASYMMETRIC SEQUENTIAL FOREPERIOD EFFECTS

5:1 Aims and methods

The asymmetric sequential foreperiod effect is investigated in two stages. The first emphasises the sequential aspect of the effect. In it, we plot what we shall call the *predecessor foreperiod curve*. This is analogous to ordinary FP curve used in the inspection of the FPD effect above, except, rather than plotting mean RT as a function of the FP on the current trial, it is plotted as a function of the FP on the preceding trial. Aside from this change in the independent variable, the methods used to analyse the data are the same as those used to produce the ordinary FP curves. The predecessor FP curves at the individual and aggregate levels are displayed in figures 7.6 and 7.7 respectively. Aggregation is performed both by considering the data from all the subjects together as a group, and by averaging (again using both the mean and the median) across the individual subjects' curves. Not every combination of these methods is shown here, but the results shown are representative.

The second stage emphasises the asymmetric aspect of the effect. Here, trials are first segregated according to their predecessor trial's FP, with three partitions being formed. The first encompasses trials whose predecessor had a FP duration between 500 and 1000 msecs; the second,

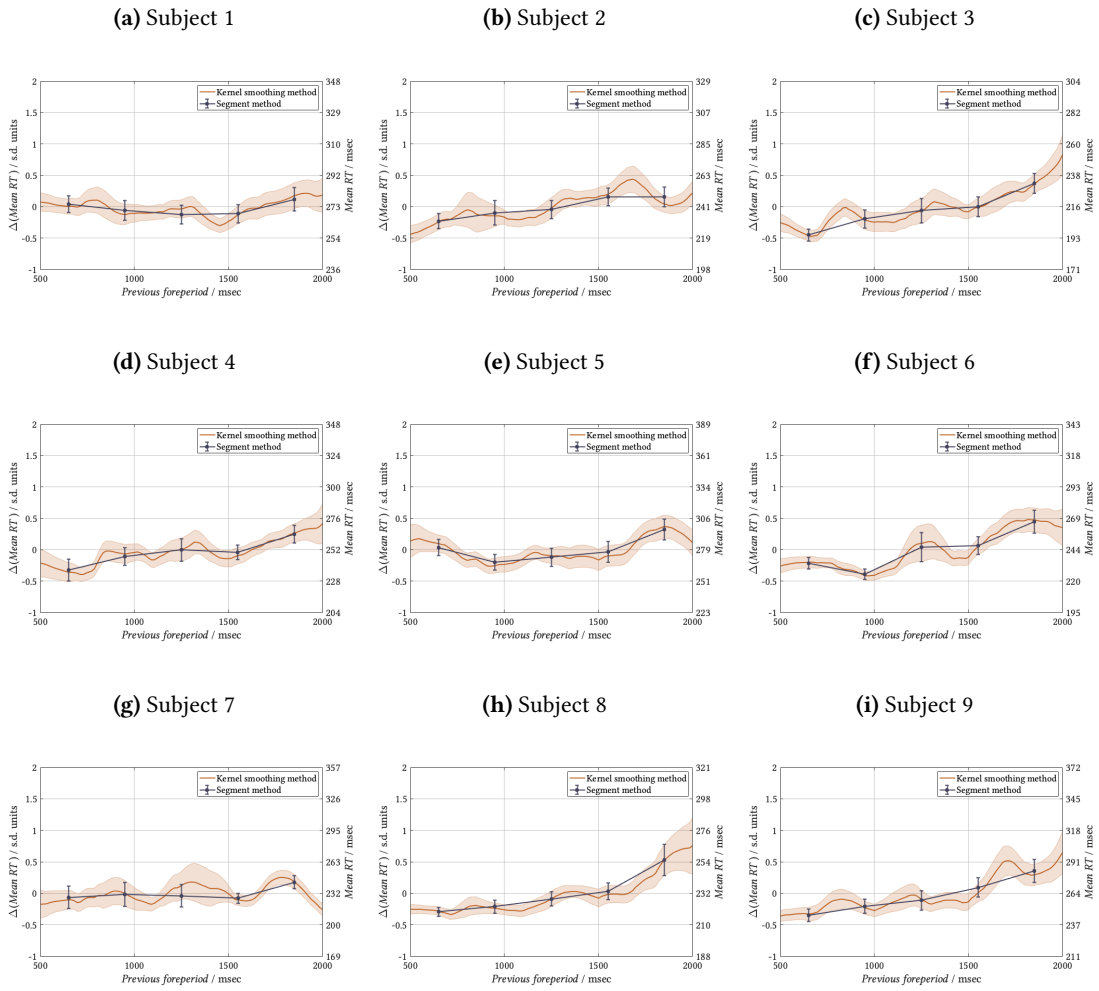


Figure 7.6. INDIVIDUAL SUBJECTS' PREDECESSOR FOREPERIOD CURVES (PART 1 OF 2). Namely, mean response time conditioned on previous trial's foreperiod. Note the presence of two scales on the ordinate axis, one for normalised and the other for absolute response time.

1000 and 1500 msecs; and the third, 1500 and 2000 msecs. Ordinary FP curves plotted separately for the subset of data in each partition; we shall refer to this set of curves the *sequential effects curves*. These curves are analogous to those shown in figure 4.5(b), which showcases results from the prior literature. The sequential effects curves obtained from the present data are shown in figures 7.8 and 7.9 (individual and aggregate results respectively). As with the first stage of this analysis, aggregation is performed both across the grouped data and by averaging across individuals, with the same provisos as before. Note that, to maintain visual clarity, only the segment method (per the FP curve fitting methods described above) is shown for this second stage.

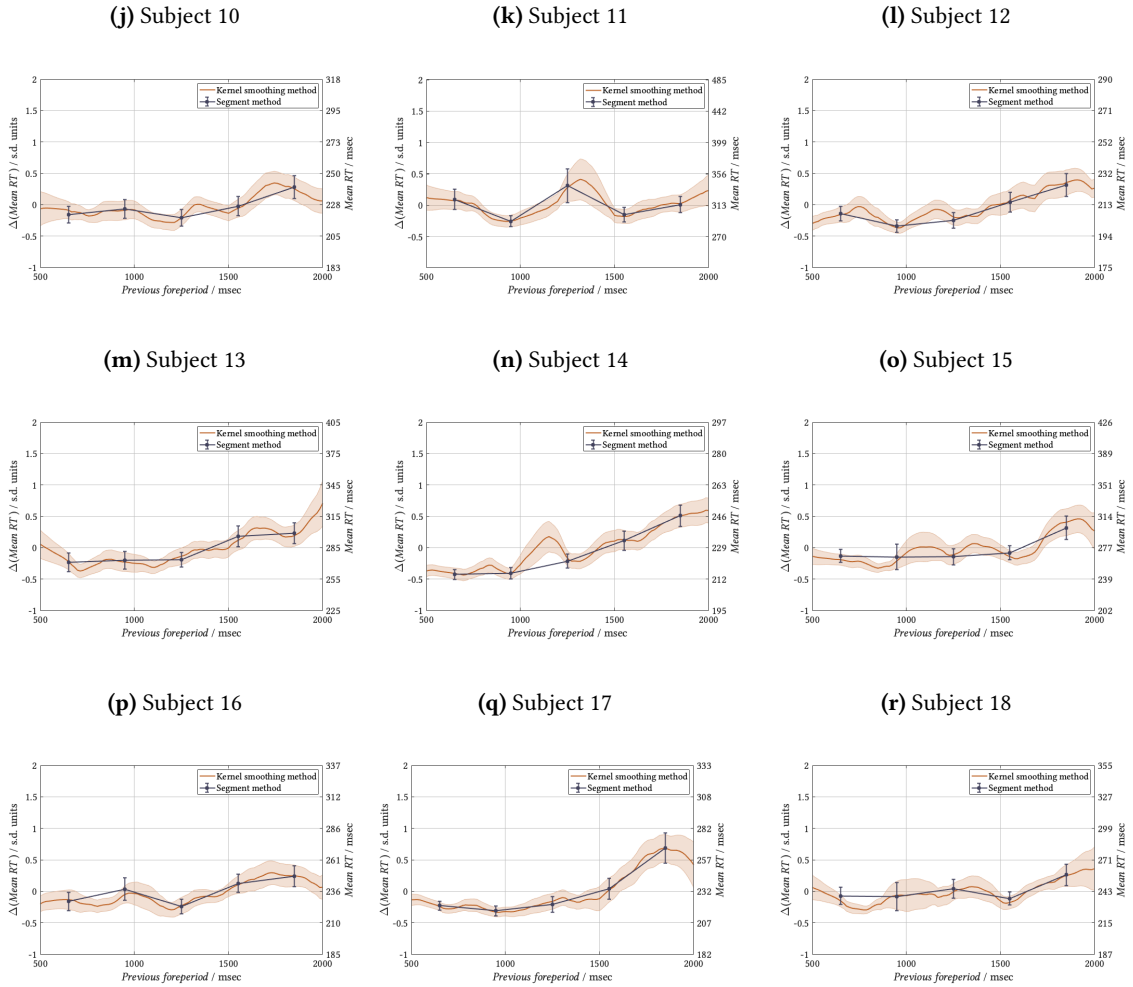


Figure 7.6. INDIVIDUAL SUBJECTS' PREDECESSOR FOREPERIOD CURVES (PART 2 OF 2). Namely, mean response time conditioned on previous trial's foreperiod. Note the presence of two scales on the ordinate axis, one for normalised and the other for absolute response time.

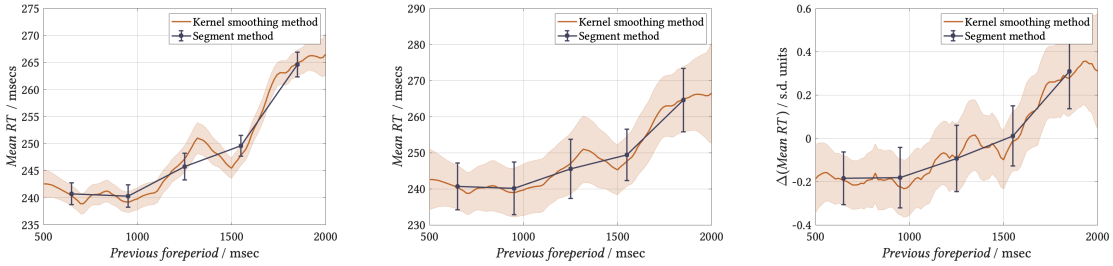
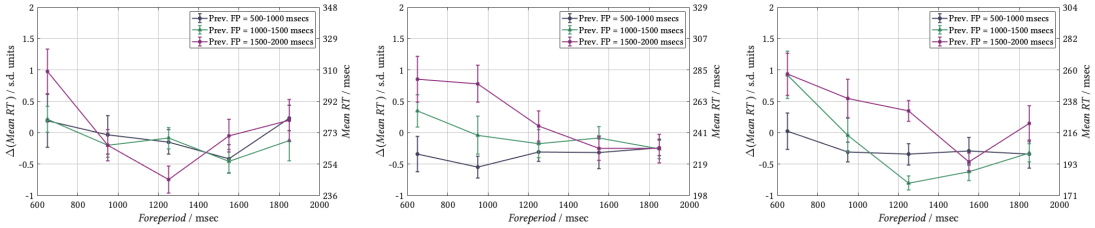
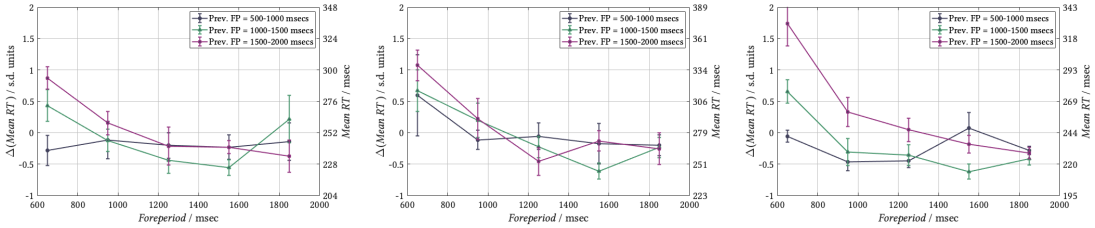


Figure 7.7. AGGREGATE PREDECESSOR FOREPERIOD CURVES.

(a) Subject 1 **(b)** Subject 2 **(c)** Subject 3



(d) Subject 4 **(e)** Subject 5 **(f)** Subject 6



(g) Subject 7 **(h)** Subject 8 **(i)** Subject 9

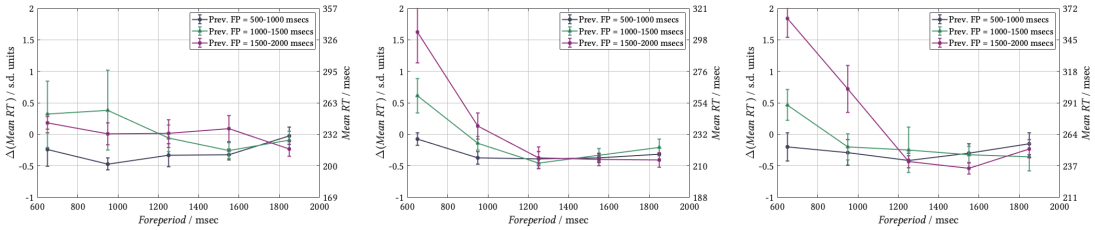
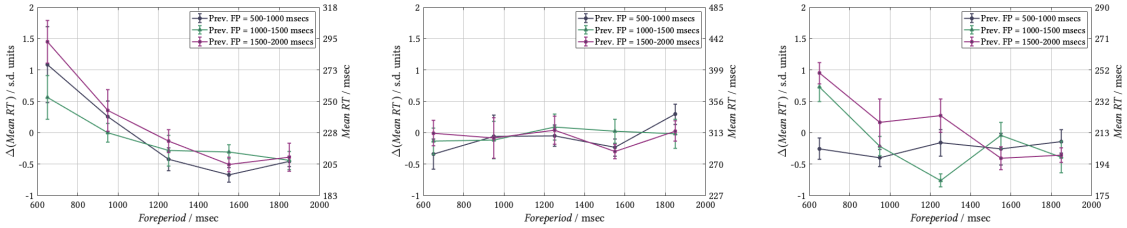
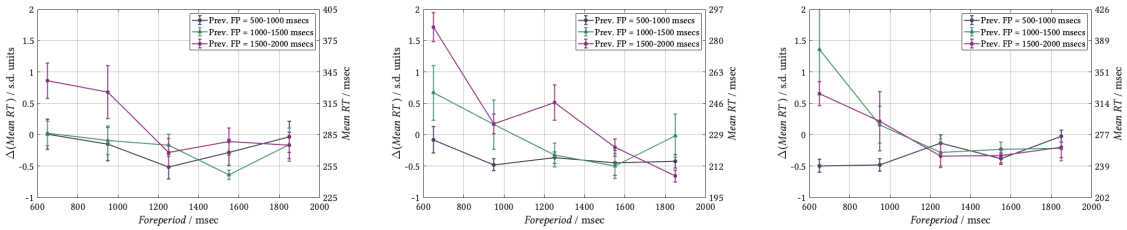


Figure 7.8. Asymmetric sequential foreperiod effect in individual subject data (part 1 of 2).

(j) Subject 10 **(k)** Subject 11 **(l)** Subject 12



(m) Subject 13 **(n)** Subject 14 **(o)** Subject 15



(p) Subject 16 **(q)** Subject 17 **(r)** Subject 18

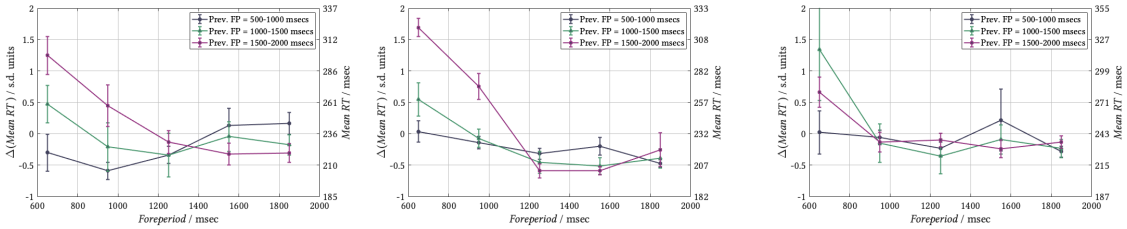


Figure 7.8. Asymmetric sequential foreperiod effect in individual subject data (part 2 of 2).

(a) Absolute RT, grouped data **(b)** Absolute RT, mean across subjects **(c)** Normalised RT, median across subjects

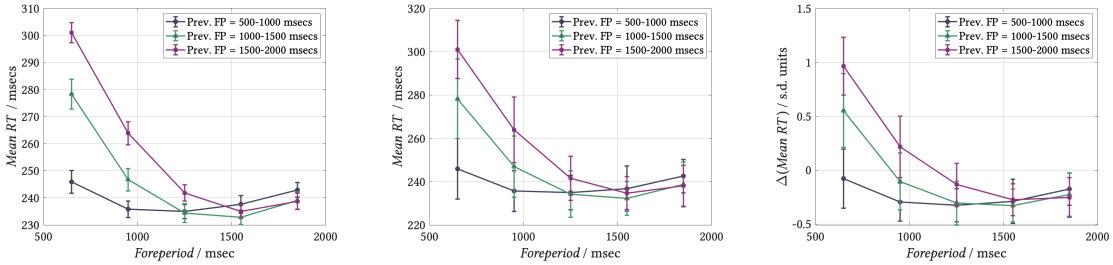


Figure 7.9. Asymmetric sequential foreperiod effects in aggregate data.

5:2 Results and discussion

The aggregate predecessor FP curve shows increasing mean RT with longer predecessor FP, as expected. The effect can be seen most pronounced in figure 7.7(a), which shows the predecessor FP curve calculated from the *grouped* RT data, so that the error bars represents the standard error of the mean of all trials being considered, across all the subjects. For comparison, figure 7.7(b) represents the average subject's results, that is, the result of taking the mean of the predecessor FP curves obtained for each individual subjects; the basic curve is therefore the same as in 7.7(a), but the error bars larger, since they represent the average standard error of the mean of a single subject's trials.

Figure 7.7(c) takes the median across subjects for the curve, in which representation the effect is still visible. Further, as with the ordinary FP curves before them, the aggregate curves are similar to many of the individual FP curves (figure 7.6), although the individual predecessor FP curves are less pronounced than the ordinary FP curves. As before, the concordance of the aggregate curves obtained across the two averaging methods suggests that the ASF effect observed in the aggregate is due to the presence of real sequential effects in individuals, and not due to distortion by a few outliers.

Figure 7.9 shows analogous results for the sequential effects curves. In particular, the asymmetric effect of long predecessor FPs on short- and long-FP successor trials – amplifying the ordinary foreperiod curve in the latter, diminishing it in the former – can be seen clearly at the group level (7.9(a)). Again, the effect is seen fairly clearly when reckoning the median subject (7.9(b)); and although it is not always clear at the individual subject level, it can be seen in a significant number of subjects despite the low number of trials collected (subjects 2, 6, 8, 9, 16, and 17 in figure 7.8; it is worth bearing in mind that each predecessor FP partition represents about 85 RTs per subject). The presence of the effect at the median levels again leads us to conclude that the effect under consideration exists at the level of individual subjects.

6. DISCUSSION

This chapter presents a description and analysis of data from a simple reaction time task with uniformly distributed foreperiods and salient, sudden-onset response signals – features designed to elicit, in a pronounced form, the complex foreperiod effects described in §4.5. The data here, collected from eighteen subjects, will be used in later chapters to makes Bayesian inferences regarding new models of simple responding designed to account for these complex effects; the purpose of this chapter has been to describe and analyse the data available to establish that it does indeed exhibit the complexities reported in earlier studies.

The analysis shows that, considered in light of all the major empirical phenomena in the

SRT, the present data replicates the findings of earlier studies. Considering both the PDFs and hazard functions, the response time data is typical: the PDF exhibits an early peak of fast responses followed by a long tail of slow responses, and the hazard function takes on an early peak followed by a fall to a relatively constant level, as expected for highly salient, sudden-onset stimuli. With regards to the foreperiod distribution effect, the average foreperiod curve across subjects shows decreasing mean RT with increasing foreperiod, as expected of a uniform FPD and its rising hazard; further, this form of FP curve is visible at the individual subject level. Finally, the FPD effect is more pronounced on those trials whose predecessor had long FP than on those whose predecessor had short FP – consistent with the set of findings dubbed the asymmetric sequential foreperiod effect in the previous literature. As with the FPD effect, the ASF effect is visible both at the aggregate and individual subject level.

Since there are no discrepancies between the present data and that reported in earlier studies, we conclude that the present data are typical of their kind. Further, since the data exhibit pronounced foreperiod effects, they are suitable targets for fitting by mathematical models which attempt to account for those effects. On the strength of these findings, we conclude that the present data will be appropriate for the forthcoming Bayesian assessment of new models of simple responding that attempt to incorporate foreperiod effects.

It was mentioned in the introduction that an uncommon but distinct effect in the data, which we shall call *corrected responding*, would receive some attention. Having detailed in §7.2 a general classification scheme used to group the responses observed, and which allows corrected responses to be identified, we are now ready to discuss this effect.

Of the 4500 trials collected across the eighteen subjects, 4428 fell into the hit class; 11 in the multiple-hit; 4 in the late; 1 in the miss; 5 in the premature; and 51 in the corrected class. Thus, even though they are very rare occurrences compared to ordinary hit trials, corrected responses are by far the second-most common class of response behaviour recorded when compared to the remaining possibilities.

The roughly ten-fold factor by which corrected responses outnumber plain premature responses, and the roughly five-fold factor by they outnumber multiple-hits, strongly imply that the second response in the corrected response trials is driven by a deliberate attempt to fix an earlier response made in error. If they were not deliberate, but accidental, we would trials with a premature-then-hit pattern of response actions to be outnumbered by premature-only responses (by independent probabilities multiplying). Further, the tiny number of multiple-hit trials supports the notion that accidental or stray responses are rare.

These facts about corrected responding provide some evidence for the distinction between anticipatory and reactive response processes detailed in §4.2, with the earlier response being anticipatory and the corrected response reactive. It may also be speculated that subjects are continually monitoring their performance, and are able to discern the difference between anticipatory

responses – perhaps as failures of inhibition – and reactive ones, since corrected responses are given even when the RT of the (putatively) anticipatory response comes after the imperative moment. A full investigation of this phenomenon is beyond the scope of this work and the capacity of the data presented here, so we will no longer dwell on it, except to say that closer study of this phenomenon holds some promise for revealing more about the cognitive mechanisms underlying simple responding.

Having established in this chapter the correspondence of the data to that found in earlier studies, the next chapter turns to tackling issues regarding the data's mathematical modelling.

—ω—

CHAPTER 8

New models of simple responding

Abstract. Simple reaction time tasks (and one-choice reaction time tasks more generally), in which a subject must and need only respond as quickly as possible to the appearance of a signal, are of great significance to cognitive science: they are ubiquitously employed in cognitive testing; have important analogues in everyday life; and – at least theoretically – the cognitive mechanisms involved in their performance form the nucleus of those involved in other, more complex cognitive tasks. Moreover, the data from these tasks forms a meaningful prototype for the data collected in numerous other cognitive tasks, since the dependent variable, response time (RT) is shared by them.

Despite all this, mathematical models of simple and one-choice responding are underdeveloped. The most notable omission in our current theoretical understanding of simple responding is the absence of well-defined mathematical models which encompass two hallmark empirical phenomena: the stereotyped *response time distribution* and *foreperiod effects* covered in chapter 4. This one omission appears in two bodies of literature. The first concerns itself with the detailed investigation of processes of temporal preparation in the the simple reaction time task; in this branch of research, mathematical modelling has been secondary to detailed empirical investigation. The other concerns mathematical models of response time and human decision-making, which, while it has treated one-choice tasks, has yet to bring the aforementioned foreperiod effects into its scope.

The aim of this chapter is to begin filling this gap. This will be achieved by elaborating existing drift-diffusion models from the response time literature to allow them to accommodate the phenomena mentioned above. Two new classes of response time model are introduced, *foreperiod parameter-dependence* and *collapsing threshold* models, and the manner in which they can overcome the shortcomings of existing models is explained. The development of these models sets up the work of the final investigations chapter, where the models will be fitted to the empirical data described in the previous chapter and assessed against extant models of RT through the application of MultiNest.

