

Simple Two-stage Algorithms for Fitting Hierarchies of Complex Models with Applications in Astrophysics

A thesis presented for the degree of
Doctor of Philosophy of Imperial College London
by

Shijing Si

Department of Mathematics
Imperial College London
London SW7 2AZ

JANUARY 15, 2018

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: **Shijing Si**

Copyright

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.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor David A. van Dyk for his incredible patience and support over the last four years.

I am very grateful to Imperial College for funding my studies and providing me with a Roth Doctoral Fellowship. Without the financial support, this work would not have been possible.

Thank you to Professor Ted von Hippel and Dr. David C. Stenning for many helpful discussions and patient explanations, and for the chance to participate in some interesting astrophysical projects.

Thank you to my examiners Professor David Valls-Gabaud and Dr. Roberto Trotta for your helpful suggestions.

Finally, thanks to my parents, for their support and patience over these past few years.

Shijing Si

Abstract

Hierarchical models are one of the most efficient and principled techniques to estimate the parameters of each object in a population. In astrophysics, we often aim to estimate a set of parameters for each of a number of objects in a population, which motivates the application of hierarchical models.

However, fully modelling a population of objects within a hierarchical model requires substantial computational investment and often specialised computer code (black-box), especially for complicated problems. In this study, we develop novel methods to conveniently obtain the improved estimates available under a hierarchical model. While taking advantage of the existing code for case-by-case analyses, our methods simultaneously estimate parameters of the individual objects and parameters that describe their population distribution.

There are many possible applications of hierarchical models in astrophysics. In this research we deploy our novel computational algorithms to three projects: the determination of distance modulus to the Large Magellanic Cloud (LMC), the precise estimates of a group of Galactic halo white dwarfs (WDs), the reliable estimates of the initial-final mass relations (IFMRs). These three circumstances correspond to the applications of Bayesian hierarchical models to more precise determination of a common parameter, a set of parameters of one kind and a set of functions of one kind, respectively.

The main contribution of this work lies in the last two—WDs and IFMRs—projects, which are built upon a few computer models. Our innovation is to fold these complex computer-based astrophysical models into a Bayesian hierarchical model that allows us to simultaneously fit individual models for multiple stars or even multiple clusters of stars to obtain estimators with superior statistical properties (e.g., reduced mean squared errors). Our easy-to-implement computational algorithms make these computationally challenging tasks tractable via taking advantage of case-by-case fits, which are potentially useful to many projects in other areas. Our contribution is mainly on the statistical side and all astronomical projects are meant to be testbeds for proofs of concept of my new algorithms. Our analyses are preliminary in the astronomy side, and all the astronomical results are not definitive.

Table of contents

Acknowledgements	4
Abstract	5
1 Introduction	19
1.1 Hierarchical Models	19
1.1.1 Background	20
1.1.2 Gaussian Linear Hierarchical Model	23
1.1.3 Frequentist Estimation	25
1.1.4 Bayesian Approaches	29
1.2 Astrophysical Background	31
1.2.1 Distance Modulus to the Large Magellanic Cloud	32
1.2.2 Stellar Evolution	33
1.2.3 Hierarchical Modelling of the Age of Galactic Halo White Dwarfs	41
1.2.4 Hierarchical Modelling of the Initial-Final Mass Relations for Star Clusters	42
1.3 Review of Statistical Computational Methods	43
1.3.1 Markov Chain Monte Carlo	43
1.3.2 Expectation-Maximisation algorithm	46
1.3.3 Monte Carlo Expectation-Maximisation Algorithm	51
1.4 Novel Statistical Algorithms Illustrated in Astrophysics	53
2 Statistical Methodology and Computation	59
2.1 Introduction	59
2.2 Fully Bayesian Approach	61
2.3 Empirical Bayes Approach	64
2.4 Prior Distributions	68
2.4.1 General Methods to Create Non-informative Priors	69
2.4.2 Priors for Variance Parameters in Hierarchical Models	71
2.5 Numerical Validation of Gaussian-Gaussian Hierarchical Model	74
2.5.1 Data Simulation	74
2.5.2 Case-by-case Approach	75
2.5.3 Fully Bayesian Approach	76