1. SHORTCOMINGS IN THE THEORY OF SIMPLE RESPONDING

1:1 On the theory and significance of one-choice reaction tasks

Perhaps the most elementary cognitive tasks conceivable are those which require nothing more of participants than the quickest possible response to a single signal. Such tasks have reaction time

(RT) as their most basic dependent variable, which – as the reader will recall from the discussion in §4.1 – is cognitive science's most commonly collected, and perhaps most important, experimental measure. Examples in this class of tasks, grouped under the heading of *one-choice reaction time*, include the simple reaction time task (SRT), the go/no-go task, and the psycho-vigilance motor task (Luce, 1986).

As argued earlier (§4.1), one-choice tasks are of fundamental interest to cognitive science, since all other stimulus-response tasks whose basic dependent variable is RT (that is, most tasks) *inter alia* require the abilities to (a) detect a response signal and (b) choose to respond to it; but these are the only imperative cognitive demands of one-choice reaction time tasks, which in that sense form the nucleus of other, more complicated tasks meant to probe more complex cognitive processes.

It is therefore, I argue, of some theoretical moment to determine whether our models of the performance on this simplest class of tasks are adequate: for if they are not, it is questionable whether our investigations of more complicated cognitive processes effectively isolate the processes of interest, and even more doubtful whether our models of the components of cognitive processing involved in more complex tasks are satisfactory.

If this broad-reaching view of the significance of one-choice tasks fails to move the reader, he or she might also consider the following features of one-choice tasks as rendering them worthy of close study. First, one-choice reaction tasks are ubiquitous in everyday life; to take one obvious example, we must notice and choose to respond to the changing colour of a traffic light. Second, one-choice tasks are widely used in cognitive assessments, the results being seen as a good measure of processing speed and a useful way to study the basic effects of conditions such as sleep deprivation on cognitive processes (Ratcliff & Van Dongen, 2011).

Before we frame the problem that we will take up in this chapter, let it be stipulated that we will narrow our attention to *simple responding*, that is, the processes involved in performing the SRT; appropriate generalisations of the forthcoming argument to one-choice tasks more generally will be reserved until the Discussion. Now, to recap (cf. §4.2), the SRT consists of a series of trials where on each trial, the aim is to respond as quickly as possible to the appearance of a response signal (RS), but not before. Every trial has the same structure: a warning signal (WS) starts the trial and is followed by a foreperiod (FP, usually of variable and random length) during which no signal is present, which in turn ceases at the onset of the RS; finally, the cessation of the RS concludes the trial. The simple reaction time task differs from other one-choice tasks in that the RS is presented frequently, with proximate presentations on the order of seconds apart, and in that there are no distractors, catch trials, or changes in stimulus properties (such as position) from trial to trial. (It really is as simple as possible.)

Despite their multifaceted importance, the theory of simple response mechanisms – especially *qua* mathematical models thereof – remains underdeveloped. This is so despite there being two

bodies of scientific literature which, on the basis of current knowledge, could take a fruitful interest in the problem. A detailed survey of these two bodies of literature is found in chapter 4. Here, we will briefly review these two lines of inquiry, and provide more detail on how the paths they have beaten lead to the problem this chapter undertakes to solve.

1:2 Empirical phenomena in simple responding

The first body of literature comprises experimentally driven investigations of simple response behaviour, which have been especially preoccupied with exploring systematic effects in simple response time (Hohle, 1965; Los et al., 2014; Niemi & Näätänen, 1981; Woodrow, 1914). An old and extremely important set of findings here finds consistent relationships between mean response time (on the one hand) and foreperiod duration, sequence, and overall distribution (on the other). Two related families of phenomena are of central importance.

The first is the so-called *foreperiod effect*: mean RT seems to be negatively correlated with the hazard (conditional probability) of the appearance the RS during the foreperiod, with the hazard determined by the FP distribution (FPD) across the task as a whole. Thus, FPDs possessed of increasing hazard (such as uniform distributions), elicit slower responses on average on those trials where the foreperiod is shorter, and vice versa; while an "unaging" FPD (that is, exponentially distributed foreperiod lengths) will elicit the same mean RT no matter the subset of trials considered *vis-à-vis* foreperiod length.

The second centrally important set of phenomena goes by the name of the *asymmetric sequential foreperiod effect*. This effect sees trial-to-trial changes in FP act to modulate response time. Specifically, long foreperiods on a predecessor trial increase mean RT on its successor trial (hence the "sequential"), with the increases in mean RT affecting successor trials being driven by considerable slowing in the RT on successor trials with short FPs, while successor trials with long FPs remain unaffected (hence the "asymmetric").

Foreperiod effects are generally taken as evidence for a process, dubbed amongst other things *temporal preparation* (Los et al., 2014), that unfolds during the foreperiod and is modulated by hazard (a statistical property of the distribution of the foreperiods) or history (via trace conditioning) or some combination thereof, according to the particular account. However, despite the thoroughly quantitative nature of these FP phenomena, this branch of cognitive scientific research has not expressly pursued the mathematical modelling those phenomena. Furthermore, these investigations have in any case focused primarily on the relationships between foreperiod and *mean* RT, setting aside the consideration that RTs are subject to well-documented, recognisable, and non-normal random variation.

1:3 Models of response time and their shortcomings

By contrast, the question of the mathematical modelling of response time and its distribution (more broadly construed to include RTs from other tasks, especially two-choice reaction time tasks) has received extensive treatment in the second body of literature addressed here. This work has been doubly motivated by the fundamental significance of response time to the whole field of cognitive science, and the observation that a broad range of tasks reliably elicit evidently similar RT distributions (generally speaking, unimodal, with an early peak and a long tail of slow responses). Consequently, much of the mathematical treatment of response time has been applied to the question of describing the overall distribution. A general ethos behind this distribution-oriented approach can be seen in, for example, Van Zandt (2000): "Many models state explicitly the characteristics of RT by specifying it as a random variable [...] The shape of the density often provides clues to the kind of random variable and, therefore, potentially to the candidate processes underlying the execution of a particular RT task."

Furthermore, it is now widely acknowledged that in the study of tasks in which the dependent variable is RT, it is important to take into consideration the entire observed distribution of RTs, since the behaviour of aspects of RT beyond its mean often contains useful information (Whelan, 2008); the desire to analyse the data in this holistic manner has spurred the search for adequate mathematical models of RT distribution that can be used to characterise the data.

In this vein, numerous models have been proposed as accounts of the response time distributions. These models can broadly be categorised as either *distributional models* or as *cognitive models* (Schwarz, 2001), which are distinguished by their ends. Distributional models are intended to be descriptive; cognitive models are intended to be explanatory. Distributional models are often taken to have the virtue of simplicity, while the latter as seen as more cognitively plausible, but more unwieldy.

The former category contains shifted Wald, ex-Gaussian, ex-Wald, gamma, and Weibull distributions (Burbeck & Luce, 1982; Hohle, 1965; Schwarz, 2001; Stone, 1960; Van Zandt, 2000). The latter category is of which the most prominent example in the latter category is Ratcliff's one-choice drift-diffusion model (Ratcliff, 2015; Ratcliff & Van Dongen, 2011). It is based on his highly influential drift-diffusion model for choice response tasks (Ratcliff, 1978, 2014). The Ratcliff drift-diffusion model is the dominant model in the dominant *evidence accumulator* paradigm of human decision making (Evans & Wagenmakers, 2020; Ratcliff, 2014; Ratcliff et al., 2016). The evidence accumulator paradigm has been successful in explaining a number of hallmark phenomena in studies of choice response time, and has the theoretical appeal of having parameters which are

Now, in reality, RTs are collected as a time series, sampled (as it were) from the same subject across successive trials. In the construction of all the aforementioned models, however, it is assumed that that the data to be accounted for comprises the marginalised RT distribution only.

Indeed, Ratcliff and Van Dongen (2011) explicitly state that "for all [one-choice reaction tasks, *see next section*], the dependent variable is RT, and *the RT distribution provides the data for modeling*. A successful model must produce the correct shape for the RT distribution [...]." (Emphasis mine.) No further necessary conditions are stated for a model to be deemed successful. This brings us to the major issue that motivates the theoretical investigations of this chapter. By focusing on the RT distribution as representative of the information contained in RTs as a whole, RTs are treated not just as random variables (an assumption I continue with here), but as independent and identically distributed (IID) draws from a fixed generating distribution.

Aside from being intuitively unsatisfying in light of the sequential process by which such data is collected, by making the IID assumption, existing models necessarily fail to take into account certain well-established systemic effects in RT data – especially foreperiod effects. Addressing this discrepancy (and referring to the modelling of response time in general, not just of one-choice tasks), Whelan (2008) notes that while RTs are generally assumed to be IID, "this is rarely the case in practice because of factors [...] such as fatigue and sequential effects," but that these factors "are generally assumed to be of negligible impact and are therefore ignored." All in all, then, despite well-established findings of systematic variation in SRT tasks, mathematical models of SRT data (and of RTs more broadly) nevertheless tend to to assume that RT can be treated as an IID random variable; *a fortiori*, the manifest presence of foreperiod effects is, in our best theoretical efforts, flatly rebuffed. Indeed, to the best of the author's knowledge, not a single mathematical model has been reported in the literature which attempts to simultaneously account for both the distribution of and the influence of FP on RT in the SRT.

1·4 The specific aims of this chapter

Thus, taking either of two approaches, we arrive at the same problem: quantitative models which accommodate the full spectrum of phenomena present in even the simplest of cognitive tasks have not been put forward. This fact provides the impetus for this chapter and sets its targets: to take the first steps towards a unified account of all the known phenomena, and thereby to restore the natural marriage of empirically driven research into the timecourse of temporal preparation and quantitatively driven research into models of human decision-making and response time. In particular, this chapter will introduce new models of SRT performance which aim to be capable of accounting for the target phenomena outlined above.

The ideas developed in this chapter form the groundwork for the next, which will further the argument of this chapter by showing that the new models of SRT performance herein introduced are not only *capable* of accounting for the phenomena of interest, but, under the terms of Bayesian model comparison, are *superior* to extant models of RT. The purpose of the next chapter will be to explore the consequences of this finding and to argue that they have scientific significance in their own right, especially for what they reveal about the nature of cognitive data and the

approach to modelling that nature dictates. But first, let us turn to the matter at hand.

2. GENERAL APPROACH AND PRELIMINARIES

In this section, the mathematical theory of simple responding will be extended to introduce two new classes of model which, per the basic goals of this study, are capable of simultaneously accounting for both the observed RT distribution and FP effects in simple responding. These two new classes, the *foreperiod parameter-dependence* (FPPD) and *collapsing threshold* (CT) classes, will be fully developed below, before which a few brief notes of a general nature will be made.

The basic strategy employed in the development of these new models is a conservative one. It can scarcely be denied that existing models of RT distribution cannot claim their fair share of successes (the reader is directed to Chapter 4 for a review); of especial note are the evidence accumulator paradigm in general and the Ratcliff drift diffusion model in particular, which have had profound successes in explaining numerous, quite subtle patterns present in the data, while providing meaningful theoretical interpretations of the components of cognitive processes involved in the generation of that data. In order to preserve the virtues of these models, they will therefore form the nucleus of the new models here; the extension out from this nucleus will be achieved by the simple device of asking what relatively minimal changes could be made to the old models in order to account for the problems identified above.

Furthermore, the development of both classes proceeds from a common point. As mentioned in the introductory section, the empirical deficiencies of present models, *viz.* their inability to account for the foreperiod effects apparent in the data, can be formally ascribed to the assumption that RT is an IID random variable. This assumption necessarily entails that, under the mechanisms proposed in theory, there be no systematic trial-to-trial variation in the processes which give rise to a response, nor, by extension, in the statistical properties of the recorded response times. Yet, if this were actually the case, mean RT clearly would not – could not – vary systematically under the manipulation of trial parameters such as foreperiod length. To overcome this, both FPPD and CT models violate the assumption of IID trials, though by different means. With these general notes made, the development of the models themselves can proceed.

3. FOREPERIOD PARAMETER-DEPENDENCE MODELS OF SIMPLE RESPONDING

Mathematically, where an RT distribution is stated as some analytical PDF, the IID assumption inheres in the constancy (across trials) of the parameters governing the exact form of the RT distribution. This observation suggests a natural remedy that can be applied to such analytical models of RT: a new model can be made simply by allowing an existing model's parameters to vary with against trial parameters in a well-defined way. By this modification, the RT predictions

for each trial will be given independently, but not identically. From this basic idea, I will develop the first class of new models of simple responding,the *foreperiod parameter-dependence* (FPPD) class of models.

Carrying the basic idea forward, let us consider (for the moment, and for simplicity's sake) the basic FP effect only, neglecting sequential FP effects. Let us also, for the time being, adopt the shifted Wald model as the *kernel* distribution: the analytical PDF assumed to underlie the RTs, whose parameters will now be made subject to the properties of a given trial. Within the limits of these assumptions and of those established by observation, we require that mean RT be related to foreperiod length. In particular, in our empirical interrogation of RT data in the previous chapter, we found that the relationship between mean RT and FP length could be well enough described by an exponential curve of the form

$$m = m_{\min} + (m_{\max} - m_{\min})\, e^{-\gamma(t-t_0)} \tag{8.1}$$

where t is foreperiod length (offset against the minimum foreperiod t_0), m is mean response time (varying from $m_{\max}$ at t_0 and decaying eventually to $m_{\min}$), and γ is a constant controlling the rate of the decay of mean RT with respect to foreperiod length.

Before we continue, a few provisos: This equation was the result of analysing data obtained by an averaging procedure across numerous subjects, and so strictly speaking is not necessarily representative of any one subject in particular (though, as we saw, there are many individuals who do exhibit, at least qualitatively, personal RT-FP relationships quite similar to the one obtained in the aggregate). Further, it it is worth bearing in mind that the data were obtained from a version of the task where the foreperiod distributions was chosen to be uniform, precisely in order to induce such clear relations, whereas it is known that in general different foreperiod distributions will produce predictable changes in the overall shape of the curve. Nevertheless, let us – as an admittedly provisional measure – adopt the exponential formulation in equation (8.1) as our standard: first, because the formal specification of the FP-RT relationship is required to continue the development of this line of argument; second, because this functional form is (under suitable variation of its controlling parameters) capable of accommodating a wide range of possible FP-RT relations at least approximately; and third, because it is the clear that this form bears some semblance to what is really going on.

Now, given the relationship between mean RT and FP length captured by equation (8.1), can we work it into a model of simple responding that has the response times on each trial drawn from an individualised shifted Wald distribution? Yes, quite easily. The reader will recall that the shifted Wald distribution has the following PDF:

$$f(t \mid \mu, \sigma, \Delta) \;=\; \frac{1}{\sqrt{2\pi\sigma^2 (t-\Delta)^3}} \exp\left\{ -\frac{[(t-\Delta)-\mu]^2}{2\mu^2\sigma^2(t-\Delta)} \right\}. \tag{8.2}$$

where μ is governs the mean of the Wald component of the distribution (*sans* shift), and is generally interpreted as mean decision time (or, more hypothetically, as ratio of the decision threshold to the the drift rate of the imputed drift-diffusion process); σ controls the variance, and can be read as the ratio of the decision threshold to the scale of the diffusive noise; and Δ is a shift parameter, usually taken to indicate the non-decision time. Furthermore, the reader will remember that that the shifted Wald distribution's mean m is simply the sum of μ and Δ. It follows that the distribution requires

$$m = \mu + \Delta, \tag{8.3}$$

that is, that mean response time can be calculated as the sum of the mean decision time and the non-decision time.

Now, to describe the findings relating to the foreperiod effect, mean RT is just the quantity we require to vary in response to FP length. Evidently, this cannot happen under the usual shifted Wald model of RT wherein μ and Δ are fixed (as per the earlier discussion of the IID assumption and the constancy of parameters). To achieve our goals, then, it is required that at least one of these parameters will have to be made a function of FP length. Under the usual interpretation of the parameters of this distribution as they relate to response time, we may well suppose that, psychologically speaking, the locus of the influence of FP on RT is found in the decision-making process (why should FP affect non-decision processes such as motor control?). Pressing forward with this supposition, any variation in mean RT must then arise from a dependency of the μ parameter on foreperiod length. Keeping Δ fixed, then, we arrive at a model whereby the response on each trial is drawn from its own shifted Wald distribution; where the Wald parameters σ and Δ are fixed across trials; and where the μ parameter is to be found as a function the trial's FP length t according to

$$\mu = (m_{\min} - \Delta) + (m_{\max} - m_{\min}) \exp\left(-\gamma\left(t - t_0\right)\right), \tag{8.4}$$

per the relationships established in (8.1) and (8.3). For later use, we shall refer to this a *linking function*, i.e., the function which links the trial-to-changes in the foreperiod dependent parameters to changes in the independent variables (here, the foreperiod of the current trial).

Actually, implicit in this formulation are two distinct foreperiod-dependent parameter models based on a Wald shift kernel. To see this, the reader is reminded that the Wald distribution arises from considering a simple drift-diffusion process proceeding with drift at a constant rate, and with an unchanging characteristic level for the zero-mean Wiener noise process; the Wald distribution is that of first passage times across a fixed threshold. A change in the mean can come about from two distinct manipulations: a change in the drift rate, with slower rates increasing the mean, or a change in the threshold level, with higher levels having a similar effect. Further, the three-parameter Wald distribution (with drift, noise, and threshold parameters) is over-specified,

being reducible to two parameters by re-expressing any two of the original parameters in the appropriate (inverse) ratio to the third. Adopting the convention stated in §4.4:1, whereby we parameterise the Wald distribution mean and noise-to-threshold ratio, the first manipulation (changing drift rate) results in a change to the mean parameter only; the second manipulation (changing threshold level) also affects the noise-to-threshold ratio. it can be shown quite easily that, alongside the change in μ, the change in σ is given by

$$\sigma(m) = \frac{\gamma\mu(t)}{m_{\min} - \Delta}. \tag{8.5}$$

Using the terminology introduced above, equation (8.5) is a linking function for σ, which supplements the linking function for μ. For later convenience, we will designate these two models the (shifted Wald kernel) *varying drift with hazard linking function* and *varying threshold with hazard linking functions* respectively.

If one desires, a psychological interpretation can be made of these models. In the varying drift variant, sensory evidence accumulation rate is decided by the foreperiod length; in the varying threshold variant, sensory evidence decision threshold is decided by the foreperiod length.

From this one example, the generalisation to the FPPD class is clear: to make an FPPD model, one simply chooses an appropriate kernel distribution and formulates the linking functions for the parameters. The class as a whole, of course, is defined by stipulating that every FPPD model consists of a combination of a kernel and linking functions, and that every combination of kernel and linking functions is an FPPD model. One could, for example, make an FPPD model with an ex-Gaussian kernel, and adopt alternative linking functions which, unlike the one explicated above which deals with the FPD effect only, incorporates the ideas of – say – trace conditioning. Endowed with parameter interpretations, each FPPD model would then represent a specific set of theoretical commitments about the processes underlying simple responding. For now, though, with the basic idea explained, we shall not consider any further FPPD models.

4. COLLAPSING THRESHOLD MODELS OF SIMPLE RESPONDING

The second novel class of model introduced here, *collapsing threshold* (CT) models, seeks to account for systemic effects in mean response time not by having model parameters vary, but by augmenting the behaviour of one of the basic aspects of evidence accumulator models: the response threshold. The general idea is that, if the threshold itself over the range of allowable foreperiods, trials with different foreperiods will effectively "see" different threshold, causing the desire systemic effects. This general idea will now be developed through a series of specific examples.

4:1 The basic collapsing threshold model

Of the complex empirical phenomena reported in the SRT, this model is intended to account for the basic within-trial FPD effect only. The response threshold is set to decay exponentially with foreperiod duration as

$$\theta(\tau) = \theta_{\min} + (\theta_{\max} - \theta_{\min}) \exp(-\beta(\tau - \tau_0)), \tag{8.6}$$

for $\tau > \tau_0$, where τ is foreperiod duration, τ_0 is the minimum FP in the task's FPD, β controls the decay rate of the threshold, $\theta_{\max}$ is maximum value of the threshold (obtained at the very shortest FP), and $\theta_{\min}$ is the asymptotic lower limit of threshold.

Besides this collapsing threshold, the model in every respect follows the strictures of the basic shifted Wald model covered in §4.4:1: a constant drift rate ρ, supplemented by noise σ and a fixed non-decision time *Delta*. The FPD effect emerges from this model as follows: on trials with short FPs, the threshold is high, so responses are on average slowed. At longer FPs, the effective threshold is lowered, on average speeding the responses. Further, past a certain point, the threshold remains stable, leading to a flattening out of the FPD effect past a certain point, as observed in the data.

4:2 The collapsing threshold model with sequentially varying span and decay rate

The basic CT model only accounts for within-trial foreperiod effects, and fundamentally cannot account for asymmetric sequential foreperiod (ASF) effects. The key feature of the ASF effect, at least with uniform FPDs, is that the FP curve obtained is exaggerated when following long-FP trials ones while being flattened following short-FP trials.

To account for this within the CT framework, we note that the FPD effect can be damped by reducing the span (in the sense of distance from starting value to asymptote) of the exponential decay, making a flatter curve overall; or by increasing the decay rate, so that the flat portion of the curve is reached sooner. If either the span or the decay rate parameter were linked to predecessor FP duration, then it is feasible the ASF might be recovered.

As basic implementations of the varying span idea, we consider a model where the span varies by setting $\theta_{\max}$ on a given trial n according to the FP of the predecessor trial τ^{n-1}, in particular as

$$\hat{\theta}^n_{\max} = \frac{\theta_{\max} - \theta_{\min}}{\tau_{\max} - \tau_{\min}} \left(\tau^{n-1} - \tau_{\min}\right) + \theta_{\min}. \tag{8.7}$$

Similarly, for the varying decay rate mode, we might let

$$\hat{\gamma}^n = \gamma_{\max} - \frac{\gamma_{\max} - \gamma_{\min}}{\tau_{\max} - \tau_{\min}} \left(\tau^{n-1} - \tau_{\min}\right). \tag{8.8}$$

Evidently, many variations are possible on this basic theme, but this is not the place to explore them all. It will be taken that the above proposals form reasonable starting points for the development of these new classes of model. The next chapter will continue the theoretical development of these models by fitting and assessing them to real SRT data using MultiNest.

—ω—

CHAPTER 9

A Bayesian assessment of models of simple responding

Abstract. This chapter contains the confluence and conclusion of the two major lines of argument developed so far in this thesis, namely, the argument for the adoption of nested sampling techniques in cognitive science, and the argument for the need for new and better models of one-choice – especially simple response time (SRT) – task performance.

To mount these arguments, I apply MultiNest – the nested sampling technique introduced and verified in earlier chapters – to test models of simple responding against real simple reaction time data. The models include traditional models of the SRT as well as the new, more complex models introduced in the last chapter that are designed to account for foreperiod effects. This thorough model comparison shows that existing models do not provide the best accounts of the data; rather, the novel, more functionally complex models, based on more elaborate theories of cognitive processing, consistently obtain higher evidentiary support. While doing so, I highlight the ways in which situation embodies many of the important conceptual and practical challenges commonly encountered in cognitive scientific modelling.

From the basic results obtained here, three immediate conclusions are drawn. First, MultiNest is vindicated in practice as an valuable tool for modelling work in cognitive science; it can be hoped that the algorithm's combination of accuracy and efficiency might open up new lines of scientific inquiry. Second, the dominant evidence accumulator framework of response time and decision making requires revision and extension if it is to continue to be scientifically useful. Third, our understanding of the basic mechanisms of stimulus response is deepened through the insights gleaned from the new models; in particular, it seems that temporal preparation effects directly modulate the signal-detection and decision-making aspects of cognition.

These results, obtained in the context of the SRT, lay the groundwork for the broader argument of the final chapter: that cognitive data exhibits irreducible complexity that cannot be ignored in modelling work, and that future advances in understanding cognition can be driven by adopting a combination of (a) advanced tools for inference of the sort used here and (b) new guiding principles for theoretical work based on the acknowledgement of the irreducible complexity of cognitive data.

1. INTRODUCTION

1:1 General background

Human cognitive abilities have been extensively studied through the use of a wide range of behavioural tasks, and data obtained from them analysed in the hope that they might reveal useful clues about the cognitive processes which underlie the data. The most commonly measured dependent variable in these experiments is *response time* (RT; Luce, 1986), the time between the onset of some signal and a participant's overt response to that signal (often a key press). Response time is so commonly used both because it is easy to collect and because, theoretically, it ought to be highly informative. After all, it is the very nature of processes to take time, and response time is perforce composed from some combination of processes (often construed as as perceptual, cognitive, and motor). The characteristics of the response time data elicited by a particular task, and the changes in those characteristics in accordance with experimental manipulations, must therefore reflect the activities of the underlying processes.

1:2 Model selection in cognitive science

Since it is not possible to deduce models directly from data, to make progress on the question of the processes underlying the performance of a particular task, it is necessary to propose candidate models that might plausibly account for the data, fit the models to the data, and test the candidates to determine (at least provisionally) which provide adequate accounts of the data. Broadly speaking, the initial step is done by noting certain characteristic features of the data (such as the shape of the RT distribution, whether in its PDF, CDF, or hazard function representation; or the general form of the relationship between certain summary measures of the data; or the way in which aspects of the data predictably change according to experimental condition) and finding or creating some mathematical model capable of reproducing the noted features. For the present discussion, we shall take it for granted that the initial step, the conceptual formation and mathematical formalisation of plausible candidate models, has been taken.

The latter steps in the process give rise to a number of practical challenges. The first is the simple need to fit the data – that is, to find those parameter values which produce the closest accord between the model's predictions of the data and the actual observations – or the problem of *parameter estimation*. Second is the problem of *model selection*: given competing candidate models of the data, which does the best job at accounting for them? As has often been noted (and as was discussed at length in the General Introduction), when data contain any degree of indeterminacy – random effects due to sampling, noise, measurement error, and the like – the best model is not necessarily, as one might initially suppose, that which produces the closest quantitative fit with the data. Models possessed of sufficient functional flexibility can be made

to over-fit any given dataset, so that the model's structure and parameters encode not general features of interest in the data, but the minutiae of the random effects the data happen to contain. Where such indeterminacy is sufficiently large, statistical models, which treat key dependent variables as random variables, must be employed.

A principled, consistent, and mathematically rigorous approach to both parameter estimation and model selection for statistical models is found in *Bayesian inference* (BI), which has been covered extensively already (chapters 2, 3, and 5.). In brief, Bayesian inference provides a solution to the problem of parameter estimation in the form of *posterior estimates*, and solution to model selection in terms of the *Bayes factor*. The problems of parameter estimation and (especially) model selection have driven an increase in the use of Bayesian methods for model inference in the cognitive sciences, with a commendable cadre of researchers expounding the approach and promoting its techniques (Andrews & Baguley, 2013; Annis et al., 2019; Evans & Wagenmakers, 2020; Gronau et al., 2017). The present work thoroughly supports this general approach.

However, as I have been arguing thus far, the particular BI techniques (such as standard MCMC approaches, information criterion based approaches, and, more recently, steppingstone and thermodynamic integration methods) that have been imported into the field to date do not represent the state-of-the-art (be it for posterior parameter estimation or for Bayesian evidence calculation); furthermore, cognitive science stands to benefit from the adoption of more sophisticated methods. In earlier chapters, I have shown that MULTINEST – a highly advanced BI algorithm developed for use in cosmology, astronomy, and particle physics, where it has supplied hundred-fold efficiency increases in the calculation of Bayesian evidences, achievable in a single run and with accurate error and posterior estimates produced as a by-product –

1. produces precise and accurate evidence estimates, as demonstrated in situations where exact solutions are available for comparison;

2. does so to a degree which is, in practical terms, as good as other algorithms which have recently been promoted in the cognitive psychology literature (the aforementioned steppingstone and thermodynamic integration methods), while incurring far less computational cost;

3. is capable of making accurate and useful model selection choices from synthetic data which plausibly resembles real RT data, consisting of multiple virtual subjects contributing relatively few trials each;

4. and does so quite feasibly with models that would normally be considered computationally intensive.

Continuing this line of argument, the first major aim of the chapter is to extend the claim of MULTINEST's usefulness to cognitive science to its maximum limit. In particular, by applying it to

current challenges in the modelling of simple reaction time, I will show that MultiNest can be used to answer questions of cognitive scientific interest, and that – on account of an accuracy and efficiency that facilitate the consideration models of greater functional complexity than has hitherto been the norm – it allows those questions to be probed to a greater depth than they have been before. Moreover, the answers obtained from these deeper probings are found to be revealing, suggesting that fundamental changes ought to be made in the enterprise of cognitive modelling – in particular, a shift in emphasis towards complex, time-dependent, and task-adaptive models of cognition is hinted at.

By demonstrating for the first time the definite theoretical significance of such models, MultiNest shows its worth as a tool for answering active questions in cognitive science; furthermore, by making such models susceptible to analysis, it opens up new avenues of future research. But let us not get too far ahead of ourselves here; first, let us turn our attention to the particular theoretical challenges to which MultiNest will be applied.

1:3 Simple responding, foreperiod effects, and models of response time distribution

Recall from earlier that reaction time (RT) is cognitive science's most commonly collected dependent variable. Amongst the most elementary of cognitive tasks are that have RT as their dependent variable are those requiring nothing more of participants than the quickest possible response to a single signal. Such *one-choice reaction time* tasks include the simple reaction time task (SRT), the go/no-go task, and the psycho-vigilance motor task (Heathcote, 2004; Luce, 1986; Ratcliff and Van Dongen, 2011).

As argued extensively earlier (§4.1), one-choice tasks are of fundamental interest to cognitive science. On a practical level, one-choice tasks present many analogues in everyday life, and are commonly used in cognitive testing. On a purely scientific level, since all other stimulus-response tasks whose basic dependent variable is RT (that is, most tasks) *inter alia* require the ability to detect and choose to respond to the response signal; these are the only imperative cognitive demands of one-choice reaction time tasks, which in that sense form the nucleus of other, more complicated tasks meant to probe more complex cognitive processes. It is therefore of theoretical significance to determine whether our models of the performance on this simplest class of tasks are adequate: if they are not, doubt is cast on our investigations of more complicated cognitive processes – such as those involved in choice response and stop signal tasks – since it becomes unclear whether these tasks properly isolate the processes of interest, and even more unclear whether our models of cognition in these more complex tasks are satisfactory.

Amongst one-choice tasks, we focus our attention here on the simple response time (SRT) task. To recap (cf. §4.2), the SRT consists of a series of trials; on each trial, the aim is to respond as quickly as possible to the appearance of a response signal (RS), but not before. Every trial has the same structure: a warning signal (WS) starts the trial and is followed by a foreperiod (FP) during

which no signal is present, which is itself ended by the onset of the RS; the cessation of the RS concludes the trial. The simple reaction time task differs from other one-choice tasks in that the RS is presented frequently, with proximate presentations on the order of seconds apart.

There are a number of major outstanding issues in our understanding on the processes of simple responding (§4.6,8.1:3). The one we take up here is the inability of current models of RT to account for extremely well-established empirical phenomena in the SRT, especially so-called *foreperiod effects* (Los et al., 2014; Niemi & Näätänen, 1981; Teichner, 1954; Woodrow, 1914).

To recap, foreperiod effects have been established by a long-standing findings of consistent systemic relationships between mean response time and foreperiod duration, sequence, and overall distribution. Two families of phenomena are of central importance. The first is the so-called foreperiod distribution (FPD). Mean RT seems to be negatively correlated with the hazard (conditional probability) of the appearance the RS during the foreperiod, with the hazard determined against the FP distribution as a whole. Thus, FP distributions possessed of increasing hazard (for example, uniform distributions), elicit slower responses on average on those trials where the foreperiod is shorter, and vice versa. The second is the *asymmetric sequential foreperiod* (ASF) effect. This refers to the phenomenon whereby trial-to-trial changes in FP modulate response time. Specifically, predecessor trials with long FPS produce pronounced slowing on successor trials with short FPS. These findings are generally taken to imply the existence of a process of *temporal preparation* during task performance. (These findings are reviewed in great detail in §4.5.)

Concurrently, as has also already been well covered (§4.3), response time exhibits substantial randomness, and RT distributions across numerous tasks share certain pronounced commonalities: an early peak of fast responses is offset by a distinctive skew in a long tail of slow responses. The fundamental importance of RT to cognitive science means that many mathematical models of RT distribution have been proffered (Van Zandt, 2000), including both straightforward distributional models intended as descriptions of the data, and more substantive *cognitive models* which embody a set of theoretical commitments regarding the underlying cognitive processes leading to response (Farrell & Lewandowsky, 2018). Shifted Wald, ex-Gaussian and ex-Wald models are amongst the distributional models that have been forwarded as descriptions of RT data (Burbeck & Luce, 1982; Hohle, 1965; Schwarz, 2001; Stone, 1960). The dominant cognitive modelling framework for human decision-making and response time is that of *evidence accumulator models*, which posit that responses occur via the crossing of a decision threshold by a quantity known as *sensory evidence* which accumulates noisily over time in the presence of a response signal. The most famous EAM model is Ratcliff's drift-diffusion model, which was originally designed to account for choice response and has been extended into the domain of one-choice tasks (Ratcliff, 1978, 2014; Ratcliff & Van Dongen, 2011).

Despite this wealth of empirical and theoretical work, models of simple responding which

simultaneously capture both the systemic and random aspects of simple responding are lacking. The work on temporal preparation and foreperiod effects has rarely offered concrete mathematical models of the phenomena it studies, and the focus has been on the relationship between foreperiod and *mean* RT, setting the random characteristics of RT to one side. Conversely, RT modelling work has often "assumed [factors which produce systematic influences on RT] to be of negligible impact" (Whelan, 2008), building models on the simplifying assumptions that the sufficient and necessary criterion in RT modelling work is to "produce the correct shape for the RT distribution" (Ratcliff & Van Dongen, 2011), and that this can be achieved by models which treat RT as IID variable.

1:4 Aims and plan of this chapter

Thus, taking either of two approaches, we arrive at the problem of incorporating foreperiod effects and response time distribution in a single mathematical framework. To begin to address just this problem, the previous chapter introduced to two new classes of mathematical model. These are the *foreperiod parameter-dependence* (FPPD) and *collapsing threshold* (CT) models; each is an extension of existing models of RT distribution designed to account at least partially for foreperiod effects.

This situation is ripe for picking by Bayesian inference: there are numerous competing statistical models of a phenomenon and a complex set of data to be accounted for, making qualitative model evaluation difficult. Despite the range of RT models proposed, none has emerged as a clear best candidate. Many of the models are difficult to fit due to their computational complexity, demanding efficient methods of parameter estimation; this need is compounded by the fact that there is no guarantee that the proposed models do not have difficult posterior geometries (cf §3.1). Moreover, while the new models of simple responding are appealing in that they appear to be able to account for FP effects, at least under certain conditions, there is no guarantee that they do in fact provide better overall accounts of the data. They may, for example, forego the quality of their fit to the marginal RT distribution in their attempt to account for FP effects; or they may simply prove to be over-fitted. Model selection through the use of the Bayes factor provides the only rigorous and principled means of evaluating the overall quality of the models.

Answering the question of model selection would concomitantly address another issue. Following on from the discussion a few paragraphs ago, it is indubitable that producing the correct shape for the RT distribution is a necessary condition any successful model of response time must meet, but is it sufficient? More pointedly, given the patent presence of foreperiod effects, is the IID assumption of RT modelling justified, and do the other factors which affect RT truly make only a "negligible impact"? Since the FPPD and CT models do not make the IID assumption, the model analysis here *ipso facto* provides a critical examination of the basic IID

assumption of response time modelling and the prevalent form of the EAM framework.

The second major aim of this chapter follows from all this, and complements the first: it is to determine which model of simple responding amongst those currently available best accounts for the data. Both of this chapter's major aims will be met by a single scientific investigation: a comparison of new and existing models of simple responding, tested against data from the simple response task (SRT) using MULTINEST. The results of this procedure will then be used to assess the models of simple responding, and the experience of the procedure will be used to assess the usefulness of MULTINEST and Bayesian inference to cognitive science.

2. BAYESIAN ANALYSIS OF MODELS

2:1 Fit data and sampling parameters

The fit data for this study is the simple response time data fully described in chapter 7. In brief, the data consists of the response times collected from eighteen subjects performing a SRT with a uniform FPD, with the critical moments distributed continuously and uniformly from 500 to 2000 msecs and a response signal lasting 800 msecs. There was no inter-trial interval, and the sudden offset of the RS was used as an implied warning signal. Each subject performed five blocks of 50 trials; the first trials from each block are discarded for this analysis because they are uniquely not preceded by another trial (which may affect the reaction process) and because the foreperiod on the first trials was considerably longer (3000 msecs) than the others. In the terms introduced earlier, late responses (up to 350 msecs) were included in the analysis; multiple hit trials took their RTs from the first recorded response; corrected responses took their RTs from the first recorded hit response; and missed and premature response trials were excluded.

2:2 Models, likelihood functions, and priors

Nine models of RT are chosen for analysis. These have been introduced throughout the thesis, so their formal aspects and the justification for their inclusion here will not be rehashed; references to these details are given below. This space is used to give an overview of the priors and the means of calculating their log-likelihoods for the present investigation. We emphasise in advance the computational complexity of the collapsing threshold models.

The following notation is adopted for the description of the prior distributions: $X \sim \mathcal{N}_+(a, b, c, d)$ means X is drawn from a truncated normal distribution with mean a, standard deviation b, and lower and upper limits c and d; $X \sim \mathcal{U}(a, b)$ means X is drawn a uniform distribution with lower and upper limits a and b; and $(X, Y) \sim \mathrm{Tri}(x_0, y_0, x_1, y_1)$ means (X, Y) is drawn uniformly from a right-triangle shaped region of the plane defined by the lower-diagonal half of the rectangle with corners at (x_0, y_0), (x_0, y_1), (x_1, y_0), and (x_1, y_1), with the diagonal drawn

from (x_0, y_1) to (x_1, y_0). We set the convention that $x_0 < x_1$ and $y_0 < y_1$; if $x_0 = y_0$ and $x_1 = y_1$, this ensures that $X \leqslant Y$ always.

SHIFTED WALD MODEL. As described in §4.4:1. The likelihood was calculated using the analytical form of the PDF, as given in equation (4.6). The prior is defined over the moments of the distribution ξ_1 (mean), ξ_2 (standard deviation) and ξ_3 (skewness):

$$\begin{aligned} \xi_1 &\sim \mathcal{N}_+(180, 50, 215, \text{inf}) \\ \xi_2 &\sim \mathcal{N}_+(55, 10, 0, \text{inf}) \\ \xi_3 &\sim \mathcal{U}(0, 3) \end{aligned} \tag{9.1}$$

and the model parameters derived from the moments using the relationships established earlier.

EX-GAUSSIAN MODEL. As described in §4.4:2. The likelihood was calculated using the analytical form of the PDF, as given in equation (4.6). The prior is the same as that for the shifted Wald model; again, the model parameters are derived from the moments using the relationships established earlier.

EX-WALD MODEL. As described in §4.4:3. The likelihood is calculated using the parexic method described in §6.2:5. The prior is the same as that for the shifted Wald model; again, the model parameters are derived from the moments using the relationships established earlier.

RATCLIFF ONE-CHOICE DRIFT-DIFFUSION MODEL. As described in §4.4:4. The likelihood is calculated using the parexic method described in 6.2:5; the parameter notation given below should be referenced to the description in §4.4:4. The prior is given by

$$\begin{aligned} 1/\hat{v} &\sim \mathcal{U}(50, 200) \\ \phi_1 &\sim \mathcal{U}(0, 1/3) \\ \hat{\delta}_t &\sim \mathcal{U}(50, 200) \\ \phi_2 &\sim \mathcal{U}(0, 0.5) \\ \sigma &\sim \mathcal{U}(0, 0.2) \end{aligned} \tag{9.2}$$

where the parameter $\eta = \phi_1/\hat{v}$ and $s = \phi_2\hat{\delta}_t$.

SHIFTED WALD MODEL WITH FOREPERIOD-DEPENDENT THRESHOLD. As described in §8.3. The likelihood is calculated on each trial by first taking the trial's foreperiod and using the linking functions of equations (8.3) and (8.4) to derive the model parameters for that trial, after which the shifted Wald PDF can be used normally with the trial's RT datum. The parameter notation is

below is the same as that of the equations mentioned above. The prior is given by

$$\begin{aligned}(1/\rho_{\min}, 1/\rho_{\max}) &\sim \text{Tri}(190, 190, 400, 400)\\ 1/\beta &\sim \mathcal{U}(0, 3000)\\ \xi_2 &\sim \mathcal{U}(0, 50)\\ \Delta &\sim \mathcal{U}(50, 200)\end{aligned} \tag{9.3}$$

where once again ξ_2 represents a standard deviation (the smallest possible in the model) and we derive the parameter

$$\sigma = \sqrt{\frac{\xi_2^2}{(\rho_{\min} - \Delta)^3}} \tag{9.4}$$

analogously to the normal shifted Wald distribution.

SHIFTED WALD MODEL WITH FOREPERIOD-DEPENDENT THRESHOLD. As described in §8.3. The likelihood is calculated on each trial by first taking the trial's foreperiod and using the linking functions of equations (8.3)–(8.5) to derive the model parameters for that trial, after which the shifted Wald PDF can be used normally with the trial's RT datum. The prior is the same as that for the previous model.

BASIC COLLAPSING THRESHOLD MODEL. As described in §8.4. The likelihood is calculated using a parexic method similar to the one used for the ROCDD and ex-Wald models. However, the violation of the IID assumption by this model requires that an individual PDF be estimated for each and every trial in the data, with the parameters derived according to the trial's FP; 50 000 sample paths are generated per trial to be used as the basis of the kernel density of estimation for that trial's PDF. The prior is given by

$$\begin{aligned}1/\hat{\rho} &\sim \mathcal{U}(50, 200)\\ \sigma &\sim \mathcal{U}(0, 0.05)\\ \Delta &\sim \mathcal{U}(50, 200)\\ (\theta_{\min}, \theta_{\max}) &\sim \text{Tri}(0, 0, 2, 2)\\ \gamma &\sim \mathcal{U}(0, 300)\,.\end{aligned} \tag{9.5}$$

Once again, the parameter notation is consistent with that used in the text where the model was introduced.

COLLAPSING THRESHOLD MODEL WITH SEQUENTIALLY VARYING SPAN. As described in §8.4. The likelihood is calculated as with the basic CT model, except that the trial parameters are derived according to both the current and the predecessor trial's FPS. The first trial was treated as having a long predecessor FP. The prior is the same as that for the basic CT model.

COLLAPSING THRESHOLD MODEL WITH SEQUENTIALLY VARYING DECAY RATE. As described in §8.4. The likelihood is calculated in a similar manner to the varying-span CT model. The prior is the same as that for the basic and varying-span CT models, except that the decay rate is now split into two parameters $\gamma_{\min}$ and $\gamma_{\max}$, which are covered by the prior

$$(\gamma_{\min}, \gamma_{\max}) \quad \sim \quad \mathrm{Tri}(0, 0, 300, 300)\,. \tag{9.6}$$

2:3 Overview of analysis

The principal analysis in this phase is meant to determine which model provides the best overall account of the simple responding. Since the models above are meant as formalisations of hypotheses regarding the performance of individual subjects, the models must be fitted to each subject's dataset individually. This results in 18 separate Bayesian evidence scores and posterior estimates. We thus face the issue of making an inference regarding model selection across the so-called *data corpus*, that is, the collection of the individual subject datasets.

This problem is discussed in depth in §6.4:1 and appendix C. We recall from there that, at the level of a data corpus, the inference is not in fact over a model *per se*, but over what we called a *model ensemble* – an assignment of a particular model to each subject (or unit) in the corpus. The evidence for a model ensemble is found by summing the evidence contributions from each individual; the proof for this simple method is given in appendix C. The question we are considering – of which single model provides the best overall account – is, from this perspective, shorthand for which *pure* model ensemble best accounts for the data corpus, a "pure" ensemble being one in which the same model is assigned to every unit in the corpus. The first analysis will therefore consist of calculating the Bayes factors of the various pure model ensembles. The second part of the analysis breaks down the model comparison at the subject level, to see whether the aggregate results are indicative of patterns at the subject level.

2:4 General notes

PRIOR SENSITIVITY ANALYSES AND ABSOLUTE MODEL SUPERIORITY. This method will be employed in both the group- and subject-level model comparisons.

TECHNICAL NOTES. MULTINEST was run using 1000 live points, the stopping criterion or tolerance parameter set to 0.5 (standard for estimates of the evidence), and with mode separation activated. The runs were performed on the Imperial College London's High Performance Cluster. As most of the individual subject fits were run on nodes containing 48 cores, so MULTINEST's MPI capabilities were aligned to this. Priors and likelihood functions were coded in MATLAB. All study code will be made available.

Model	$\ln \mathcal{Z}$	$\Delta \ln \mathcal{Z}$	$2 \ln B_{10}$	Min. $2 \ln B_{10}$
Varying span CT	−22 227.68	1.73	0	463.69
Varying decay rate CT	−22 337.10	1.67	−218.84	219.36
Basic CT	−22 364.00	1.75	−272.64	202.82
FPPD shifted Wald – varying drift	−22 486.64	1.64	−517.91	−126.70
FPPD shifted Wald – varying threshold	−22 558.46	1.61	−661.56	−283.63
Ex-Wald	−22 559.39	1.39	−663.42	−350.53
Ex-Gauss	−22 607.73	1.36	−760.12	−498.01
ROCDD	−22 626.75	1.61	−798.13	−414.58
Shifted Wald	−22 725.74	1.34	-996.12	−739.74

Table 9.1. BAYES FACTORS OF THE PURE MODEL ENSEMBLES. The first column specifies the model; they are ranked from highest to lowest Bayes evidence, i.e. from best to worst. The second and third contain the ensemble log-evidence and log-evidence error estimate calculated as the sum of individual subject log-evidences and errors as returned by MULTINEST. The fourth column contains deviance-scaled Bayes factor (twice the difference of the natural logs of the evidences). The final column contains the results of the prior sensitivity analysis described in the methods; the value given is the result of taking the maximum possible log-evidence obtainable from the maximum likelihood value, and calculating the Bayes factor between that maximum possible evidence and the evidence of the best-performing model.

3. RESULTS

3:1 Model comparison at ensemble or group level

The results of the group level (*alias* model ensemble) analysis is shown in table 9.1. From these results, it can be seen that – taking into account all the data available – the varying span CT model receives by far the greatest evidentiary support as measured by the Bayes factor. Using the interpretative scale of Kass and Raftery (1995) displayed in table 2.1, compared to even the next-best performing model, the varying span CT model receives decisive support. (Translated into straightforward probability terms, the marginal likelihood of the varying span CT model is a factor of 2×10^{47} higher than that of the next-best model.) Indeed, even the average (deviance-scaled) Bayes factor per subject (=12.1) would be considered very strong evidence against that scale.

Furthermore, the models which fall into the collapsing threshold class are, in the sense defined above, absolutely superior to all the other models; that is, the non-CT likelihoods would be unable to produce comparable Bayesian evidence scores, no matter the prior selected.