2.5.4	Two-stage Sampler for Fully Bayesian	77
2.5.5	Empirical Bayes Approach	78
2.5.6	Result of Simulation Study	79
3	Distance Modulus to the Large Magellanic Cloud	91
3.1	Introduction	93
3.2	Hierarchical Analysis	94
3.3	Simulation Study	97
3.4	Data Analysis	103
3.5	Concluding Remarks	111
4	Estimating the Ages of a Group of Halo White Dwarfs	113
4.1	Case-by-case Analysis	114
4.2	Hierarchical Modelling of a Group of White Dwarfs	115
4.2.1	Prior Distributions	116
4.2.2	Empirical Bayes Method	119
4.2.3	Fully Bayesian Method	120
4.3	Simulation Study	122
4.4	Data Analysis	141
4.4.1	Case-by-case Analysis	143
4.4.2	Hierarchical Analysis	145
4.5	Sensitivity Analysis for the Assumption of the Hierarchical Model . .	154
4.6	Concluding Remarks	166
5	Initial-Final Mass Relations	168
5.1	Case-by-case Analysis	169
5.2	Hierarchical Models for a Group of Star Clusters	174
5.3	Two-stage Algorithm for Hierarchical Modelling	177
5.4	Simulation Study	179
5.5	Data Analysis	185
5.5.1	Case-by-case Analysis	186
5.5.2	Hierarchical Analysis	190
5.6	Sensitivity Analysis	208
5.6.1	Sensitivity to Prior Distribution	209
5.6.2	Sensitivity to Membership of WDs in M67	213
5.7	Concluding Remarks	214
6	Conclusion & Future Work	218
6.1	Conclusion	218
6.2	Future Work	219
	Appendix A Case-by-case Fits of WDs	240

Appendix B	Figures and Tables for the IFMR Project	251
B.1	Photometry for WDs in Clusters	251
B.2	Cluster-specific Results	251

List of Figures

1.1	The CMD plot for Hyades. Each dot indicates one star in the Hyades cluster. The units on y-axis are magnitudes on V filter with lower values indicating brighter sources. The units on x-axis are difference of magnitudes between B and V filters. The solid red and blue lines are the fitted MS/RG and WD models. References: DeGennaro et al. (2009); van Dyk et al. (2009); Stein et al. (2013).	35
1.2	Stellar parameters and expected photometry data.	38
1.3	A WD's 3-dimensional scatterplot of its marginal posterior distribution of mass, distance modulus, and $\log_{10}(\text{Age})$. Its shape is complicated, and the parameters have non-linear correlations. The specific values of the parameters are not included, because they are not relevant to the purpose of this figure. This WD, J0822+3903 (Dame et al., 2016), is included in analysis in Chapter 4.	40
1.4	The figure of a Galaxy. Source: http://csep10.phys.utk.edu/OJTA2dev/ojta/course2/milkyway/components/components_tl.html-save	41
2.1	Comparisons of the RMSEs for μ, γ and τ from different approaches of fitting the Gaussian-Gaussian model.	88
2.2	Quantile-quantile plots for the two-stage sampler versus Gibbs sampler of $\mu_i, i = 1, 2, \dots, 10$ in the Gaussian-Gaussian model.	89
2.3	Quantile-quantile plots for the two-stage sampler versus Gibbs sampler of γ, τ in the Gaussian-Gaussian model.	90
3.1	The profile posterior distribution of τ (left panel) and of $\xi = \ln \tau$ (right panel). The modal estimate of τ is zero and is less representative of the distribution than is the modal estimate of ξ	97
3.2	The EM iterations of γ, τ and ξ	98
3.3	The biases, standard deviations, and RMSEs of point estimates of μ_{ij} versus true values of μ_{ij} . Estimates are obtained from three approaches: the case-by-case (black plus signs), FB-Gibbs (blue circles) and EB-log (green triangles).	100

3.4	Posterior density functions of the LMC distance modulus based on hierarchical fitting (Red dashed lines) and based on the case-by-case analysis (Blue solid lines). Those based on hierarchical fitting shrink towards the centre relative to those based on the case-by-case analysis.	104
3.5	The point estimates of distance moduli from the hierarchical (upper line) and case-by-case (individual in the lower line) fittings. Each arrow connects the point estimate from one certain case-by-case analysis with its hierarchical counterpart.	105
3.6	The point estimates and 68.3% credible intervals for the distance moduli to LMC from the EB-log, FB-Gibbs and case-by-case analyses.	106
3.7	Conditional posterior means of distance modulus $E(\mu_i \tau, \mathbf{D})$, as a function of the standard deviation of the systematic errors τ . The red dashed line indicates our estimate $\hat{\tau} = 0.045$. The 13 solid round dots of different colours show the hierarchical estimates of the distance modulus to the LMC using various methods, whereas the plus signs are estimates from the case-by-case analyses. (Plus signs and round dots of the same colour correspond to the same published result; see Table 3.1.)	109
4.1	Results for Simulated Setting I, with $\tau = 0.05$ (standard deviation of $\log_{10}(\text{age})$) and $\Sigma = \Sigma_0$ (photometric error variances). The rows correspond to various fitting methods: EB, EB-log, EB-inv and FB, respectively. Each row includes: i) a histogram of the fitted values of $\hat{\gamma}$, across the $N_2 = 25$ simulation replicates, ii) a histogram of $\hat{\tau}$, and iii) a comparison of the difference in absolute errors in the fitted WD ages obtained with the case-by-case analysis and with each of the hierarchical methods. The thick dashed red lines in each histogram indicate the true values of γ or τ . In the right-most column, positive vertical values indicate larger error for the case-by-case fit, and the thin horizontal lines correspond to equal errors for the two methods.	126
4.2	Similar to Fig. 4.1, but for the case $\tau = 0.03$ and $\Sigma = 0.8^2\Sigma_0$.	127
4.3	Similar to Fig. 4.1, but for the case $\tau = 0.03$ and $\Sigma = 1.2^2\Sigma_0$.	128
4.4	Similar to Fig. 4.1, but for the case $\tau = 0.06$ and $\Sigma = 0.8^2\Sigma_0$.	129
4.5	Similar to Fig. 4.1, but for the case $\tau = 0.06$ and $\Sigma = 1.2^2\Sigma_0$.	130
4.6	Projections of the joint posterior distributions onto the two dimensional planes (from top to bottom) of age–metallicity, age–distance, age–ZAMS mass, and distance–ZAMS mass planes for five of the Galactic halo WDs. The posterior distribution are computed under the case-by-case fits.	144
4.7	Quantile-quantile plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 8.6. The red line represents that quantiles from samples in x and y axes are equal.	146

4.8	Quantile-quantile plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 8.8. The red line represents that quantiles from samples in x and y axes are equal.	147
4.9	Quantile-quantile plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 9.0. The red line represents that quantiles from samples in x and y axes are equal.	148
4.10	Quantile-quantile (QQ) plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 9.2. The red line represents that quantiles from samples in x and y axes are equal.	149
4.11	The comparison of the posterior distributions of ages of WDs from hierarchical analyses and case-by-case study. The left panel uses FB approach to fit hierarchical model, while the right panel uses EB-log.	151
4.12	The EB-log, FB, and case-by-case estimates of the age of the ten halo WDs. The interval estimates are 68.3% credible intervals of the posterior distributions. The dashed, solid, and dotted intervals are computed from FB, case-by-case, and EB-log, respectively. The triangle, circle and cross signs are the posterior means of the ages of WDs from FB, case-by-case, and EB-log, respectively.	152
4.13	Histograms of the Monte Carlo sample for the population mean age and population standard deviation obtained using FB fit to ten candidate galactic halo WDs. Here the population mean of Age (Gyr) = 10^γ and the population standard deviation of Age (Gyr) = $10^{E(\gamma)} \log_e(10)\tau$.	153
4.14	Time series plots of samples of $\log_{10}(\text{Age})$ for 10 WDs from the EB-log approach.	155
4.15	Histograms of samples of $\log_{10}(\text{Age})$ for 10 WDs from the EB-log approach.	156
4.16	Iterates of population-level parameters γ and τ from the EB-log approach.	157
4.17	Time series plots of samples of $\log_{10}(\text{Age})$ for 10 WDs from the FB approach using the Two-stage algorithm.	158
4.18	Histograms of samples of $\log_{10}(\text{Age})$ for 10 WDs from the FB approach using the Two-stage algorithm.	159
4.19	Time series and autocorrelation plots of population-level parameters γ and τ from the FB approach using the Two-stage algorithm.	160
4.20	The scaled gamma distribution used to generate $\log_{10}(\text{Age})$ in our sensitivity analysis. It has a long left tail from 9.9 to 10.06 and peaks at 10.06, then it plunges to 0 at 10.09. In astronomy, this population distribution of $\log_{10}(\text{Age})$ means that more WDs have their $\log_{10}(\text{Age})$ smaller than 10.06 and 10.09 is the upper limit for $\log_{10}(\text{Age})$.	161