3:2 Model comparison at subject level

Table 9.2 breaks down the aggregate results to the subject level. In all but three of the subjects, the varying-span CT model is selected as the best model at the subject level, just as it is at the group level. Furthermore, in many subjects, the Bayesian evidence in found to be strongly in favour of the varying span CT models, or failing that the CT class in general. However, it is worth

	Subject								
Model	1	2	3	4	5	6	7	8	9
Varying span CT	0	-9.15	0	0	0	0	0	0	0
Varying decay rate CT	-0.71	-20.06	-9.22	-21.22	-1.16	-19.95	-6.47	-16.17	-21.76
Basic CT	-1.82	-21.64	-12.29	-24.99	-3.16	-23.74	-9.54	-19.83	-26.36
FPPD shifted Wald – varying drift	-2.98	-8.07	-33.55	-49.95	-10.43	-33.21	-32.79	-9.77	-25.59
FPPD shifted Wald – varying threshold	-4.56	0	-51.97	-59.54	-18.93	-46.62	-32.79	-13.32	-39.72
Ex-Wald	-9.18	-12.51	-45.82	-55.37	-28.06	-51.80	-22.70	-56.55	-63.61
Ex-Gauss	-8.98	-12.18	-45.58	-57.00	-28.82	-51.93	-23.42	-55.88	-64.12
ROCDD	-12.59	-12.32	-54.82	-54.60	-27.42	-50.99	-11.06	-51.50	-70.76
Shifted Wald	-14.69	-21.62	-56.54	-53.22	-35.56	-68.17	-24.88	-84.64	-72.74
	Subject								
Model	10	11	12	13	14	15	16	17	18
Varying span CT	-19.15	0	0	0	0	0	0	0	-72.49
Varying decay rate CT	-7.45	-1.85	-20.83	-7.47	-29.01	-22.48	-14.64	-18.47	-80.70
Basic CT	-10.26	-3.04	-23.67	-10.95	-33.26	-26.38	-17.90	-21.97	-82.60
FPPD shifted Wald – varying drift	0	-19.81	-40.39	-39.69	-64.77	-29.33	-35.66	-82.37	-100.34
FPPD shifted Wald – varying drift	-4.35	-14.01	-46.99	-47.16	-87.27	-34.17	-36.36	-115.62	-108.98
Ex-Wald	-60.61	-12.50	-36.34	-28.63	-93.11	-45.10	-22.40	-119.92	0
Ex-Gauss	-55.89	-16.63	-33.69	-23.05	-93.63	-46.33	-14.49	-120.08	-109.21
ROCDD	-74.82	-7.42	-46.33	-43.97	-96.37	-40.23	-37.63	-120.22	-85.86
Shifted Wald	-84.11	-18.13	-57.79	-57.73	-103.45	-42.97	-49.32	-144.94	-106.40

Table 9.2. BAYES FACTORS OF THE MODELS FOR THE INDIVIDUAL SUBJECTS. The first column specifies the model; they are ranked from highest to lowest average Bayes evidence, i.e. from best to worst at the group level. The remaining columns, one for each subject, list the deviance-scaled Bayes factor for the subject; each model's Bayes factor is calculated against that of the best performing model for that subject, which can therefore be identified by having zero for its Bayes factor.

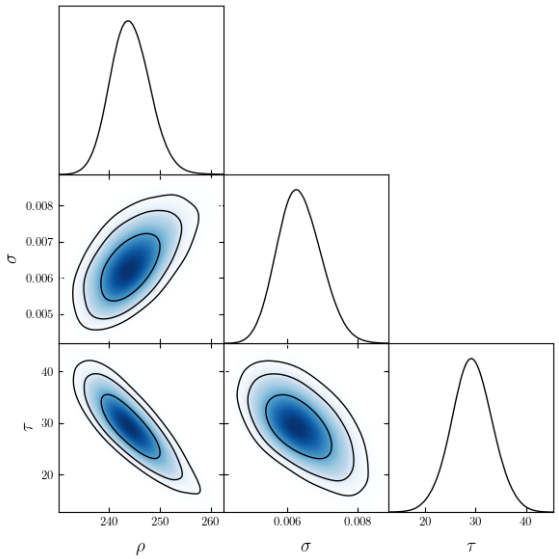


Figure 9.1. MARGINAL POSTERIOR PLOTS FOR THE EX-WALD MODEL FOR SUBJECT 1 The plots along the top (diagonal) edge are the marginal posteriors plots for each of the parameters in the model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

noting that a number of subjects fare much better under different models.

3:3 Posterior estimates

Figures 9.1-9.6 showcase the marginalised posterior distributions for six of the nine models tested here. Of particular note is the fact that many of these distributions exhibit the kind of difficult posterior geometries which, in comparison to standard MCMC methods, modern nested sampling techniques efficiently explore (these were discussed extensively in chapter 3). For example, curving degeneracies and mild multimodality can be seen in the ROCDD model (figure 9.2, and curving degeneracies can be seen in the FPPD models (figures 9.3 and 9.4). Non-curving degeneracies and unusual (*viz.* non-Gaussian) posterior geometries are common throughout. The fact that MULTINEST finds simple, Gaussian posterior geometries for the simple and well-behaved ex-Wald model (figure 9.1) suggests that the difficult posterior geometries in the remaining models are indeed products of the model and data, and not of an errant sampling process.

3:4 Comparison of model postdictions to empirical data

Since Bayesian model analysis is fundamentally comparative and relative in nature, it cannot tell us whether the models being assessed provide good fits to the data. This section is dedicated to showing that the best-performing of the new models of simple responding do, in fact, provide adequate fits to the data, and fulfil their stated aim of accounting for both marginal RT distribution

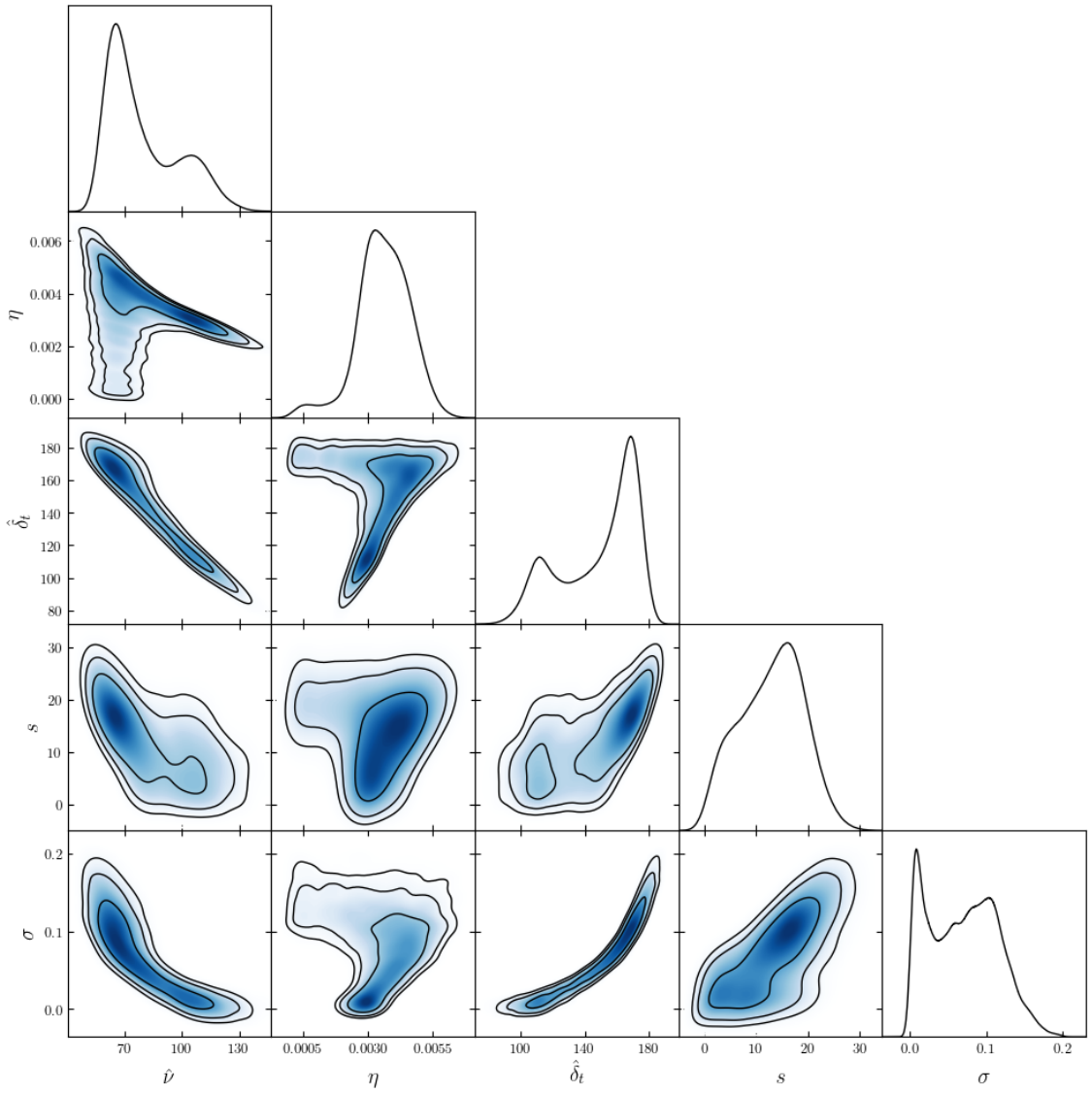


Figure 9.2. MARGINAL POSTERIOR PLOTS FOR THE ROCDD MODEL FOR SUBJECT 7 The plots along the top (diagonal) edge are the marginal posteriors plots for each of the parameters in the model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

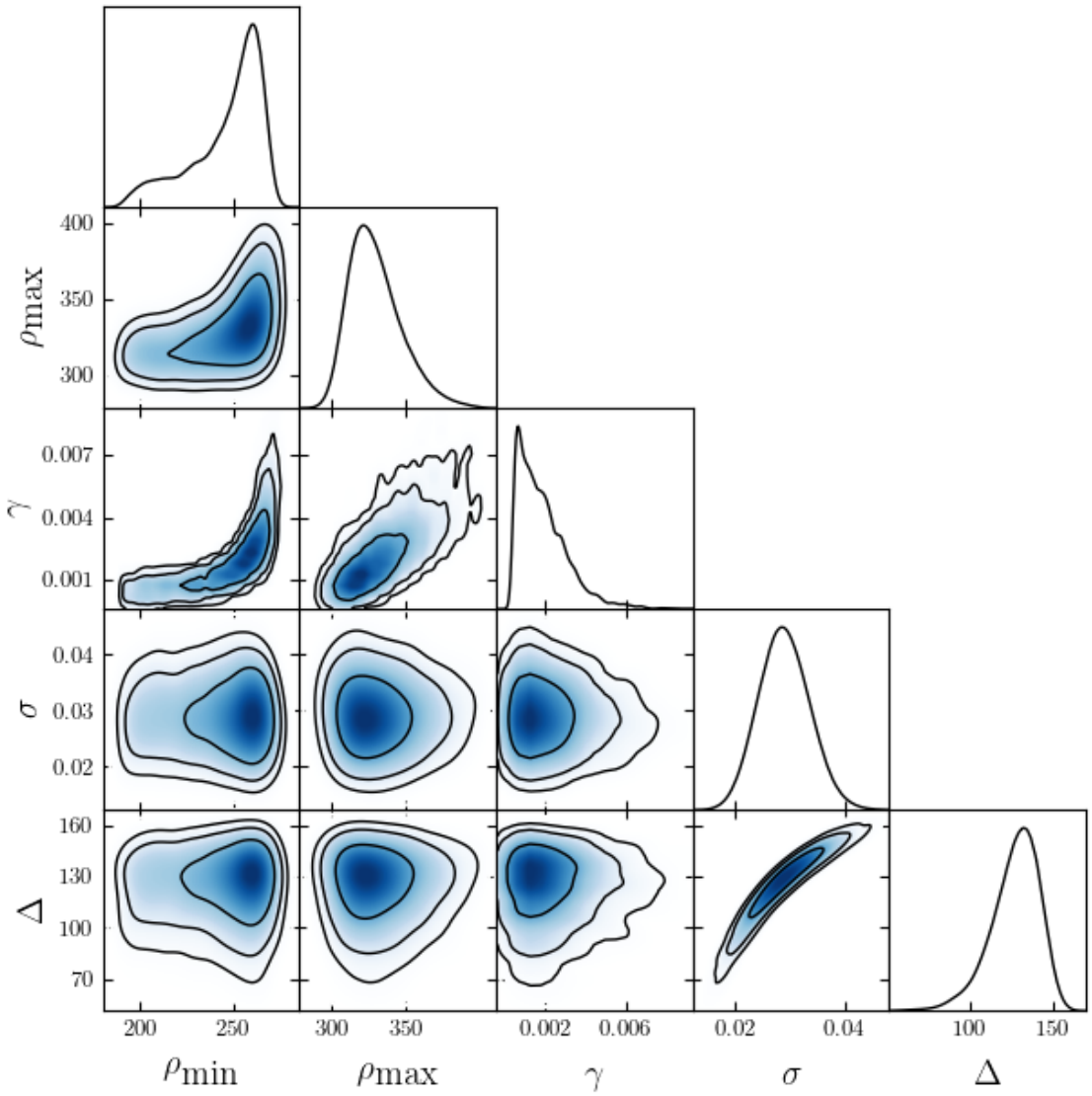


Figure 9.3. Marginal posterior plots for the fppd model with a shifted Wald kernel and foreperiod-dependent drift for subject 5. The plots along the top (diagonal) edge are the marginal posteriors plots for each of the parameters in the model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

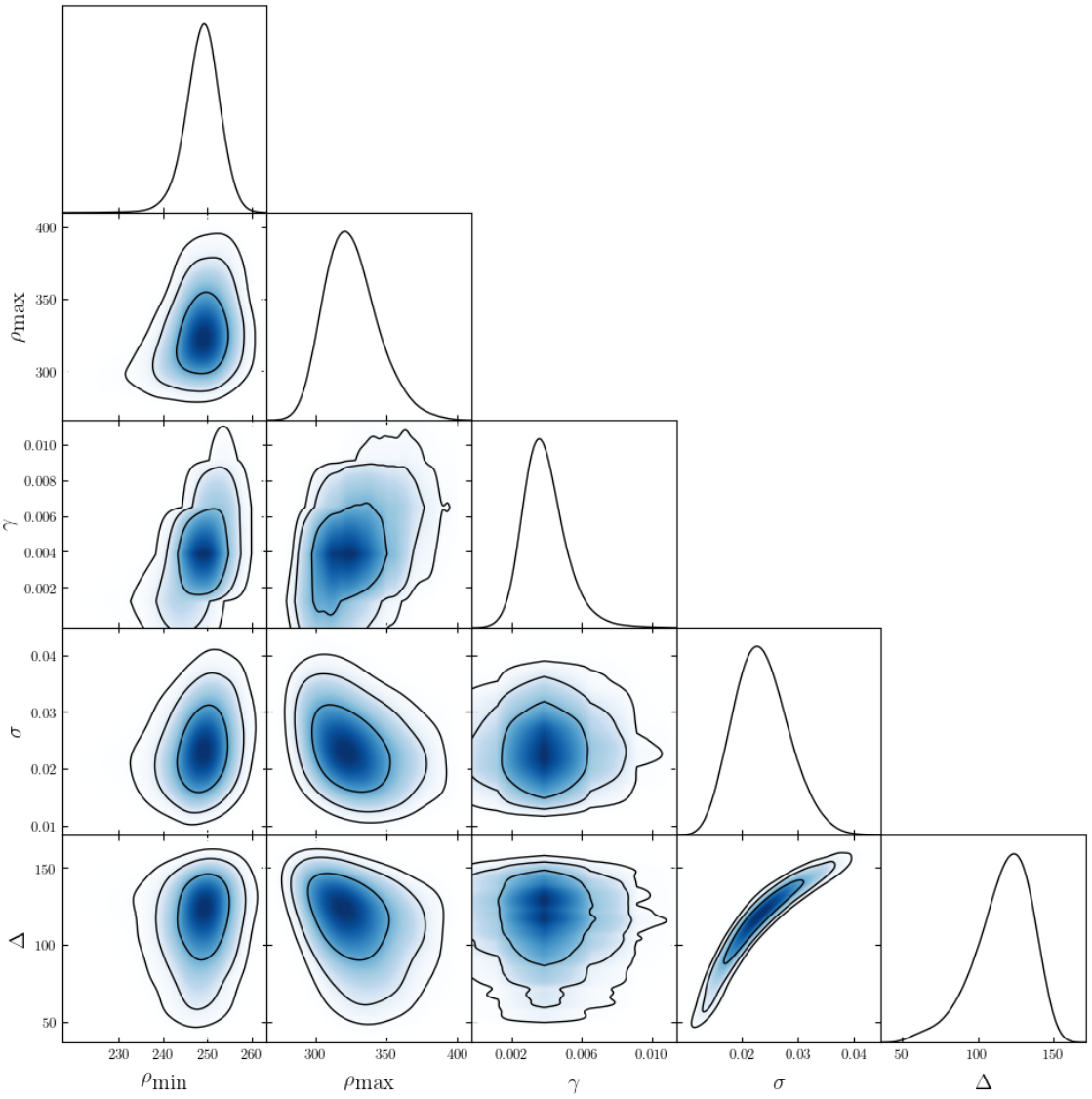


Figure 9.4. MARGINAL POSTERIOR PLOTS FOR THE FPPD MODEL WITH A SHIFTED WALD KERNEL AND FOREPERIOD-DEPENDENT THRESHOLD FOR SUBJECT 9. The plots along the top (diagonal) edge are the marginal posteriors plots for each of the parameters in the model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

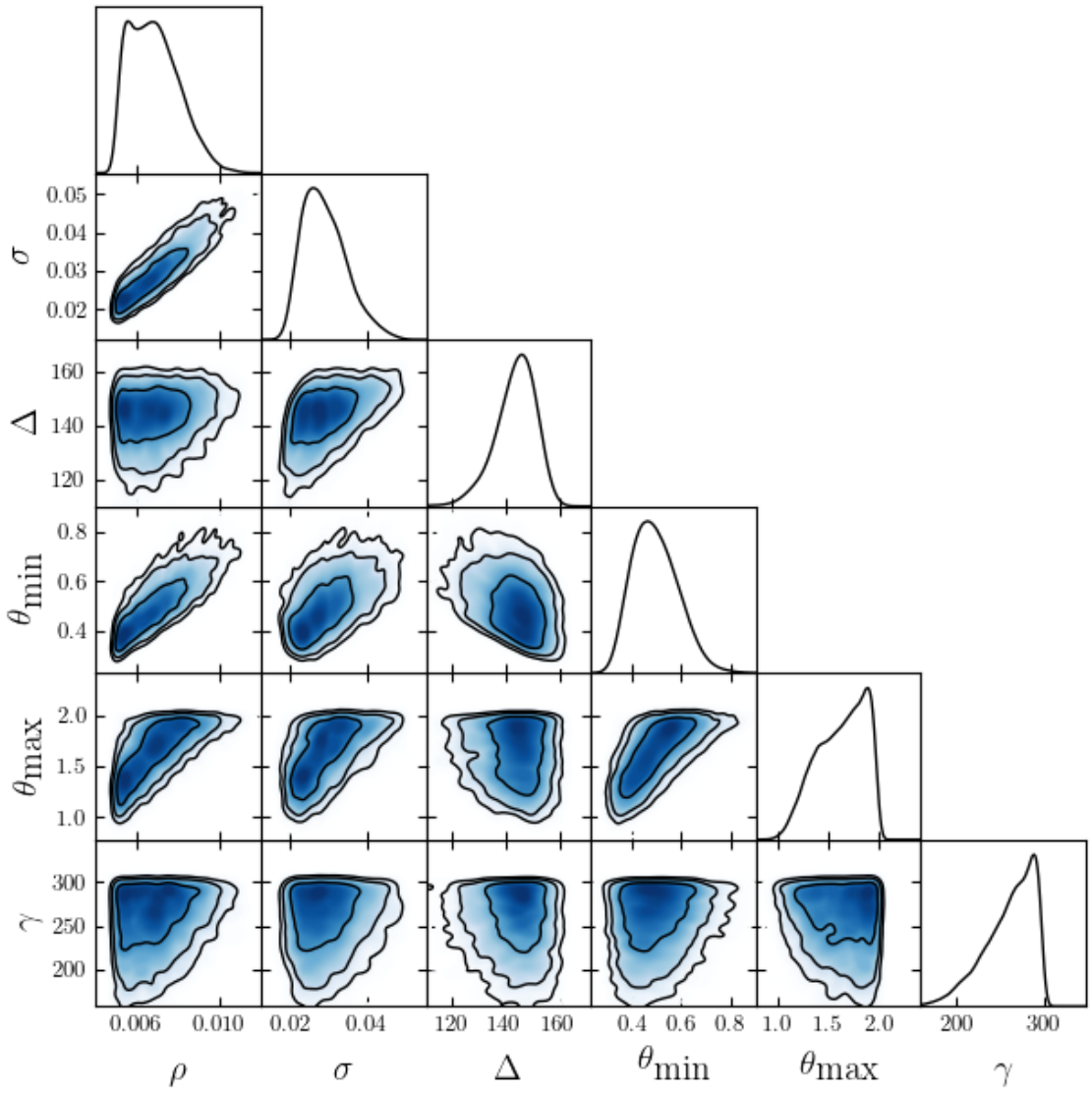


Figure 9.5. MARGINAL POSTERIOR PLOTS FOR THE BASIC COLLAPSING THRESHOLD MODEL FOR SUBJECT 10. The plots along the top (diagonal) edge are the marginal posteriors plots for each of the parameters in the model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

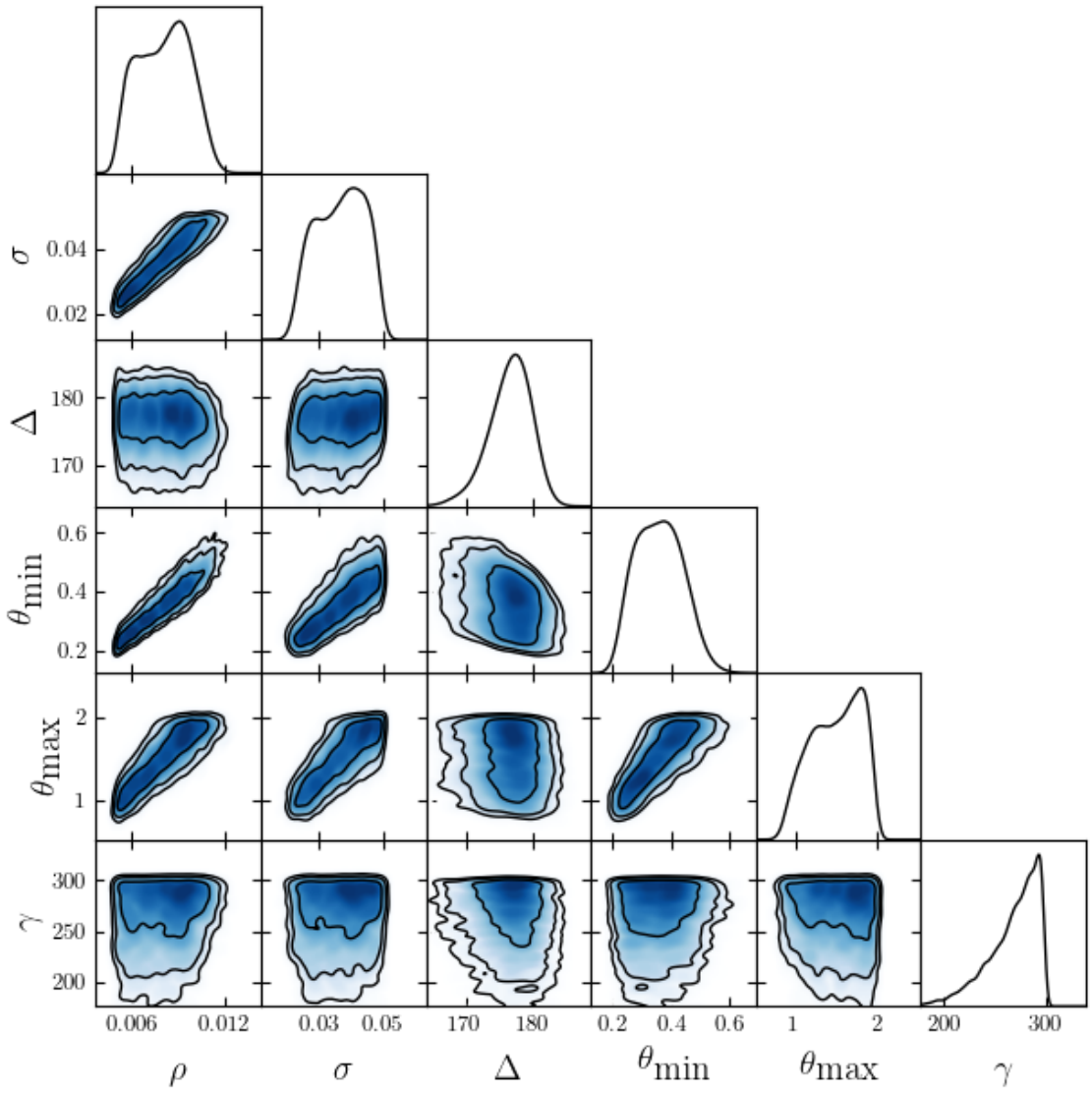


Figure 9.6. MARGINAL POSTERIOR PLOTS FOR THE VARYING SPAN COLLAPSING THRESHOLD MODEL FOR SUBJECT 8. The plots along the top (diagonal) edge are the marginal posteriors plots for each of the parameters in the model; the remaining plots are the 2d marginal posteriors for every pair of parameters. The depth of shading is proportional to the density in the 2d plots; the contour lines, from innermost to outermost, indicate iso-probability lines for the 68%, 95%, and 99% highest confidence regions of the marginalised posterior.

and FP effects.