4.21	The absolute value of errors from the star-by-star fits minus that of the hierarchical fits under four simulation settings. Each plus sign represents one WD and dashed lines indicate 0. Plus signs above the dashed lines correspond to WDs with ages that are better estimated with the hierarchical fits. For plus signs below the dashed lines, star-by-star fits produce better estimates.	163
5.1	Estimated IFMR parameters for 10 simulated clusters using both the hierarchical (red solid lines) and cluster-specific (blue dashed lines) analyses.	184
5.2	Cluster-specific results: projections of the joint posterior distributions onto the two dimensional planes of (from left to right) age–metallicity, age–distance, age–absorption, and IFMR intercept–IFMR slope for five clusters.	189
5.3	Colour magnitude diagrams for the Hyades based on both the hierarchical and cluster-specific estimates. The solid (red) lines are from the hierarchical fits and dashed (blue) lines are from the cluster-specific fits.	194
5.4	Colour magnitude diagrams for (from top to bottom) M67, NGC 188, NGC 2168, NGC 2477, based on both the hierarchical and cluster-specific estimates. The solid (red) lines are from the hierarchical fits and dashed (blue) lines are from the cluster-specific fits. The solid points in the left panel of M67 are inferred as members of the cluster with BASE-9, while the hollow points in the CMD of M67 are inferred as non-members.	195
5.5	The average IFMR for five clusters obtained from both the hierarchical and cluster-specific analyses. The ovals are 68.3% contours of the joint distribution of ZAMS and WD masses for WDs in each cluster. The colour scheme of red, blue, green, purple, and black corresponds to the Hyades, M67, NGC 188, NGC 2168, and NGC 2477, respectively. The colour scheme of Williams, Salaris and Salaris piecewise are in solid, dashed and dotted grey lines.	196
5.6	The IFMR estimates and joint distributions of the initial and final masses for WDs in each cluster from both the cluster-specific (case-by-case) and hierarchical analysis. The lines are IFMR estimates and the ovals are 68.3% contours of the joint distribution of ZAMS and WD masses for WDs in each cluster. The dashed lines and ovals are based on cluster-specific (case-by-case) fits, while their solid counterparts are based on the hierarchical analysis.	197
5.7	The time series plots of IFMR parameters from the case-by-case (cluster-specific) analysis.	199
5.8	Histograms of IFMR parameters from the case-by-case (cluster-specific) analysis.	200

5.9	The time series plots of IFMR parameters from the hierarchical analysis.	201
5.10	Histograms of IFMR parameters from the hierarchical analysis. . . .	202
5.11	Time series plots for population-level parameters in the IFMR hierarchical model.	203
5.12	Histograms for population-level parameters in the IFMR hierarchical model.	204
5.13	Quantile-quantile plots of MCMC samples of IFMR parameters of five clusters. The x-axis and y-axis are quantiles corresponding to fits with $\Lambda = 10^3$ and $\Lambda = 10^4$, respectively. The red line represents that quantiles from samples in x and y axes are equal.	210
5.14	Quantile-quantile plots of MCMC samples of IFMR parameters of five clusters. The x-axis and y-axis are quantiles corresponding to fits with $\Lambda = 10^3$ and $\Lambda = 10^5$, respectively. The red line represents that quantiles from samples in x and y axes are equal.	211
5.15	Quantile-quantile plots of MCMC samples of IFMR parameters of five clusters. The x-axis and y-axis are quantiles corresponding to fits with $\Lambda = 10^3$ and $\Lambda = 10^6$, respectively. The red line represents that quantiles from samples in x and y axes are equal.	212
5.16	Colour-magnitude diagram (CMD) plots of three fits. The black lines are from the original fit, red lines from the fit that includes all WDs in M67, blue lines from the fit that removes nine non-members. In the left panel, hollow points represent non-member WDs from the original fit and solid points are members. Observed error bars are added to the WDs.	215
A.1	The scatterplots of parameters of WD J0301-0044 from the case-by-case fit.	241
A.2	The scatterplots of parameters of WD J0346+246 from the case-by-case fit.	242
A.3	The scatterplots of parameters of WD J0822+3903 from the case-by-case fit.	243
A.4	The scatterplots of parameters of WD J1024+4920 from the case-by-case fit.	244
A.5	The scatterplots of parameters of WD J1102+4113 from the case-by-case fit.	245
A.6	The scatterplots of parameters of WD J1107+4855 from the case-by-case fit.	246
A.7	The scatterplots of parameters of WD J1205+5502 from the case-by-case fit.	247
A.8	The scatterplots of parameters of WD J2137-1050 from the case-by-case fit.	248