To do so, the best-performing models – the three collapsing threshold models – will have their postdictions of subject performance at their maximum *a posteriori* parameters compared to the real data the Bayesian analysis was performed against. Three fits of the models to the data will be assessed: to the overall RT distribution, the average FP curve, and the average predecessor FP curve. These comparisons are shown in figures 9.7 through 9.9; the captions to those figures contain detailed information about how the fits were performed.

Figure 9.7 shows the fits of the best-performing models to the overall RT distributions. As can be easily seen, each of the models – when considering this aspect of the data only – provides a good fit to the data. Figure 9.8 shows that the models are capable of simultaneously providing highly accurate fits to the FP curves caused by the foreperiod distribution effect. Figure 9.9 further shows that the two elaborations of collapsing threshold model which were designed to account for the sequential FP effect – the varying span and varying decay rate models – can indeed model the sequential effects in the data. The models, therefore, can be considered a success, accounting for systemic effects of foreperiod on mean RT while preserving good fit to the overall RT distribution, which no other models of RT can do.

Notably, the results of the Bayesian analysis accord well with the conclusion that that can be drawn from the above graphs, that the varying span model provides a better account of the data, having closer fits to the observations for both the within-trial and predecessor FP curves. On top of what can be concluded from this simple inspection, the weight of evidence in favour of this model measured by the Bayesian analysis assures us that this result is not due to overfitting.

4. DISCUSSION

4:1 Summary of work

In this chapter, we have performed a Bayesian analysis of new models of simple responding, comparing them to a range of commonly used models. The new models, which fundamentally extend the evidence accumulator framework of decision making, distinguish themselves from their forebears in three related ways. First, and most importantly, they simultaneously address both the random and systematic variability evinced in SRT data, accounting for both the marginal response time distribution and (to varying degrees) foreperiod effects. Second, in order to make these predictions, the new models abandon the assumption that RT can be considered an IID random variable, instead treating between-trial variability – usually attributed to random processes – as the result of tightly constrained systematic processes. Third, as a consequence of abandoning the IID assumption, these models are of considerably greater computational (as well as conceptual) complexity than the norm, with PDFs to predict the outcomes being calculated individually for

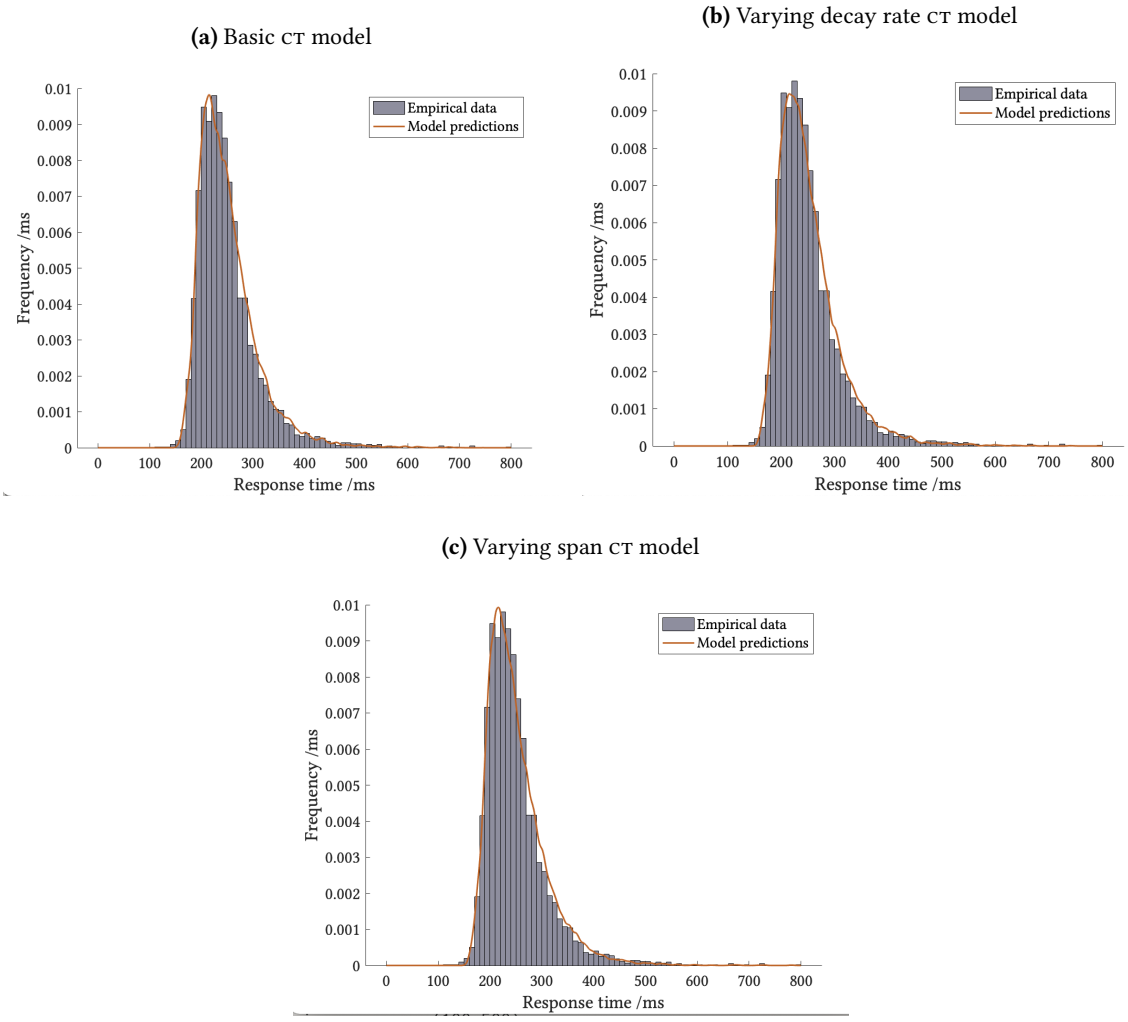


Figure 9.7. RESPONSE TIME DISTRIBUTION FITS OF THE BEST-PERFORMING MODELS. In each plot, the histogram (shaded blue) represents the RT distribution marginalised across all subjects; the gold line is the average, across subjects, of the kernel density estimate of the RT distributions predicted by the indicated model taking the maximum *a posteriori* values of the parameters as returned by MULTINEST.

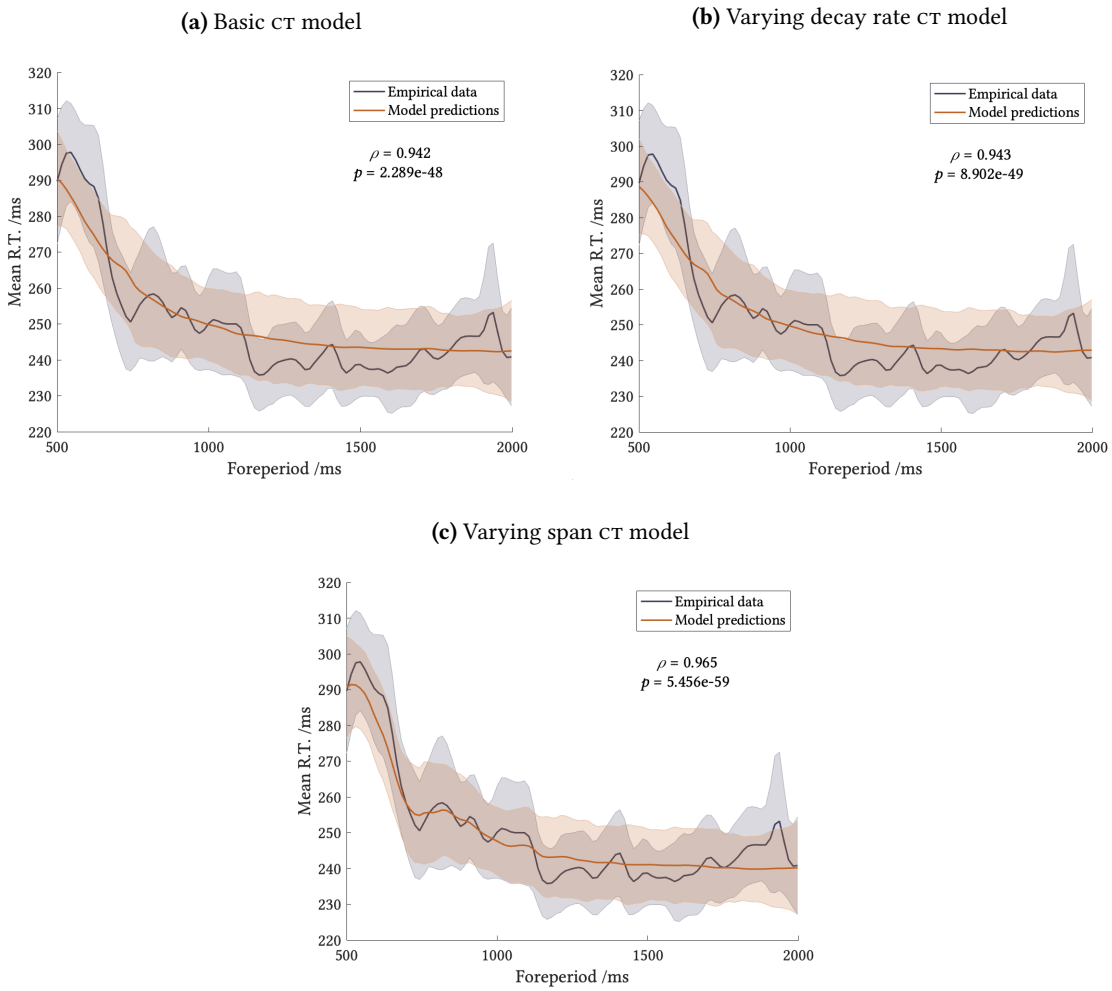


Figure 9.8. FOREPERIOD CURVE FITS OF THE BEST-PERFORMING MODELS. In each plot, the blue lines represents the average within-trial FP curve across subjects, calculated using kernel density regression with a Gaussian kernel; the shaded area represents the standard error of the mean RT thus calculated. The gold line for each model is obtained in two steps. First, for each subject, the average FP curve of 500 runs of the model – each simulating a single run of the task – is calculated. These individual FP curves are then themselves averaged to give the mean prediction of the model across subjects. As with the RT plot, figure 9.7, the models are run using the maximum *a posteriori* values of the parameters as returned by MULTINEST. The ρ and p values represent the Pearson correlation coefficient and associated p-value when comparing the blue and gold lines, that is, the mean subject FP curves as observed and as predicted by the model.

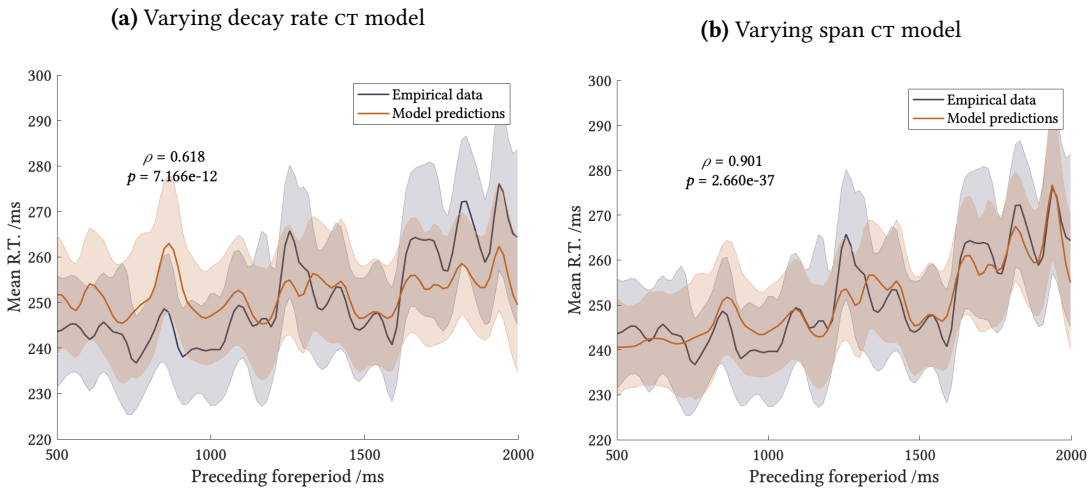


Figure 9.9. PREDECESSOR FOREPERIOD CURVE FITS OF THE BEST-PERFORMING MODELS. In each plot, the blue lines represents the average predecessor FP curve across subjects, that is, the mean RT taken against the foreperiod on the preceding trial. These are calculated in a similar fashion as described for the the standard FP curves in figure 9.8, Once again, the ρ and p values represent the Pearson correlation coefficient and associated p-value when comparing the blue and gold lines, that is, the mean subject curves as observed and as predicted by the model. The predictions of the basic collapsing threshold are not shown since it is fundamentally incapable of producing sequential foreperiod effects.

each trial.

The models were analysed against data collected from multiple subjects who undertook a simple response time task; the fashion of the data collection was typical for cognitive experiments (see chapter 7). Posterior estimation and model comparison were performed using MULTINEST, the advanced nested sampling algorithm for performing Bayesian model analysis introduced and validated in earlier chapters.

This work set out to achieve two main goals. The first was to demonstrate MULTINEST's utility to cognitive science by applying it to solve an outstanding problem in cognitive science; the problem is an exemplar of the kind of model selection commonly encountered in the field and involved typical cognitive data and sophisticated cognitive models. The second was to test whether the new models of simple responding (that encompass both RT distribution and FP effects) can better account for the RTs observed in the SRT than extant models can, and *ipso facto* to test the viability of the common simplifying assumption made in RT modelling, that RT can be treated as an IID variable.

The principal finding of this work is that one of the most complex of the newly introduced models of simple responding, the *varying span collapsing threshold* model, provides a far better account of the data – as determined by the Bayes factor – than other models when weighed against all the available data, and furthermore that this result holds for a clear majority of individual subjects. The superior model is set apart from the others in that it was deliberately designed to

take into account many of the known phenomena in simple responding, especially the foreperiod distribution and asymmetric sequential foreperiod effects.

With reference to the main goals of this chapter, I argue that these findings have at least three significant ramifications. First, the extent of the superiority of the collapsing threshold models – especially the varying span CT model – to those traditionally employed renders their use in modelling the RT distribution of simple responding not just preferable, but scientifically necessary (until even better models can be found, either in extant models not considered here or in new models not yet developed). Second, the IID assumption in RT work generally cannot be sustained, even as a knowing simplification; future RT modelling work should therefore abandon this assumption. Third, the use of MULTINEST in cognitive modelling is vindicated and its further use recommended. These conclusions are now developed and fleshed out in full.

4:2 On the use of collapsing threshold models of simple responding

The first claim – that, currently and compared to extant models, the use of collapsing threshold models to model SRT data is not just preferable, but scientifically necessary – rests on two grounds. The first is that the Bayes factor provides the most principled way of performing model comparison, and as such that a Bayes factor in favour of a model should be taken as constitutive, and not just indicative, of that model's superiority *qua* evidentiary support. Second, large Bayes factors mean much greater support from the evidence, and evidentiary support is the scientific criterion for adopting a model. In the present case, then, the fact that the CT models (and the varying span version in particular) obtained much higher Bayesian evidence in many cases (amounting to absolute superiority, in the sense mentioned earlier) should be interpreted to mean that this model provides a far better account of the data than the alternatives, and so – amongst the alternatives tested – its use is mandated by the data itself.

This claim deserves a few caveats. First, it should be pointed out that not every extant model of response time was considered here, and many variants of the new classes of model introduced were not considered. It is possible that there exists a simple distributional model that can perform as well as the more complex models, though that seems unlikely given that numerous popular and simple models of RT, such as the ex-Gaussian, generally fared very poorly in comparison to the new classes of model. Of the new classes of model, FPPD models were not tested in forms which could account for all aspects of the known foreperiod effects; however, where the models were designed to account for the same features in the data, the CT models did consistently perform better. In particular, the basic CT model and both FPPD models are meant to account for the FP distribution effect only, and the basic CT model consistently outperforms the FPPD models. It thus seems unlikely that the extension of the FPPD framework to account for ASF effects would close the explanatory gap sufficiently.

(We note, in passing, that this situation serves as a prototype for a potential use of MULTI-

NEST in model development: where multiple extensions of a model are available, the inference framework might be used to weed out the weaker contenders at earlier stages, preventing an combinatorial explosion of possibilities from emerging which would all have to be tested.)

Second, while the varying span CT model was selected in 15 of the 18 subjects, and CT models in general fared better than the alternatives, the data for a few subjects roundly supported other models. Most surprisingly, the data of subject 18 is best accounted for by an ex-Wald model, and that by a significant margin; especially surprising is that this subject's data, as shown in chapter 7, appears to exhibit foreperiod effects which the ex-Wald model fundamentally cannot account for. (Meanwhile, the data from the two subjects mentioned in that chapter without very evident FP effects still supported the more complex models; we note these anomalies here as being worthy of closer inspection, especially if more data about similar anomalies could be collected, but – given the paucity of current data in this regard – offer no other interpretation.) Less surprising than the support for the ex-Wald model given by the data of subject 18, but still notable, are the selections of the FPPD models in two subjects. This result suggests that it might be the case that different subjects employ different cognitive processes in the performance of the SRT, and that the ultimate theoretical account of simple responding will have to admit varying models – and not just varying parameters – when addressing the question of individual differences. In terms of the language mentioned earlier and introduced in appendix C, it may be that a mixed model ensemble outperforms a pure one. The relatively small number of subjects in this study makes it difficult to determine the answer to this question with the current data, but future work might fruitfully explore this question.

Finally, the adequacy of the new models presented here, while representing considerable improvements on existing models of RT, should not be overstated: they are still in many ways theoretically unsatisfactory. The most important shortcoming of these models is that they are adapted towards explaining the data in an SRT task with uniform FPD, and it is unclear how the model could be organically extended to account for other FPDs. Recalling the discussion in chapter 4, the means by which the FPD effect is accounted for is essentially *ad hoc*; no natural mechanism by which the overall FPD distribution can be learnt from experience, and translated into sensitivity to temporal hazard, is provided by these models. Less importantly, but still relevant, is that the current models do not provide space to account for the phenomenon of corrected responding reported in chapter 7. Future work should attempt to find a more principled account of these phenomena. Despite these caveats, it should not be ignored that the varying span collapsing threshold model represents a significant improvement over previous models of response time.

4:3 On the IID assumption in RT modelling

Notwithstanding the results for subject 18, models which adopt the IID assumption for RT trials fare poorly when tried against the evidence using the Bayes factor and when compared to those

models – both of the FPPD and CT classes – which do not make the IID assumption. The best performing IID model, the ex-Wald model, has on average ~ 8 deviance units of evidence (strong evidence) against it per subject when compared to the better of the two FPPD models; and and average of ~ 22 deviance units of evidence (very strong evidence) against when compared to even the worst performing CT models. Put differently, including mechanisms for systematic intertrial variation in RT affords a great deal of additional predictive power. When considered alongside the ample empirical evidence of intertrial foreperiod effects in the SRT, these results strongly imply that the IID assumption cannot be sustained when attempting to model SRT performance.

Future modelling work in the SRT ought, therefore, to abandon the IID assumption, even as a knowing and useful simplification. This, in turn, requires a change in the general attitude towards response time modelling work as expressed before: it is not the case that effects which result in systematic intertrial variability can "generally assumed to be of negligible impact and [...] therefore ignored" (Whelan, 2008). Nor is the sufficient criterion for fitting RT data that the "successful model must produce the correct shape for the RT distribution [...]" (Ratcliff & Van Dongen, 2011). Rather, cognitive modelling should be attempted with the aim of capturing the full richness and complexity present in the data. (This argument will be expanded on significantly in the next and final chapter.)

This is not to diminish the scientific importance of the IID-based models which have been used so far; after all, the new classes of model introduced are direct extensions of those older ones. To understand the scientific context for switching from the former to the latter, then, it is worthwhile to consider the reasons that drove the development of models based on the IID assumption. First, it is a common approach in scientific modelling to aim for simplicity, and it always necessary to carve off a subset of phenomena for explanation while stipulating that other closely related phenomena lie beyond the scope of the work; by focusing on the marginal RT distribution and making the IID assumption, these aims are arguably achieved. Second, there have been very real limits to the computational power that can be brought to bear in the pursuit of modelling; the IID assumption drastically reduces the computational complexity of the models while still providing *prima facie* meaningful results.

Against the first of these reasons, the results of this study demonstrate that the simplicity achieved by focusing exclusively on the marginal RT distribution through the IID assumption amounts to an *over*-simplification; they carve off too much of importance from consideration, in other words, so that the resultant models underfit the data. Against the second, the work presented here demonstrates that we are now possessed of the computational means – both in terms of hardware and algorithms – which allows us to develop more complex models. Given that the scientific means and motive to develop more complex, more powerful models now exist, the opportunity to do so should be taken.

4:4 On the use of MultiNest in cognitive science

The results obtained here were only possible because of the use of a highly efficient instrument of Bayesian inference, MultiNest. By providing such a striking result, the use of MultiNest in cognitive modelling work more broadly is vindicated: if MultiNest is capable of effectively adjudicate between competing cognitive models in this case, in combination with the more general reasons given for its adoption earlier, strongly suggest that it could be usefully applied to other problems of model selection in cognitive science, of which there are numerous important examples. For instance, linear ballistic accumulator and drift-diffusion models compete as accounts of choice responding, and independent horse race and leaky competing accumulator models compete as accounts for stop signal tasks; both these problems of model selection could be efficiently addressed using the methods presented here.

Three aspects of the analysis performed in the current study particularly recommend MultiNest (and Bayesian inference more generally) when considering it for broader use. First – regarding Bayesian methods more generally – the data used here was of a very typical scope for cognitive studies, involving a moderate number of participants contributing relatively few trials each. Even this data can be used to draw very strongly conclusive inferences regarding model selection (and potentially parameter estimation). This strongly suggests that Bayesian model analysis can be used to draw meaningful inferences within the kind of experimental paradigm that we already use, and therefore the data we already have; this means the approach has scope for wide, immediate, and meaningful application.

Second, – regarding MultiNest and other state-of-the-art methods of Bayesian inference specifically – some of the models used here were of considerably greater computational complexity than is the norm. The CT models, in particular, required that an individual predictive RT distribution be calculated using the parexic sampling method for each individual trial in the data; even so, MultiNest manages to produce successful fits. When applied to the simpler models, the fitting procedure was completed in the space of a few minutes. The hitherto unavailable efficiency of MultiNest and similar techniques opens up new avenues for scientific research, by allowing for the testing and development of models of unprecedented complexity.

Third, some of the models exhibit difficult posterior geometries; of particular note is the pronounced curving degeneracy. These are exactly the kind of difficulties that nested modern sampling methods are designed to resolve with maximum efficiency; alternative methods will run into problems of efficiently, accurately, and conveniently performing posterior and marginal likelihood estimation with such geometries (see chapters 3 and 5). The fact that these difficult posterior geometries were found with a model that does not obviate this point; the finding still demonstrates that plausible cognitive models can exhibit such difficult posterior geometries. Since such geometries cannot be ruled out in advance of actually fitting a model, it must be assumed prior to a model analysis that models might exhibit such difficulties, and the appropriate

techniques used accordingly.

With that being said, the fits of the most complex models here were still computationally costly, requiring (in some cases) days of intensive computer time to complete. Given the success of the most complex models, it would be highly desirable to bring the computational cost of evaluating them down to more normal levels. While this is far from a trivial challenge, a number of potential remedies will be suggested in the next chapter.

This computational challenge should not be downplayed, but, I argue, should not be resolved by retreating to simpler models which can be more easily fitted: to do so would be to set aside the scientifically more pertinent fact, that the most complex models received, generally speaking, vastly greater evidentiary support than the simpler ones. Ultimately, while the use of MultiNest has revealed issues with the current modelling of RT and simple responding, it also provides the tools needed for their resolution.

4:5 Epilogue

This chapter has argued that sophisticated models of the SRT which account for the full spectrum of FP effects provide superior accounts of SRT data; that, therefore, future modelling work in the SRT and in RT tasks more generally should abandon the assumption of IID trials, instead attempting to bring more of phenomena seen in the data into the scope of the models; and that, through its instrumental role in obtaining these results, the use MultiNest for Bayesian model analysis in cognitive science is vindicated. The next and final chapter generalises from these results to argue for, and outline the consequences of accepting, one of the major claims of this thesis: that cognitive data exhibits irreducible complexity that is not fully understood, and that future advances in understanding cognition can be driven by adopting a combination of (a) advanced tools for inference of the sort used here and (b) new guiding principles for theoretical work based on the acknowledgement of the irreducible complexity of cognitive data.

—ω—

Part III

Conclusions

In fact, the whole machinery of our intelligence, our general ideas and laws, fixed and external objects, principles, persons, and gods, are so many symbolic, algebraic expressions. They stand for experience; experience which we are incapable of retaining and surveying in its multitudinous immediacy. We should flounder hopelessly, like the animals, did we not keep ourselves afloat and direct our course by these intellectual devices. ***Theory helps us to bear our ignorance of fact.***

— George Santayana, The Sense of Beauty

CHAPTER 10

General discussion

1. SUMMARY OF WORK

The fundamental aim of cognitive science is to provide rational accounts of the processes of cognition, an aim pursued – amongst other things – by the proposal of mathematical models of cognitive data, including those psychologically informed and interpretable models known as cognitive models. The conclusion of this thesis – and the proposition argued for in this chapter – is that cognitive data exhibits irreducible complexity that we are far from fully understanding, and that advances in our understanding of cognitive data can be caused and catalysed by adopting a combination of (a) new methodologies for model fitting and assessment in the form of nested sampling techniques of Bayesian inference and (b) new guiding principles for theoretical work based on the acknowledgement of the irreducible complexity of cognitive data.

To support this proposition, we recall that this thesis has thus far forwarded two major claims. The first, methodological claim is that state-of-the-art methods of nested sampling, exemplified here by MultiNest, should be adopted for parameter estimation and model selection in cognitive science. The second, theoretical claim is that the foreperiod parameter-dependence and especially collapsing threshold models of one of the simplest cognitive tasks, the simple response time (SRT) task, provide superior accounts of the data of simple responding when compared to standard extant models; importantly, the new models are distinguished by their attempts to account for a family of *foreperiod effects* (FP effects) discernible in well-known forms of systematic variability in SRT data. These claims are themselves supported by two converging branches of argument, summarised below.

On the methodological branch, I began (chapter 5) by demonstrating the fundamental efficacy of MultiNest for posterior and marginal likelihood estimation in a case where the analytical solutions are known. The same study also established MultiNest's competitiveness as a method of marginal likelihood estimation when set against rival techniques, especially *thermodynamic integration* and the *steppingstone* method. This competitiveness was demonstrated in two particular cases, the aforementioned analytically tractable data and a more demanding problem involving synthetic cognitive data and a popular cognitive model, the linear ballistic accumulator of choice response; general reasons to prefer MultiNest were also highlighted. Armed with this knowledge,

I next demonstrated (chapter 6) the ability of MultiNest to adjudicate accurately and precisely between alternative models of simple response time that are popular within the cognitive sciences, with reference to plausible simulated data where the ground truth regarding the generating model was known. This series of results, taken together with past evidence of MultiNest's capabilities, were taken to justify the reliability of MultiNest when applied to draw inferences from real cognitive data, which work will be covered after the survey of the second line of argument.

The theoretical branch began by identifying a shortcoming in the current theory of simple responding, namely, the absence of mathematical models which simultaneously account for both the marginalised RT distribution and for the well-known foreperiod effects present in SRT data (chapter 4). It was further argued that this gap is significant to cognitive science, not only because of the intrinsic interest of simple responding, but also because the SRT – and its demands to detect and respond to a signal – represents the nucleus of many other cognitive tasks. To begin to fill this gap, I developed two new classes of model, the *foreperiod parameter dependence* and *collapsing threshold* classes, which can account for the overlooked phenomena to varying degrees (chapter 8). These models act as extensions of the popular evidence accumulator model framework of human decision making, but, in the process of accounting for foreperiod effects, abandon the common assumption that RT can be treated as an IID random variable. This results in models of considerably greater computational complexity than the norm, since predictive distributions are required to be calculated for each trial in the data individually; extant models in general entail that only a single, global marginal RT predictive distribution be calculated. The bevy of extant models, as well as the new ones, constitute a typical problem of model selection.

The two branches converged in the final investigation reported here (chapter 9), wherein I applied MultiNest to perform Bayesian model selection from this large pool of models using real SRT data. The data (introduced in chapter 7) were typical of many cognitive experimental designs, consisting of numerous subjects each contributing a relatively modest amount of data; furthermore, given the motivation for the development of the new models, the particular form of the SRT used was designed to elicit noticeable FP effects. The principal result of this Bayesian model analysis was that the best account of the data was provided by one of the most complex of the models of simple responding, the *varying span collapsing threshold* model, and that, as measured by the Bayes factor, the evidence in favour of this model over the others is decisive.

Thus, having earlier established MultiNest's *prima facie* efficacy and competitiveness for use with cognitive data, the first major claim – that state-of-the-art nested sampling techniques should be adopted for use in cognitive science – was further vindicated by demonstrating its successful application to help solve an outstanding problem in cognitive modelling; crucially, the problem exemplified the general class of model selection problems often encountered in cognitive modelling, and used typical cognitive data and sophisticated, computationally intensive cognitive models.

The second major claim, that of the superiority of the new models of simple responding introduced here, was demonstrated directly through the Bayes factors of the new models when compared to standard extant models. It was concomitantly argued that, since models which abandon the assumption that RT can be treated as an IID variable across trials in general perform worse than those that subscribe to it, it would be imperative in future RT modelling work to move away from assumptions of trial independence and to adopt stricter standards for evaluating models, incorporating the full range of effects seen in the data.

2. ON THE IRREDUCIBLE COMPLEXITY OF COGNITIVE DATA

The first part of the proposition that this chapter argues for states that "cognitive data exhibits irreducible complexity that we are far from fully understanding." We begin by arguing that, per the results obtained in the last chapter, the data of the SRT can be said to possess considerable *irreducible complexity*.

To flesh out the notion of "irreducible complexity", we start from the notion that the Bayes factor directly reflects the predictive power of one model over another, and so by default how well the models reflect the intrinsic properties of the data being modelled. This can be said of the Bayes factor because it is defined in terms of the Bayesian evidence, which formalises a definition of the explanatory quality of a model as fundamentally being equal to its predictive power, in turn fundamentally dependent on how probable the observations are under the model.

Consider that Bayesian evidence depends only on the prior and likelihood: the prior is given before the analysis begins, and when evaluating the likelihood, all the Bayesian method can do is take a datum and calculate its probability according to the model's predictions. Assuming that there is sufficient data for the Bayesian evidence to primarily depend on the contribution of the likelihood and not the prior, then if the Bayes factor is higher for a given model amongst a set, it is only because the data in aggregate is more probable – to wit, better predicted – by that model.

But a model's predictions can only be better in this way if they better reflect the intrinsic structure of the data; thus, if a more complex model (*viz.* the one with a greater interrelation of the parts) is receives greater evidentiary support – of predicting the data better – then the model complexity must be reflecting complexity endogenous to the data itself. No other source can lend credence to the more complex model. This complexity is *irreducible* insofar as it must be accounted for by any successful model of the data; roughly speaking, insofar as the elaborations of the more complex model are necessary to produce the desired gains in predictive power.

Now, it was established earlier that the more complex models of SRT gained decisive evidentiary support over the simpler ones, per their Bayes factors, and that the complexity of the models arises from an explicit attempt to encompass the full spectrum of FP effects in SRT data. From the considerations given above, it follows that the decisive evidentiary support received

by the more complex models arises *only* from the fact that the inclusion of the mechanisms to account for systematic intertrial variation in RT affords a great deal of additional predictive power *measured directly as* how probable the data as a whole is under the model. In turn, this entails that SRT data is possessed of at least equal complexity as the model which (to date) best describes it, and that – whatever the truth about how it emerges – the systematic intertrial variability in RT embodied in FP effects are essential to adequate descriptions of the data: they are irreducibly complex aspects of the phenomenon.

It should be noted that the imputation of irreducible complexity in this case is not certain, although I argue it is very likely. The imputation could, in principle, be refuted by finding a model that does not account for FP effects which produces similar Bayesian evidence to the best-performing models that do. However, given the general separation in evidentiary support between models which attempt to account for FP effects from those that do not – which include many of the already most-popular RT models – the chances of this sort of refutation seem slim.

So much for how irreducible complexity can be identified in general and its presence in SRT data. The proposition being argued for here has two further sub-claims: that irreducible complexity is (probably) present in cognitive data more generally, and that this complexity is currently poorly understood. The first sub-claim is based on the fact that the irreducible complexity demonstrated here has been found in the context of *one of the simplest* of all cognitive tasks. There are two reasons to believe that the insights about the importance of irreducible complexity in SRT data will generalise to other cognitive tasks.

On a theoretical level, as was argued for extensively earlier (see §4.1), the SRT forms the nucleus of many other cognitive tasks, since *inter alia* almost every task requires the ability to detect and respond to signals; since these are the only imperative demands of the SRT, it is intuitively reasonable that the cognitive processes implicated in its performance are utilised across a broad range of other tasks. If we cannot understand the cognitive processes at play in the simpler task, it stands to reason that we will not be able to understand the processes in more complicated tasks which build on it.

On an empirical level, there are well-known sequential effects in a range of other cognitive settings (such as choice reaction and stop signal tasks) which bear a strong resemblance to FP effects in that they create evident intertrial systematic variability in the resultant data. Given the findings presented here in the SRT, it is highly plausible that these effects indicate the presence of irreducible complexity in these kinds of data, too.

The second sub-claim – that the irreducible complexity of cognitive data is poorly understood – rests on the fact, already noted (§9.4), that despite the superiority of the varying span CT model to others, it is still unsatisfying as a general model of the SRT. It remains unclear how the model could be organically extended to account for FP distributions other than the uniform one. Further, the model still relies on a subject having implicit prior knowledge of the overall FPD; no natural

mechanism by which the overall FPD distribution can be learnt from experience, and translated into sensitivity to temporal hazard, is provided by the model. In other words, the exact nature and full extent of the FP effects in the SRT have not been well-characterised; so the complexity present in the data has not been well understood. This being the case for the SRT, and in combination with the considerations above, it is likely that the irreducible complexity of cognitive data in general is similarly poorly understood.

If this line of argument is accepted, it has important ramifications for how we study and model human cognition in general. First and foremost, the view of cognitive data propounded here impels us to an approach to theoretical work which takes the irreducible complexity of cognitive data as a serious guiding principle. The next section will lay out some of the differences that follow from adopting this view of cognitive data in juxtaposition to the views implied – or explicitly stated – in current practice. Beyond that, this view, in combination with the practical aspects of the work presented in this thesis, naturally suggests a program of research dedicated to gaining a deeper understanding of the full richness of cognitive data through more sophisticated models, and using state-of-the-art methods of Bayesian inference as crucial tools in the development of these models. The final section will provide suggestions for future work which could be undertaken under the rubric of such a program.

3. ATTITUDES AND APPROACHES TOWARDS MODELLING COGNITION

We start by outlining how the view of data as irreducibly complex and poorly understood affects the attitudes one adopts towards the meaning and capabilities of models of cognition. Recall that a major aim in theoretical cognitive science is the development of so-called *cognitive models* (Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2013), which are understood not merely as representations or descriptions of the data, but – by formalising a set of theoretical commitments about cognitive processes – are meant to provide precise mechanistic accounts of processes of mind such as "perception, memory, decision making, and other fundamental aspects of cognition" (Annis & Palmeri, 2018). This is an important scientific aim; arguably, its pursuit is essential if we are to make progress in understanding the not-directly observable processes of cognition.

Indeed, significant work has advanced the view that cognitive models can be used as "measuring tools", as some recent studies have phrased it (Evans, 2019; Evans & Wagenmakers, 2020). This refers to the idea that, by fitting cognitive models and leveraging the psychological interpretations imputed on them, we can infer the values of individual's latent cognitive parameters from those fits (Dutilh et al., 2019; Forstmann et al., 2011). While such an interpretative function is certainly attractive and potentially useful, the desire to ascribe cognitive meaning to data through mathematical models requires tempering in light of the limitations of our present understanding; evidently, if our present cognitive models are inadequate, as they appear to be, the validity of any

such interpretation is unclear at best.

Rather, the results presented here suggest that current models of cognition should be viewed as fundamentally and demonstrably limited, warranting a healthy scepticism regarding how representative they are of the cognitive or psychological causes of the observable phenomena, *viz.* the underlying cognitive components and processes. Indeed, it is not even clear that our current models are adequately representing the complexity inherent to the data; where traditionally, overfitting has been a major concern in modelling work, the fact that the most complex models provide the best account here, in the simplest of tasks, suggests that *underfitting* may be a greater concern. Even in the case of our best models, their practical role should be better understood as *engines of prediction*, and their theoretical significance construed as wayfinders to deeper theories of the laws governing cognition.

4. FUTURE DIRECTIONS

Finally, we turn our attention to future work that could be done inside the program of research suggested above, based on adopting (a) new methodologies for model fitting and assessment in the form of nested sampling techniques of Bayesian inference and (b) the view that cognitive data exhibits irreducible complexity which requires further investigation and explanation.

4:1 Modelling and model comparison in other cognitive tasks

As has already been stressed, the complexity of the models required to provide the best accounts of the SRT data used here casts doubt on the sufficiency of models used to explain performance in more complex tasks, many of which presumably rely on the same or similar mechanisms as the SRT. Most immediately, this forces the need for the reassessment of our models of more complex tasks, such as the choice reaction and stop signal tasks. Models such as the Ratcliff drift-diffusion model for choice response and the Logan horse race model for stop-signal tasks make many of the same simplifying assumptions in the models struck out by the model selection process here (Logan & Cowan, 1984; Logan et al., 2015; Ratcliff et al., 2016). This suggests a fruitful line of future work extending the methods and ideas developed here into other cognitive domains.

4:2 Simultaneous modelling of and inference from multiple tasks

The argument that there are cognitive processes that are shared across related tasks suggests another avenue of future research, based on building models that capture how those cognitive systems operate across different contexts; more plainly, to build models which can simultaneously account for an individual's performance in, say, all three of the SRT, the choice reaction time task, and the stop signal task. Relatedly, models could be built which provide principled accounts of

how an individual's performance varies across different versions of the same task (e.g., under a change of the foreperiod distribution), accounted for by some meaningful mechanism for the variation of model parameters across the myriad versions of the task. This avenue of research would also require extending the Bayesian inference framework to be able to properly calculate the Bayesian evidence, taking into account the fact that the data against which the models are tried are drawn from the multiple sources of measurement (tasks).

4:3 Modelling in cognitive neuroscience with multimodal data

Extending the theme of multiple sources of measurement, there is naturally a great interest in understanding the relation between cognitive processes (strictly construed) and neural processes. This has led to the rise of model-based cognitive neuroscience, which attempts to relate cognitive models to brain activity, or – even more ambitiously – to build explicitly neural models of complex cognitive processes (Kriegeskorte & Douglas, 2018; Turner et al., 2016). In order to assess these models, and thence to draw inferences about how neurological systems support cognition, it will be necessary to extend the Bayesian framework presented here to allow proper inferences to be drawn about the explanatory power of models, when those models must simultaneously the data obtained from multiple, incommensurable modalities – behavioural and imaging data especially.

4:4 Relating novel behavioural models of the SRT to brain measures

Even without the development of fully fledged neural-behavioural models, the new cognitive models developed here could be extended into cognitive neuroscience research. In particular, it could be telling to explore the relationship between individual differences in the fitted model parameters from the newer, more successful models and measures of brain dynamics. Insofar as we can be confident that the model parameters represent *something* cognitively meaningful (and not forgetting the earlier cautions about the psychological interpretation of mathematical model parameters), such research could begin to reveal the neural substrates underlying that meaningful cognitive feature. Less ambitiously, finding consistent relationships between model parameters, behavioural measures, and brain measures would provide some level of evidence that a model is, in fact, capturing something meaningful about brain and cognitive function; this would be especially true if less successful behavioural models failed to show such consistent relationships.

4:5 Improving the implementation of current models

The development of new and ever-more complex models calls for increases in computational efficiency. It was mentioned in the previous chapter that the fits of the most complex models here were still computationally costly, requiring (in some cases) days of intensive computer time to complete. The principal cause of inefficiency is the expense of performing a single likelihood

evaluation. A number of potential remedies can be suggested. On a technical level, the refactoring of the parexic likelihood algorithms to be run on GPUs is likely to be helpful, since this could exploit the embarrassing degrees of parallelism inherent in the process of generating samples for the density estimation portion of the algorithms. It may also be possible to find more effective means of estimating the than the naïve approach employed here, say by using adaptive methods of kernel density estimation. Finally, it can be hoped that, despite their apparent complexity, the new models here can be recast into an analytically tractable form.

4:6 The utilisation of other advanced nested sampling algorithms

On a broader methodological note, the success of MULTINEST in this study recommends the exploration of other advanced nested sampling methods of Bayesian inference. As tentative suggestions, POLYCHORD (Handley et al., 2015) and BAMBI (Graff et al., 2012), both developed within the same nested sampling framework, offer promise. POLYCHORD extends nested sampling to very high-dimensional (>50) models, which may be of interest for the analysis of hierarchical models. BAMBI uses machine learning techniques to deal with difficult-to-evaluate likelihood functions, which may be helpful in light of the likelihood evaluation problems outlined above.

4:7 Nested sampling for other purposes

Nested sampling algorithms have been presented here as a means of calculating the Bayesian evidence. However, since the essence of the method is to efficiently explore the prior-weighted likelihood function to find those regions which have the highest values, on a mathematical level, nested sampling is just a means of finding the maxima of an arbitrary function. This insight has already seen application in materials science, where nested sampling is used to approximate the complex potential energy surfaces of atomistic systems (Pártay et al., 2021). Given that finding function maxima (or, equivalently, minima) is a general problem of very broad applicability, it stands to reason that nested sampling approaches could be applied to this end very widely. In particular, optimisation problems in machine learning applications are expressed as the minimisation or maximisation of an often-complex function, and might benefit from being tackled with this approach.

4:8 Translation to clinical application

The final suggestion for future work steps outside the purely scientific bounds we have ranged inside up to this point, and considers a practical – and potentially valuable – ramification of the key findings of this thesis. Currently, clinical cognitive assessments often involve relatively complicated tasks which may not be suitable for all patient populations, especially those in cognitive decline because of, say, dementia. However, if even the SRT exhibits rich and informative

patterns about basic processes such as attention and learning, this presents the possibility that much richer inferences could be drawn about patients' cognitive states using simpler tasks and more complex models. This means that, by fitting a patient's performance data to more sophisticated models, the resulting parameters – if associated with clinically relevant information – might be used to provide valuable clinical indicators of a patient's cognitive state.

This suggestion may, at first, appear to contradict the earlier cautioning against the use of cognitive models as measurement tools. However, care should be taken to distinguish the predictive and analytic functions of mathematical models from their theoretical and explanatory ones, with the corresponding differences in technique borne in mind carefully. A patient's data can be fitted to a complex model to yield useful clinical measurements – in much the same manner as when using cognitive models as measurement tools – but, crucially, useful clinical measurement need not be convincingly shown to correspond to latent cognitive constructs. Rather, they need only be useful clinical indicators and diagnostic predictors.

5. CLOSING STATEMENT

This final suggestion illustrates that, while this thesis has framed the value of more sophisticated models of cognition in terms of scientific necessity, such models may well turn out to have widespread practical value, too. Even so, to the sceptic, adopting the view of cognitive data as irreducibly complex, and the new, more sophisticated models and techniques of model inferences they entail as illustrated in this thesis, may initially seem like a great deal of over-complication. After all, much progress has been made without the adoption of these views, models, and techniques. However, the ultimate aim of cognitive science is to provide rational accounts of the processes of cognition as they really are; and, I argue, cognition really is complicated.

Progress towards the ultimate aim of the science depends on reconciling ourselves with the actual nature of the cognitive processes under study, and using techniques appropriate to the demands of the data acquired and models built in the course of their study. It is hoped that the argument presented in this thesis shows that, fortunately, equipped with current understanding and techniques, there is great potential for progress towards this end; and more modestly, it is hoped that the suggestions advanced here can serve – however little – as cause and catalyst of this progress.

—ω—

Acknowledgements

Theses are funny things. From the perspective of those who write them, they can be all-consuming; from the perspective of science and other scientists, they're drops in the ocean of knowledge that will anyway soon evaporate. For the student, successfully presenting the Ph.D. is a moment of triumph, the culmination of a long process of learning and recompense for the effortful expenditure of so much of his or her creative and critical mental energies; for the academy, it represents little more than the hurried passing of a newcomer through some of the outer gates around the ivory tower. There's the rub: the thesis is decreed to be absolutely necessary, and destined to be utterly ignored.

No matter. I hope the work in this thesis has some lasting meaning beyond me, but, even if not – and despite all the stresses – it was a worthwhile experience. That I had the experience I had can be attributed to the people around me; I must thank them, though my gratitude runs deeper than I feel I can adequately express. I only hope these few words are enough.

To my supervisors, Professors Adam Hampshire and Claudia Clopath, thank you for the guidance, support, and encouragement over the years. Dr Bojan Nikolić has my lasting gratitude for introducing me and Adam to MultiNest, as well as for the time he spent answering my naïve questions about Bayesian inference and helping me with various technical issues. To my examiners, Professor Payam Barnaghi and Dr Tristan Beckinstein, thank you for a stimulating and enjoyable *viva voce*. I'm glad I managed to win you over.

Thank you also to the many friends and colleagues who have helped me: from the Hampshire Hub, Maria Bălăeț, Eyal Soreq, Richard Daws, and William Trender deserve special mention; from the Clopath Lab, Jacopo Bono, Sara Zannone, and Wilten Nicola. Nicola Davis, thank you for helping me through a very difficult stage early on; it will be forever appreciated. To the others, too numerous to name: Liam Brown, Alex Morton, Amy Bryce… You all take a share of this. Dodo, your memory does not fade.

To the therapists of the Imperial College Counselling Service: your work matters; your work helped. Thomas Gibson, for the example you set with your diligence and calm, for the intellectual stimulation, and for your listening ear, I am grateful. Diana Lucaci, I will never forget that it was you who helped me onto this path, nor that you always let me know how you believed in me. Lotte Weerts, for your unfailing support; for the many kindnesses – and kicks in the ass; for holding me accountable, especially at the end – thank you. Thore Mainart Bücking, in our

two years living together and many more as friends, past and future, I have benefited from you companionship and sanguine wisdom; they helped me through this all. Reuben Nixon-Hill, for the example of your faith and courage and compassionate concern, which I stand in awe of, and which helped me keep *some* perspective in all this, you have my praises. Ioana-Raluca Tănase, for everything.

Baka, your memory is inextricably linked to this period of life in my mind. I feel proud to come from this heritage of scientists. You will be in my thoughts, always. *Ja te volim.*

Finally, to my parents and sister: I don't know how to thank you enough. Ayna, your unflagging belief in me and my abilities, so far beyond what I was able to muster up myself, gave me the energy I needed to keep going countless times. Tata, for all the advice I failed to listen to; you've still taught me more than you can imagine. Mama, I really don't have the words.

Needless to say, any mistakes which are in this thesis – and I know they're there – can and should be laid squarely at my feet. And finally, thank you, dear reader.

—ω—

Appendices

A. THE ORIGINS AND OBSTACLES OF MODELLING

Author's note: This supplemental essay represents the author's working out, in his own thoughts, the difficulties which science faces in general when trying to move from data, to theory, to understanding. It is by no means conclusive or authoritative, but is provided in the hope that it might better contextualise the concerns of this thesis about proper inference, and that it might lead more students and scientists to think about these deep and fascinating issues. Interested readers are directed to Chalmers (2013) for an excellent introduction to these issues in the philosophy of science. The references to pieces about inductive logic in the Internet Encyclopedia of Philosophy and Stanford Encyclopedia of Philosophy given at the start of Chapter 2 may also be consulted (Hawthorne, 2021; Henderson, 2020; Huber, 2022; IEP Staff, 2022).

It is a basic precept of science that data – accurately reported information obtained via well-specified procedures of observation – comprise our only certain and direct knowledge of the external world. Data are not, however, self-explanatory. That is, the mere possession of data does not by itself confer an adequate understanding of the phenomena the data record; nor, though we wish to *infer* some such understanding from the data, can it be directly *deduced* from them. For ultimately, the definite content of a dataset amounts to nothing more than a series of hyper-specific statements of the form, "*This* happened *then* under *these circumstances*," and no power of deductive logic is capable of alchemically transmuting statements from specific to universal in nature, however approximately. By itself, data is important – but inert.

We nevertheless perceive regularities, structure, and correspondences in data that lead us to believe in some lawful, comprehensible, and persistent data-generating process which underlies the data *qua* pure information – a process which cannot be deduced from the data, for sure, but *from which*, if its principles were understood, we take it the data *could* be deduced. Where such intriguing hints of inner order are seen, we naturally wish to describe this underlying process – or something like it, or, at the very least, the characteristic patterning of data that results from the process – since obtaining this description satisfies the basic curiosity which motivates us to science (or, more pragmatically, since a well-specified description allows us to make useful predictions).

Yet if data cannot yield answers directly, how are we to proceed in creating such descriptions? To move beyond the borders of data alone and into the region of understanding, we must – in cognitive science as elsewhere – engage in speculative theoretical work, the chief aim of which is to provide rational, conceptually clear explanations and descriptions of the phenomena at hand, however complex they outwardly seem. In particular, when the data are quantitative, theorising means mathematical modelling, *viz.* imagining an adequate mathematical representation of the data, or more ambitiously, of the natural processes which putatively underlie the observed phenomena.

Modelling, while necessary, presents unavoidable challenges. Two intimately related issues will be of particular interest to us here: the problem of quantifying and qualifying the degree of support the data lends to a model; and the problem of determining the proper degree of complexity for a model. Let us examine these in turn.

The most basic obstacle in modelling is the challenge of establishing and applying effective *laws of inductive inference*. Let us give these words some content. Recall the basic precept that data are our only reliable sources of knowledge of the processes which give rise to them. However, as already noted, a dataset is *per se* particular, tightly bound to the specific circumstances under which it was collected, whereas the supposed underlying process which begat the dataset is presumably general, applying to all similar circumstances. This disjuncture between the natures of underlying processes and the data they give rise to forms an issue: how does one pass with any degree of confidence from the implicit knowledge embodied in data, which is concrete and bounded, to explicit knowledge of the underlying processes, which must be abstract and general?

We are taking it for granted that we are at least able to conceive of possible explanations for the observations we see, namely, to propose hypotheses about the general underlying processes. Once equipped with hypotheses, in response to the issue at hand, it seems self-evident that there should be some consistent and logical way of making statements about the validity or plausibility of a hypothesis in light of the available data, despite data's inherent narrowness – grading some hypotheses as good and others as bad, or at least some as better and others as worse. More precisely, it seems that there must be a way of reasoning about hypotheses in conjunction with data in order to arrive at some sort of conclusion about the degree of *evidentiary support* the hypotheses receive from the data. We call such reasoning *inductive reasoning*, and the rules which govern correct reasoning of this sort *laws of inductive inference*.

If data were exact – a perfect record of the object under study, and nothing else – the basis of inductive inference for mathematical models would be relatively straightforward: the degree of support that a model receives from data would be in direct proportion to the closeness of the quantitative fit between the model and the data. With ideal data, any deviation between a model's predictions and the observations would indicate a flaw on the model's part, and it is natural to

take as axiomatic that larger deviations represent greater flaws in the model, or equivalently, less evidentiary support.

Even under these ideal conditions, the fact that it is impossible to deduce a model from data means that the inference that a model is "correct" could never be made in an absolute sense. It is always possible to produce multiple models which fit the data equally well, simply because the data available will always be finite, and the number of predictions a model is capable of making, infinite: models will always make predictions regarding future or other unseen data, which of course have not been collected (yet). As products of the imagination, models can, with sufficient creativity, always be modified or generalised to generate new models which make the same predictions of the data actually observed, while differing in some detail about something(s) yet to be observed. Such modifications and generalisations may come from good reasoning or simply be *ad hoc*, may please or trouble us, may seem plausible or ridiculous; but in any case, it must be admitted that there is always an outside chance that an alternative model, whatever violence it does to our good sense and however unlikely it seems, will turn out to make the correct prediction of an as-yet unrealised observation in a way that our preferred hypothesis simply cannot.

(This grounds the scientific truism that our current understanding of things must always be understood as provisional – see the quote from Russell's *Skeptical Essays* in the epigraph to the Foundations section. We have talked of definite alternatives obtained through the modification and generalisation of hypotheses, but this logic taken to its fullest extent demands the awareness that an as-yet unimaginable hypothesis might be required to account for as-yet unseen observations – observations which in light of current understanding would seem absurd. The history of science is replete with examples of surprising empirical results resulting in such dramatic shifts in understanding, the quantum revolution in physics being perhaps the best-known.)

So, though data can be used to determine *some* measure of the degree of support for the model by how well the model approximates the data, and more data can be helpful in restricting the possibilities, we are still in principle left with a potential infinity of models capable of accounting for the data to the same degree. How, then, are we to retreat from the uncertainty inherent in these endless possibilities? To arbitrate between competing models of similar quality, we fall back on the intuition of Occam's Razor: "entities should not be multiplied without necessity." That is, given two models which equally well account for the available data, the model which uses fewer "entities" (often cast even more vaguely as the "simpler" model) should be preferred until such a time as sufficient information is found that can apparently only be accounted for by a more complex model. "Entities" or "simplicity" are usually taken to mean fewer parameters, assumptions, or other identifiable elementary components, though the exact meaning is often a point of contention.

These difficulties notwithstanding, we can see that the process of inference would be quite

comprehensible – if not always easy to carry out – if our observations of the world were reliably exact: the subset of models with the closest fit would be preferred, with the appropriately defined "simplest" model amongst the subset being considered the current best. In practice, however, data is not exact. Measurement is always an operation, and the measured value a proxy for the object of measurement, not a perfect copy of it. Through measurement, all data is subject to some degree of indeterminate variation – random effects due to measurement error, noise, sampling effects, and so on – which obscures, to some extent, the "ideal" observation behind a veil of irrelevant happenstance. Mere intimacy of quantitative fit cannot, therefore, be taken as the sole factor of evidentiary support: models endowed with sufficient flexibility (whether through the addition of more parameters, or more complex functional forms, or more elaborate modifications) can be made to conform arbitrarily closely to the data, resulting in a model that accounts for the minutiae of the aberrations, anomalies, and other uncontrolled aspects of the data. One of the key challenges of scientific inference, then, is assessing theories in light of this inherent and ineluctable uncertainty in our observation of the world.

The second challenge is one of finding the appropriate complexity of a model. Broadly speaking, a good – a useful – model captures all the essential features of its target system inside the simplest possible rational scheme. However, as the phenomena susceptible to scientific investigation have increased in complexity, it has become correspondingly more difficult to discern whether a given model constitutes a "good" explanation or description of its target. This is especially true in fields, such as cognitive science, which deal with inherently limited and varied data sets that nevertheless reveal the presence of complicated phenomena; and in fields in which the theoretical and mathematical frameworks appropriate to the phenomena are not yet agreed upon.

This kind of complex data poses a number of significant obstacles in the pursuit of modelling. The more complex the system under investigation, the more difficult it is to identify the "essential features" of that system; to put it another way, it becomes less clear when the data *as a whole*, as opposed to in some particular aspect, have been well captured by the model. In a typical cognitive science data set, many interlaced phenomena will be found, and it will rarely be clear which of these phenomena are essential and which consequential features of the underlying system. Further, there is no guarantee that all the important features of the system will be found after analysis; in modifying a model to explain one aspect of the data, we may unwittingly do damage to a model's ability to explain some other part of the same data.

It is often remarked that models are intended to simplify, but the consequences of this are not always spelled out. Most meaningful amongst them is that, when considering data in terms of a given model, that data is by necessity interpreted entirely through the devices available inside the model; this means that any properties and relationships which are not expressible in terms of the framework that the model provides are lost. Now, this can be highly desirable: it

turns out, for example, that Earth's property of being the only known planet bearing life is not relevant to understanding its orbit about the Sun. In fact, it turns out that the situation is even more remarkable than that: equipped with the proper knowledge of the relationships between them, just three properties for each each planet effectively account for the motion of all the planets around the Sun: their mass, their positions, and their velocities. Whatever else may be going on in the Solar System – and there is a lot, not least of all, all known cogitation – it is of little consequence to the motions of the heavenly bodies. A model is meant to do away with irrelevancies and leave us only with what is essential.

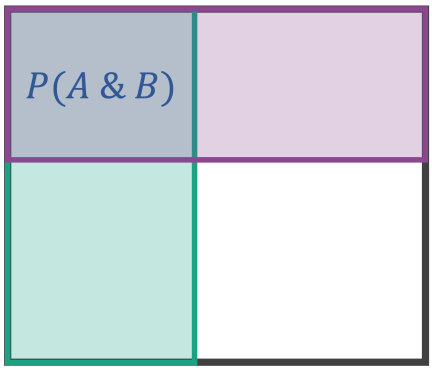
Once an effective model is obtained, this ability to encapsulate all that is relevant and discard the rest is powerful for the analysis of novel situations. Indeed, this essential feature of models gives models their power, but it also acts as a defence against the discovery of effective models. Before we have a model to guide us in doing so, how are we to discern what are the relevant and irrelevant, essential and coincidental features of the system? The modeller is left in a perpetual Catch-22: We need a model to inform us which aspects of something can be carved off for relatively independent study; we need to know what things can be subjected to relatively independent study to build the models.

These challenges exist in cognitive science, and such conditions make for an environment in which it is straightforward to come up with multiple – and potentially vastly different – models of the selfsame data. The major difficulty facing modelling work in cognitive science, then, is not with the *creation* of models, but with their *assessment*. Alas, as we have seen, there is no absolute, objective measure of model goodness when referenced to the data alone; we cannot say how well a model describes the world. We therefore require techniques which bestow the capacity to *effectively adjudicate between competing models* with respect to their (shared) target data, and in a way which accounts for the complexity present in that data. This is where Bayesian statistics plays it piece.

B. A VISUAL PROOF OF BAYES' THEOREM

The largest box in figure A.1 is a rectangle of unit area. Imagine that each point in the rectangle represents some unique basic observable outcome, such that the whole box represents all possible observable outcomes in the situation under consideration. The outcomes in box A are all those that satisfy some conditions C_A; this collection of outcomes is called the *event A*. Similarly, event B outcomes (box B) satisfy some other condition C_B. The outcomes in the region of overlap, in the top-left corner, satisfy both conditions; we will refer to this region as "$A \,\&\, B$".

Now, by definition, the probability of A, $p(A)$ is just the proportion of the total outcomes which satisfy condition C_A, i.e., the area of box A, since the large square of all outcomes has



$$P(A \mid B) = \frac{P(A \,\&\, B)}{P(B)} \quad \textit{(def.)}$$

$$P(B \mid A) = \frac{P(A \,\&\, B)}{P(A)} \quad \textit{(def.)}$$

$$\therefore P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

!tbp

Figure A.1. A simple visual proof of Bayes' theorem. See main text for full explanation.

unit area; likewise for B. Furthermore, the probability of A *given* B, $p(A \mid B)$, is by definition the proportion of outcomes in B which *also* satisfy A; evidently, this is the ratio of the area of $A \,\&\, B$ to area of B. As indicated in the figure, by applying the same logic to reckon out the calculation of $p(B \mid A)$, and doing some simple algebra, Bayes' theorem appears as a direct consequence of these basic definitions.

C. INFERENCE FROM MULTIPLE DATASETS

It is common in cognitive science (and elsewhere) for data of the same kind to be collected from multiple independent sources drawn from a population. To choose a pertinent example, we may collect a dataset consisting of a few hundred response times from each of a dozen or two subjects.[i] When the complete body of data can be partitioned into multiple sources in this manner, it is often incorrect to aggregate the data into a single set to act as the basis of model inference, but it may still be desirable to draw inferences from the totality of the data available (see, e.g., Chapter 5 of Farrell and Lewandowsky, 2018).

In particular, assume (a) that there is a phenomenon for which the data is recorded across multiple sources in the manner described; (b) that there are two competing models of that phenomenon, where each model's predictions are intended to account for the observations of a single source; and (c) that it is expected or known that there will be variability in the models' parameters from source to source, rendering it impossible to fit all the data simultaneously under a single model (as two Gaussian datasets with differing means and standard deviations could not

[i] Subjects as independent sources of the same kind of data are one obvious example. Less obvious examples include data from different blocks in a task (each block being a source), data from different task or treatment conditions (each condition a source), and so on. As an aside, the significance of this for the testing of whether different factors effect data, *à la* the traditional ANOVA of frequentist statistics, is obvious.

be effectively fitted in the aggregate by a single Gaussian curve). In such cases, it nevertheless seems sensible to ask for which of the two models there is greater evidentiary support, taking into account the evidence available across all the sources. However, the methods for model comparison developed thus far, while capable of weighing the evidence in favour of one model or the other in the case of each source, are insufficient to provide a direct answer at this combined-source level. The principal aim of this section is to develop the required technique for accomplishing this kind of combined inference from multiple sources.

The first step in this development is an observation about the structure of the hypotheses being compared. In the initial statement of the problem, the evaluation being made is whether one model (A) is better supported across all the available data than another model (B). However, each model is (by presumption) designed to account for the observations of a *single* source's dataset, and the techniques thus far developed provide measures of the support that lone dataset provides to the model. We must mean things in a different sense when we talk of a model accounting for the data from *multiple* sources, and ask for a measure of the support those many datasets lend to it.

Deeper consideration makes it clear that – since we are treating each source's dataset separately, fitting the candidate models to each one individually – when we ask which model is better supported by the evidence overall, the "model" in this notion of "model selection" is really of a composite nature. In particular, the statements being weighed are *not* about models A and B directly *per se*, but collections or ensembles of models, where each source is allocated or assigned a particular model. Namely, the statements being examined are of the form "model x accounts for dataset 1; model y accounts for dataset 2", and so on . To talk of *a* model explaining the whole body of datasets is shorthand for the hypothesis that each source has the same underlying mathematical description (specific parameter values notwithstanding). In concrete terms, the hypothesis that model A is better supported than model B across all the available data amounts to the claim that, if we were to independently fit model A to each dataset, then model B to each dataset, where each analysis gives a measure of the explanatory quality of the model, model A would outperform B overall when comparing some appropriate composite of these measures.

For the sake of ease, we will henceforth refer to the complete body of available data – which is to say, the set of datasets – as the *data corpus*. The series of models allocated to the source datasets will be called the *model ensemble*. The question at issue, then, can be phrased more precisely as follows: Given a data corpus and competing model ensembles intended to account for the evidence in the corpus, which ensemble receives the highest degree of evidentiary support?

Clearly, answering the question posed requires a method with which, once each source is fitted and analysed separately, an inference can be drawn from the whole data corpus by combining the results of the independent inferences for each source into a measure of the evidence for the ensemble. In this section, I offer a short proof, using Bayes' theorem, that the correct method for

finding the total evidence for a model ensemble from a data corpus composed of independent sources is simply the product of the Bayesian evidences found for each source. Formality aside, the proof can be intuitively grasped by noting that the Bayesian evidence of a model is the model likelihood – the probability of observing the data at hand under that model (given a prior over the parameters). Since data corpora are composed of multiple independent sources, the probability of observing all the data in a corpus, given a model ensemble, must be the product of the probabilities of observing the data in each source separately. Each such probability is simply the probability of observing that data under the assigned model, i.e., the Bayesian evidence.

To give the proof proper, we begin with some preliminary definitions which will be used to frame the problem in formal Bayesian terms. We denote the n datasets we have, one per source, as $D_1, D2, \cdots D_n$, and the corpus as a whole as $\mathbb{D} = \{D_1, D_2, \cdots D_n\}$. For the source indexed by i, the model allocated to the source will be denoted M_i, the model ensemble by $\mathbb{M} = \{M_1, M_2, \cdots M_n\}$. The evidence scores obtained in the individual analysis of source i will be written $\mathcal{Z}_i = p(D_i \mid M_i)$. Recalling Bayes' theorem as written and interpreted in equation (2.4), we frame the problem of drawing inferences from multiple independent sources in the following manner. First, we observe that our hypothesis H is the statement, "The model ensemble $\mathbb{M}$ accounts for the data in the corpus $\mathbb{D}$", which is reducible to the conjunction of the series of statements, "Model M_1 accounts for the data in D_1; and model M_2 accounts for the data in D_2; and (etc.)." Second, the evidence E is simply the information contained in the data corpus. It follows that the posterior probability assigned to model ensemble $\mathbb{M}$ given data corpus $\mathbb{D}$ is

$$p(\mathbb{M} \mid \mathbb{D}) = \frac{p(\mathbb{D} \mid \mathbb{M})\ p(\mathbb{M})}{p(\mathbb{D})} \tag{A.1}$$

where we have implicitly and mildly abused the notation so that $\mathbb{D}$ and $\mathbb{M}$ stand not only for the data corpus and model ensemble *per se*, but also for the statements explicated above. By extension of the standard Bayesian terminology, we call the quantities $p(\mathbb{D} \mid \mathbb{M})$ the *ensemble likelihood* and $p(\mathbb{M})$ the *ensemble prior*. It follows, then, that the relative strength of evidence a data corpus $\mathbb{D}$ lends to model ensemble $\mathbb{M}_1$ over $\mathbb{M}_2$ is given by the posterior ratio

$$\frac{p(\mathbb{M}_1 \mid \mathbb{D})}{p(\mathbb{M}_2 \mid \mathbb{D})} = \frac{p(\mathbb{D} \mid \mathbb{M}_1)}{p(\mathbb{D} \mid \mathbb{M}_2)} \frac{\mathbb{M}_1}{\mathbb{M}_2}, \tag{A.2}$$

which is to say, the Bayes factor (or posterior odds) of one ensemble with respect to another is given by the ratio of the ensemble likelihoods multiplied by the prior odds assigned to them.

The problem of inference from multiple sources is solved if an expression for the ensemble likelihoods can be found in terms of known quantities. It is straightforward to show that the ensemble likelihood over a data corpus is simply the product of the model likelihoods (that is,

Bayesian evidences over each dataset in the corpus), thus completing the problem:

$$\begin{aligned} p(\mathbb{D} \mid \mathbb{M}) &= p(D_1 \wedge D_2 . \cdots D_n \mid M_1 \wedge M_2, \cdots M_n) \\ &= p(D_1 \mid M_1 \wedge M_2, \cdots M_n)\ p(D_2 \mid M_1 \wedge M_2, \cdots M_n) \cdots p(D_n \mid M_1 \wedge M_2, \cdots M_n) \\ &= p(D_1 \mid M_1)\ p(D_2 \mid M_2) \cdots p(D_n \mid M_n) \\ &= \prod_{i=1}^{n} \mathcal{Z}_i \end{aligned} \tag{A.3}$$

where in the second and third steps, we have have exploited the presumed independence of the sources.

Model ensembles can be of two kinds, which we shall call "pure" and "mixed". Pure ensembles assign the same model to every dataset in the corpus; mixed ensemble allow for different models to be assigned. This fact implies that this technique can be used to answer a specific and interesting kind of question: is there evidence for a plurality of mechanisms at play across a population, or is the evidence consistent with there being a common mechanism whose parameters simply vary from source to source?

D. PRIOR SENSITIVITY AND ABSOLUTE SUPERIORITY

The Bayes factor is critically sensitive to the choice of prior, since – *contra* to the more common understanding of a model being defined by its likelihood function alone – the prior over the model parameters forms an integral part of the notion of a model in Bayesian statistics. When the prior is vague or debatable, it is therefore common to perform prior sensitivity analyses, that is, to find the Bayesian evidence for a proposed likelihood function under numerous reasonable priors. However, for certain studies in this work, we introduce and employ a method for insuring against prior sensitivity that can work when certain relationships obtain between the maximum likelihood and Bayesian evidence.

The basis of this trick is the fact that the Bayesian evidence is, by definition, the marginal likelihood: the expected (mean) value of the prior-weighted likelihood function (PWLF). Now, assuming a fixed likelihood function and data, the Bayesian evidence will only be affected by changes in the prior. It is clear that, under these conditions, the Bayesian evidence will be maximised when the prior is strongly peaked (formally, as a Dirac delta function) at the point of maximum likelihood. To see this, consider that the average of the PWLF of this construct will evidently just be the maximum likelihood, since all the prior mass (equal to unity) is concentrated at that one point; now, if the prior is varied to include any points with lower likelihood, the simultaneous drop of the prior weighting at the point of maximum likelihood and the mixing in of lower likelihood values guarantees that the average of the PWLF under this variation will be

lower. Since no variation is possible which brings in points of higher likelihood, it follows that, fixing the likelihood and data, the maximum-likelihood peaked prior maximises the Bayesian evidence. (Compare this to the $2 \ln \mathcal{L}_{\max}$ term in the AIC and BIC, which, the reader will recall are expressed on the deviance scale obtained as twice the natural logarithm of the Bayes factor.)

Thus, for a given likelihood function and data, the maximum attainable Bayesian evidence is equal to the maximum likelihood. If one compares two models (call them A and B) with different likelihoods, and it is found that A has greater Bayesian evidence under its prior than B has maximum likelihood, whatever its prior, then model A can be said to be *absolutely superior* to model B. In practical terms, so long as the prior used with model A is reasonable, this sort of result would mean that there is no prospect of likelihood functions of the kind in model B offering better accounts of the data than those of model A.

E. MARKOV CHAIN MONTE CARLO

At the level of parameter estimation, the posterior distribution can rarely be analytically solved, since both it and the prior are often complex distributions unamenable to formulaic expression and manipulation. As a first step, one could evaluate the posterior density (or, rather, its unnormalised counterpart, the prior-weighted likelihood function) on a fine mesh in parameter space, but this approach becomes computationally impractical even for low-dimensional situations ($\gtrsim 5$ parameters, even in straightforward applications) because the number of mesh points – and so posterior evaluations – required for an accurate estimate of the density scales exponentially with the number of dimensions. Moreover, this inefficiency is greatly exacerbated by the fact that the most informative parts of the posterior, where the density is high, will be concentrated in a relatively small proportion of parameter space. Since this entails that most of the posterior evaluations made using a mesh approach will be relatively uninteresting, the key to efficient posterior sampling lies in finding ways of concentrating the sampling in the higher-density regions of the posterior – without knowing in advance where those regions are.

Faced with this problem, the standard approach to estimating posterior densities is to use Markov Chain Monte Carlo (MCMC) methods such as the famed Metropolis-Hastings algorithm. In broad strokes, the central idea behind MCMC methods is to create a "chain" of samples, where new links (i.e., samples) are added to the chain sequentially.[ii] At each step, the lengthening of the chain is initiated by drawing a point randomly[iii] from parameter space using a "proposal distribution" and evaluating that point's unnormalised posterior probability. If the evaluation shows the newly sampled point has higher posterior probability than (is "uphill" of) the current

[ii] For the mathematically inclined, a *Markov* chain because the probability of drawing sample $n + 1$ is dependent, in a particular sense, only on sample n.

[iii] That is, by chance, wherefore "Monte Carlo."

last link in the chain, the newly sampled point is added as the next link in the chain; if lower ("downhill"), the newly sampled point is usually, though not always, rejected, and the chain remains unchanged. By linking all of the uphill moves, and few of the downhill moves, the majority of samples accumulated in the chain will end up being from high-posterior-probability regions in parameter space; indeed, a little thought reveals that, with a sufficiently large number of samples, the ratio of the density of chain samples found in one small region of parameter space (approximated by simply counting the number of samples in the region) compared to another equally sized region elsewhere must be equal to ratio of the posterior probabilities of those two regions.[iv] Thus is an estimate of the posterior density built.

On repeated runs of the sampler, under identical conditions, each run will return a different chain because of the stochastic nature of the method of getting from one sample to the next. When using MCMC techniques, collecting multiple chains allows one to perform important checks for consistency.

Further, MCMC methods start by drawing samples very poorly (that is, not reflecting the target distribution well); as the chain grows larger, the later draws better and better approximate the true posterior. Convergence occurs when enough samples are obtained that to provide an adequate representation of the desires posterior. Checking for convergence is essential when using MCMC; however, convergence needs to be checked *a posteriori*.

For the same reason of early samples being unrepresentative of the target distribution, it is common for *burn-in* to be applied, whereby a portion of the unrepresentative samples from the beginning of the MCMC run are discarded.

Another putative problem with MCMC sampling is that of *autocorrelation*: the problem is that each MCMC sample is, in principle, dependent on the last, so there exists a certain degree of correlation, between the samples. This makes them unlike the ideal they're trying to approximate: a set of independent and identically distributed draws from the target posterior. The usual way to solve this problem is by first finding the so-called autocorrelation length – the number of iterations it takes for any dependence between samples to effectively disappear, and then *thinning*, or removing all the intermediate samples. (So, for example, if the autocorrelation length is ten samples, using only every tenth sample.)

Applying burn-in and thinning to the full chain returned by an MCMC methods makes the resultant samples satisfy the condition of being (approximately) IID draws from the target posterior.

F. THE LINEAR BALLISTIC ACCUMULATOR MODEL OF CHOICE RESPONSE

This appendix covers the basic ideas and formalisms of the *linear ballistic accumulator* (LBA), first described by Heathcote and Brown (2008). The LBA is a mathematical model widely applied to

[iv] Directly akin to counting histogram bin frequencies to estimate a probability distribution.

decision-making tasks which, like the Ratcliff drift-diffusion model and the models introduced in this thesis, is a member of the *evidence accumulator* class of models, in which decisions are made when a quantity, often dubbed sensory evidence, passes a threshold. The principal distinguishing characteristic of the LBA is that it conceives of choices and response times as emerging from a race between independent sensory evidence accumulator units.

To be specific, the LBA posits that, in a task with n possible responses, there exist n evidence accumulators, one for each choice; let us index the accumulators and their associated choices with the variable $c \in \{1, 2, \ldots, n\}$. On each trial, each accumulator begins with an initial amount of sensory evidence k_c, which proceeds to increase at a steady accumulation rate ρ_c until the first of the accumulators (the "winner") reaches the decision threshold b. At that point, an irrevocable decision is made to respond, according to choice associated with the winner. (Note that each accumulator is endowed with its own initial sensory evidence k_c and accumulation rate ρ_c, while the level of the response threshold b is common to all.) It is straightforward to see that, so specified, the *decision time* (the time to taken to reach the decision threshold) associated with accumulator c on a given trial j is

$$\mathrm{DT}^j = \frac{b - k_c^j}{\rho_c^j}. \tag{A.4}$$

Furthermore, to introduce the necessary stochasticity into the model, it is presumed that, on each trial and for each accumulator, the starting evidence and accumulation rate are randomly drawn. The starting evidences k_c are drawn uniformly between 0 and an upper level A, where $A < b$; the accumulation rates ρ_c for accumulator c are drawn from a normal distribution centred on μ_c and with a standard deviation of σ_c. (Note again that A is common to all the accumulators, while the parameters controlling the accumulation rates, μ_c and σ_c, are specific to each accumulator; thus, certain accumulators are more likely to be winners than others.) With the trial- and accumulator-specific parameters so drawn, the decision-time is supplemented by a constant non-decision time t_0 to determine the overall response time and response choice on a given trial j as

$$\mathrm{RT}^j = \min_c \left\{ \frac{b - k_c^j}{\rho_c^j} + t_0 \,, \quad c = 1, 2, \ldots, n \right\} \tag{A.5}$$

The model presented above allows for an analytical expression of the resulting response time and choice probability distribution (the LBA distribution), given the model parameters. Before the LBA distribution is stated, we note that since the unit scale for the sensory evidence quantity is arbitrary, and since the parameters expressed in these units (b, A, μ_c, σ_c) determine the decision time by the ratio of the derived quantities $b - k_c$ and ρ_c to one another, the resulting LBA distribution is equivalent under any scaling of these parameters. To restore a one-to-one mapping between individual LBA distributions and model parameters, it is customary to fix one of the parameters in a given analysis, usually by setting $\sigma_1 = 1$ (Donkin et al., 2009). It can thus be seen that the

LBA model with n accumulators is parameterised by $3 + 2n - 1$ parameters: b, A, t_0, μ_1, $\mu_2, \ldots, \mu_n$, $\sigma_2, \ldots, \sigma_{n-1}$and σ_n.

We conclude by covering the analytical aspects of the LBA distribution; derivations of the following can be found in Brown and Heathcote (2008). First, it can be shown that the the PDF of the response times of the subset of responses which occurs at a given accumulator – namely, the response time distribution of choice c – is given by

$$f_c(t) = g_c(u, v) = \tfrac{1}{A}\{\rho_c[\Phi(v) - \Phi(u)] + \sigma_c[\phi(u) - \phi(v)]\} \tag{A.6}$$

where $\Phi(\cdot)$ is the standard normal CDF, $\phi(\cdot)$ is the standard normal PDF, the variables u and v are defined as

$$u = u(t) = \frac{b - A - \mu_c(t - t_0)}{\sigma_c(t - t_0)} \quad \text{and} \quad v = v(t) = \frac{b - \mu_c(t - t_0)}{\sigma_c(t - t_0)}, \tag{A.7}$$

and the functional dependence on the parameters b, A, t_0, μ_c, and σ_c has been left implicit throughout. Continuing with the same notation, the corresponding CDF is

$$F_c(t) = G_c(t, u, v) = 1 + \tfrac{1}{A}\sigma_c t\left[u\,\Phi(u) - v\,\Phi(v) + \phi(u) - \phi(v)\right]. \tag{A.8}$$

Now, we are usually interested, not in the probability of obtaining a particular RT *given* the choice c, but the probability of obtaining a particular RT and choice combination. For some particular response choice c, this joint probability distribution is evidently the probability of obtaining the RT under that choice multiplied by the probability that no other response has yet occurred due to one of the other accumulators, namely,

$$h_c(t) = f_c(t) \prod_{\substack{i=1,\\ i \neq c}}^{n} (1 - F_i(t)). \tag{A.9}$$

The distribution in this equation has been termed the "defective" distribution of accumulator (or choice) c. The formulas in A.6-A.9 are those used to calculate the LBA likelihoods in 5.

G. KERNEL DENSITY ESTIMATION

The method of *kernel density estimation* is used in a number of applications in this thesis, including the calculation of parexic likelihoods and posteriors. The problem of kernel density estimation (KDE) is, given a discrete set of random samples X over a continuous support, to provide an approximate representation of the unknown generating probability density function $f(x)$. The basic notion behind KDE's solution to the problem is to first transform each discrete member of

the sample set to a continuous contribution over the support by convolving the sample with a smoothing function, or *kernel*, then to sum the smoothed contributions to obtain the final density estimate; regions with higher densities of samples will see overlap in the convolved contributions, leading to higher estimated densities. More precisely, kernel density estimation approximates the underlying probability distribution $f(x)$ of a set of N samples $X = \{X_i, i = 1, 2, \ldots, N\}$ as $\widetilde{f}$ according to the following equation:

$$\widetilde{f}(x|X) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right) . \tag{A.10}$$

In this equation, K is the *kernel*, a function which convolves each discrete sample to transform it into a continuous contribution to $\widetilde{f}$, and h is the *bandwidth*, which controls the width of the convolution transformation and therefore the extent to which the kernel density estimate smooths the sample set X in forming $\widetilde{f}$.

Evidently, the choice of kernel and bandwidth significantly affect the final kernel density estimate. The bandwidth especially can produce an over-broad (squashed and spread out) estimate if too high, and an overly ragged one if too low. Unless otherwise noted, when applying KDE in this work, we use a standard Gaussian kernel and a simplified form of the bandwidth as suggested in Janssen et al. (1995), whereby

$$h = 1.06\hat{\sigma}_X N^{-1/5}, \qquad \hat{\sigma}_X = \min\{\sigma_X, R_{0.4}/1.048\} . \tag{A.11}$$

Here, σ_X is the sample standard deviation of X, and R_α is the smallest parameter range enclosing proportion of α of the probability. The value of R_α is found by searching over the $\{p + \alpha, p = 0.01, 0.02, \ldots, 0.6\}$. The value 1.048 is a scaling value, equal to $R_{0.4}$ for the unit normal distribution. Further discussion of kernel density estimates and bandwidth selection can be found in Janssen et al. (1995), Jones et al. (1996), and Lewis (2019).

H. PERMISSIONS FOR REPRODUCED WORKS

Permission for the use of work from Feroz et al. (2013) in figure 3.2 (in a modified form) on page 51 kindly granted by the authors under a Creative Commons Attribution 4.0 International license (CC BY 4.0).

Use of work from Feroz and Hobson (2008) in figures 3.4 and 3.5 on pages 56 and 57 allowed by journal policy (https://academic.oup.com/mnras/pages/rights_and_new_business_development).

Use of work from Feroz et al. (2009) in figures 3.6, 3.7, and 3.8 on pages 58, 58, and 59 allowed by journal policy (https://academic.oup.com/mnras/pages/rights_and_new_business_development).

Permission for the reproduction of work from Kass and Raftery (1995) in Table 2.1 on page 41 kindly granted by Taylor and Francis on 24 June 2021.

Permission for the reproduction of work from Burbeck and Luce (1982) in figure 4.2 on page 71 kindly granted by Springer Nature on 23 June 2021.

Permission for the reproduction of work from Smith (1995) in figure 4.3 on page 72 kindly granted by the American Psychological Association on 23 June 2021.

Permission for the reproduction of work from Smith (1995) in figure 4.3 on page 72 kindly granted by the American Psychological Association on 23 June 2021.

Permission for the use of work from Los et al. (2014) in figure 4.5 (in a slightly modified form) on page 79 kindly granted by the authors under a Creative Commons Attribution 4.0 International license (CC BY 4.0).

—ω—

Bibliography

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 1–7. https://doi.org/10.1111/bmsp.12004

Annis, J., Evans, N. J., Miller, B. J., & Palmeri, T. J. (2019). Thermodynamic integration and steppingstone sampling methods for estimating Bayes factors: A tutorial. *Journal of Mathematical Psychology*, *89*, 67–86. https://doi.org/10.1016/j.jmp.2019.01.005

Annis, J., & Palmeri, T. J. (2018). Bayesian statistical approaches to evaluating cognitive models. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(2). https://doi.org/10.1002/wcs.1458

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception Psychophysics*, *32*(2), 117–133. https://doi.org/10.3758/BF03204271

Chalmers, A. F. (2013). *What is this thing called science?* (4th ed.). Brisbane, University of Queensland Press.

D'Agostini, G. ((2003). *Bayesian reasoning in data analysis : A critical introduction*. New Jersey ; World Scientific.

DeGroot, M. H. (1970). *Optimal statistical decisions*. McGraw-Hill.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin Review*, *16*, 1129–1135.

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A. M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y. S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., … Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin and Review*, *26*(4), 1051–1069. https://doi.org/10.3758/s13423-017-1417-2

Evans, N. J. (2019). Assessing the practical differences between model selection methods in inferences about choice response time tasks. *Psychonomic Bulletin and Review*, *26*(4), 1070–1098. https://doi.org/10.3758/s13423-018-01563-9

Evans, N. J., & Annis, J. (2019). Thermodynamic integration via differential evolution: A method for estimating marginal likelihoods. *Behavior Research Methods*, *51*(2), 930–947. https://doi.org/10.3758/s13428-018-1172-y

Evans, N. J., & Brown, S. D. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, *50*(2), 589–603. https://doi.org/10.3758/s13428-017-0887-5

Evans, N. J., & Wagenmakers, E.-J. (2020). Evidence Accumulation Models: Current Limitations and Future Directions. *The Quantitative Methods for Psychology*, *16*(2), 73–90. https://doi.org/10.20982/tqmp.16.2.p073

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. https://doi.org/10.1017/CBO9781316272503

Feroz, F. (2008). *Bayesian Methods for Astrophysics and Particle Physics* (Doctoral dissertation September). Univeristy of Cambridge.

Feroz, F., Hobson, M. P., & Bridges, M. (2009). MultiNest: An efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, *398*(4), arXiv, 1601–1614. https://doi.org/10.1111/j.1365-2966.2009.14548.x

Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. (2013). Importance Nested Sampling and the MultiNest Algorithm. *The Open Journal of Astrophysics*, *2*(1). https://doi.org/10.21105/astro.1306.2144

Feroz, F., & Hobson, M. P. (2008). Multimodal nested sampling: An efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, *384*(2), arXiv, 449–463. https://doi.org/10.1111/j.1365-2966.2007.12353.x

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? *Trends in Cognitive Science*, *11*(6), 272–279. https://doi.org/10.1016/j.tics.2011.04.002

Friel, N., Hurn, M., & Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, *24*(5), 709–723. https://doi.org/10.1007/s11222-013-9397-1

Friel, N., & Wyse, J. (2012). Estimating the evidence - a review. *Statistica Neerlandica*, *66*(3), 288–308. https://doi.org/10.1111/j.1467-9574.2011.00515.x

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (Third). Boca Raton, FL, Taylor & Francis.

Gordon, I. E. (1967). Stimulus Probability and Simple Reaction Time. *Nature*, (1965), 895–896. https://doi.org/10.1038/215895a0

Graff, P., Feroz, F., Hobson, M. P., & Lasenby, A. (2012). BAMBI: Blind accelerated multimodal Bayesian inference. *Monthly Notices of the Royal Astronomical Society*, *421*(1), arXiv, 169–180. https://doi.org/10.1111/j.1365-2966.2011.20288.x

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E. J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Handley, W. J., Hobson, M. P., & Lasenby, A. N. (2015). PolyChord: Nested sampling for cosmology. *Monthly Notices of the Royal Astronomical Society: Letters*, *450*(1), L61–L65. https://doi.org/10.1093/mnrasl/slv047

Hawthorne, J. (2021). Inductive Logic (E. N. Zalta, Ed.; Spring 2021). In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University.

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods, Instruments, and Computers*, *36*(4), 678–694. https://doi.org/10.3758/BF03206550

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., & Derks, K. (2020). *A Review of Applications of the Bayes Factor in Psychological Research* [Preprint]. Preprint. https://doi.org/10.31234/osf.io/cu43g

Henderson, L. (2020). The Problem of Induction (E. N. Zalta, Ed.; Spring 2020). In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*(4), 382–386. https://doi.org/10.1037/h0021740

Huber, F. (2022, February 21). Confirmation and Induction, In *Internet Encyclopedia of Philosophy*. Retrieved February 21, 2022, from https://iep.utm.edu/conf-ind/

IEP Staff. (2022, February 21). Deductive and Inductive Arguments, In *Internet Encyclopedia of Philosophy*. Retrieved February 21, 2022, from https://iep.utm.edu/ded-ind/

Janssen, P., Marron, J. S., Veraverbeke, N., & Sarle, W. (1995). Scale measures for bandwidth selection. *Journal of Nonparametric Statistics*, *5*(4), 359–380. https://doi.org/10.1080/10485259508832654

Jaynes, E. T. (2003). *Probability theory: The logic of science* (G. L. Bretthorst, Ed.; First). Cambridge, Cambridge University Press.

Jeffreys, H. (1961). *The theory of probability* (Third) [Reprint 2003. First published 1939.]. Oxford, Oxford University Press.

Johnson, R. A. (2011). Statistical Inference. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1418–1420). Berlin, Heidelberg, Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_545

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*(433), 401–407. https://doi.org/10.1080/01621459.1996.10476701

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.2307/2291091

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), arXiv 1807.11819, 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Lambert, B. (2018). *A student's guide to bayesian statistics*. Los Angeles, SAGE.

Lanczos, C. (1956). *Applied analysis*. Englewood Cliffs; N. J., Prentice-Hall.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. New York, NY, US, Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lewis, A. (2019). Getdist: A python package for analysing monte carlo samples.

Liu, P., Elshall, A. S., Ye, M., Beerli, P., Zeng, X., Lu, D., & Tao, Y. (2016). Evaluating marginal likelihood with theromodynamic integration method and comparison with several other numerical methods. *Journal of the American Water Resources Association*, *52*, 734–758. https://doi.org/10.1002/2014WR016718

Logan, G. D., & Cowan, W. B. (1984). On the Ability to Inhibit Thought and Action: A Theory of an Act of Control. *Psychological Review*, *91*(3), 295–327. https://doi.org/10.1007/978-3-642-56546-5_2

Logan, G. D., Schall, J. D., & Palmeri, T. J. (2015). Inhibitory Control in Mind and Brain: The Mathematics and Neurophysiology of the Underlying Computation (B. U. Forstmann & E. J. Wagenmakers, Eds.). In B. U. Forstmann & E. J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience*. Springer. https://doi.org/10.1007/978-1-4939-2236-9

Los, S. A., & Agter, F. (2005). Reweighting sequential effects across different distributions of foreperiods: Segregating elementary contributions to nonspecific preparation. *Perception and Psychophysics*, *67*, 1161–1170. https://doi.org/10.3758/BF03193549

Los, S. A., Knol, D. L., & Boers, R. M. (2001). The foreperiod effect revisited: Conditioning as a basis for nonspecific preparation. *Acta Psychologica*, *106*(1-2), 121–145. https://doi.org/10.1016/S0001-6918(00)00029-9

Los, S. A., Kruijne, W., & Meeter, M. (2014). Outlines of a multiple trace theory of temporal preparation. *Frontiers in Psychology*, *5*(SEP), 1–13. https://doi.org/10.3389/fpsyg.2014.01058

Los, S. A., Kruijne, W., & Meeter, M. (2017). Hazard Versus History : Temporal Preparation Is Driven by Past Experience, *43*(1), 78–88.

Los, S. A., & Van den Heuvel, C. E. (2001). Intentional and unintentional contributions to nonspecific preparation during reaction time foreperiods. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 370–386. https://doi.org/10.1037/0096-1523.27.2.370

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, N. Y., Oxford University Press USA.

Mo, S. S., & George, E. J. (1977). Foreperiod effect on time estimation and simple reaction time. *Acta Psychologica*, *41*, 47–59. https://doi.org/10.1002/1097-4679

Mukherjee, P., Parkinson, D., & Liddle, A. R. (2006). A Nested Sampling Algorithm for Cosmological Model Selection. *The Astrophysical Journal*, *638*(2), arXiv, L51–L54. https://doi.org/10.1086/501068

Näätänen, R. (1971). Non-aging fore-periods and simple reaction time. *Acta Psychologica*, *35*(4), 316–327. https://doi.org/10.1016/0001-6918(71)90040-0

Näätänen, R., & Koskinen, P. (1975). Simple reaction time with very small imperative-stimulus probabilities. *Acta Psychologica*, *39*, 43–50.

Newton, M., & Raftery, A. (1994). Approximate bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society*, *56*, 1–48.

Nickerson, R. S., & Burnham, D. W. (1969). Response times with nonaging foreperiods. *Journal of Experimental Psychology*, *79*(3), 452–457. https://doi.org/10.1037/h0026889

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin*, *89*(1), 133–162. https://doi.org/10.1037/0033-2909.89.1.133

Pártay, L. B., Csányi, G., & Bernstein, N. (2021). Nested sampling for materials. *The European physical journal. B, Condensed matter physics*, *94*(8).

Pérez-Cruz, F. (2008). Kullback-Leibler Divergence Estimation of Continuous Distributions, In *Proceedings of IEEE International Symposium on Information Theory*.

Polzella, D. J., Raey, E. G., & Bower, S. M. (1989). The effects of brief variable foreperiods on simple reaction time. *Bulletin of the Psychonomic Society*, *27*(5), 467–469. https://doi.org/10.3758/BF03334658

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R. (2014). Measuring psychometric functions with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 870–888.

Ratcliff, R. (2015). Modeling one-choice and two-choice driving tasks. *Attention, Perception and Psychophysics*, *77*, 2134–2144.

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4). https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., & Van Dongen, H. P. A. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences*, *108*(27), 11285–11290. https://doi.org/10.1073/pnas.1100483108

Requin, J., & Granjon, M. (1969). The effect of conditional probability of the response signal on the simple reaction time. *Acta Psychologica*, *31*, 129–144. https://doi.org/10.1016/0001-6918(69)90074-2

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, and Computers*, *33*(4), 457–469. https://doi.org/10.3758/BF03195403

Shaw, J. R., Bridges, M., & Hobson, M. (2007). Efficient Bayesian inference for multimodal problems in cosmology. *Monthly Notices of the Royal Astronomical Society*, *378*(4), arXiv 0701867, 1365–1370. https://doi.org/10.1111/j.1365-2966.2007.11871.x

Sivia, D. S. (2012). *Data analysis: A Bayesian tutorial* (Second edition.). Oxford, Oxford University Press.

Skilling, J. (2004). Nested Sampling (R. Fischer, R. Preuss, & U. von Toussaint, Eds.). In R. Fischer, R. Preuss, & U. von Toussaint (Eds.), *AIP Conference Proceedings of the 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, *1*(4), 833–860. https://doi.org/10.1214/06-BA127

Smith, P. L. (1995). Psychophysically Principled Models of Visual Simple Reaction Time. *Psychological Review*, *102*(3), 567–593.

Steinborn, M. B., & Langner, R. (2011). Distraction by irrelevant sound during foreperiods selectively impairs temporal preparation. *Acta Psychologica*, *136*, 405–418.

Steinborn, M. B., & Langner, R. (2012). Arousal modulates temporal preparation under increased time uncertainty: Evidence from higher- order sequential foreperiod effects. *Acta Psychologica*, *139*, 65–76.

Steinborn, M. B., Rolke, B., Bratzke, D., & Ulrich, R. (2008). Sequential effects within a short foreperiod context: Evidence for the conditioning account of temporal preparation. *Acta Psychologica*, *129*, 297–307.

Steinborn, M. B., Rolke, B., Bratzke, D., & Ulrich, R. (2009). Dynamic adjustment of temporal preparation: Shifting warning signal modality attenuates the sequential foreperiod effect. *Acta Psychologica*, *132*, 40–47.

Steingroever, H., Wabersich, D., & Wagenmakers, E. J. (2020). Modeling across-trial variability in the Wald drift rate parameter. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01448-7

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*, 251–260.

Teichner, W. H. (1954). Recent studies of simple reaction time. *Psychological Bulletin*, *51*(2), 128–149. https://doi.org/10.1037/h0060900

Trotta, R., Feroz, F., Hobson, M., & de Austri, R. R. (2013). Recent Advances in Bayesian Inference in Cosmology and Astroparticle Physics Thanks to the MultiNest Algorithm. In J. M.

Hilbe (Ed.), *Astrostatistical Challenges for the New Astronomy* (pp. 107–119). New York, NY, Springer New York. https://doi.org/10.1007/978-1-4614-3508-2_6

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, *128*, 96–115. https://doi.org/10.1016/j.neuroimage.2015.12.030

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384. https://doi.org/10.1037/a0032222

Vallesi, A. (2010). Neuroanatomical substrates of foreperiod effects (A. C. Nobre & J. T. Coull, Eds.). In A. C. Nobre & J. T. Coull (Eds.), *Attention and time*. Oxford, Oxford University Press.

Vallesi, A., Arbula, S., & Bernardis, P. (2014). Functional dissociations in temporal preparation: Evidence from dual-task performance. *Cognition*, *130*, 141–151.

Vallesi, A., Lozano, V. N., & Correa, A. (2013). Dissociating temporal preparation processes as a function of the inter-trial interval duration. *Cognition*, *127*, 22–30.

Vallesi, A., & Shallice, T. (2007). Developmental dissociations of preparation over time: Deconstructing the variable foreperiod phenomena. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1377–1388.

Vallesi, A., Shallice, T., & Walsh, V. (2007). Role of the prefrontal cortex in the foreperiod effect: Tms evidence for dual mechanisms in temporal preparation. *Cerebral Cortex*, *17*, 466–474.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, *7*(3), 424–462. https://doi.org/10.3758/bf03214357

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model Comparison and the Principle of Parsimony (J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels, Eds.). In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology*.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin and Review*, *25*(1), 1–4. https://doi.org/10.3758/s13423-018-1443-8

Whelan, R. (2008). Effective analysis of reaction time data. *Psychological Record*, *58*(3), 475–482. https://doi.org/10.1007/BF03395630

Woodrow, H. (1914). The Measurement of Attention. *Psychological Monographs*, *17*(5), i–158.

Woods, D. L., Herron, T. J., Wyma, J. M., Reed, B., & Yund, E. W. (2015). Factors influencing the latency of simple reaction time. *Frontiers in Human Neuroscience*, *9*(March), 1–12. https://doi.org/10.3389/fnhum.2015.00131

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M. H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, *60*(2), 150–160. https://doi.org/10.1093/sysbio/syq085