A.9	The scatterplots of parameters of WD J2145+1106N from the case-by-case fit.	249
A.10	The scatterplots of parameters of WD J2145+1106S from the case-by-case fit.	250
B.1	The scatterplots of Hyades parameters from the case-by-case fit. . . .	254
B.2	The scatterplots of M67 parameters from the case-by-case fit.	255
B.3	The scatterplots of NGC 188 parameters from the case-by-case fit. . . .	256
B.4	The scatterplots of NGC 2168 parameters from the case-by-case fit. . . .	257
B.5	The scatterplots of NGC 2477 parameters from the case-by-case fit. . . .	258

List of Tables

2.1	Parameters in the hierarchical model	60
2.2	Notations in the simulation study	75
2.3	The RMSE of the various estimates of $\boldsymbol{\mu}$, γ and τ under numerous simulated settings. For $\boldsymbol{\mu}$, its RMSE is the root of average MSE of $\mu_1, \dots, \mu_I$	82
2.4	The average coverage probabilities (CP) and their 68.3% confidence intervals (CI) of various credible intervals (with nominal coverage probability 68.3%) of $\boldsymbol{\mu}$ (averaged over $\mu_1, \dots, \mu_I$) under numerous simulated settings.	84
2.5	The actual Coverage Probability (CP) and their 68.3% Confidence Interval (CI) of FB estimates of population-level parameters γ and τ under numerous simulated settings.	85
2.6	CPU Times of different methods in this simulation study	86
3.1	Published Distance Indicators	94
3.2	Parameters in the hierarchical model	95
3.3	The average bias, standard deviation (SD) and RMSE of the various estimates of $\boldsymbol{\mu}$, γ and τ . Each figure has a multiplication factor 10^{-2} and we omit it to make the numbers more legible.	101
3.4	The actual coverage probabilities (CP) and their 68.3% confidence intervals (CI) of the various credible interval estimates of $\boldsymbol{\mu}$, γ and τ . The nominal coverage probability is 68.3%. We obtain these results by pooling estimates from all simulated datasets. For $\boldsymbol{\mu}$, its coverage probability is the average over $\mu_1, \dots, 13$	102
3.5	Point and 68.3% credible interval estimates of the object-level parameter $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{13})$ for 13 included studies from the EB-log, FB-Gibbs and case-by-case analyses.	107
3.6	Estimates for population-level parameters: the true distance modulus (γ) and standard deviation (τ) of systematic errors from 13 different studies.	108
4.1	Parameters in the hierarchical model	117
4.2	Prior distributions for both object-level and population-level parameters	118

4.3	Distributions used in the simulation study.	123
4.4	The percentage of WDs for which EB and FB improved age estimates over case-by-case fits among 250 simulated WDs under the five simulated settings. In other words, these actual figures are the percentages of plus signs above the red lines in Figs 4.1–4.5.	133
4.5	The MAE and RMSE of the various shrinkage and case-by-case estimates of ages ($\log_{10}$ scale) of WDs under five simulated settings. . . .	134
4.6	The RMSE of the various hierarchical estimates of population-level parameters under five simulated settings.	136
4.7	The coverage probabilities (CP) of 68.3% credible intervals of the estimates of ages ($\log_{10}$ scale) of WDs and underneath are their 68.3% confidence intervals (CI).	138
4.8	FB coverage probabilities (CP) of 68.3% credible intervals for estimates of population-level parameters and their 68.3% confidence intervals (CI).	139
4.9	CPU Times of different methods	140
4.10	Prior distributions for distance moduli and atmosphere composition for halo WD sample	142
4.11	Posterior means and 68.3% credible intervals of the ages (Gyr) of ten candidate halo WDs.	150
4.12	68.3% intervals for the age distribution, mean age and standard deviation of Galactic halo WDs	151
4.13	Summary of the relative performance of hierarchical and star-by-star fits in each of four simulated settings.	164
4.14	Coverage probabilities (CP) and their 68.3% confidence intervals (CI) of interval estimates for ages of WDs in each of four simulated settings. The nominal level is 68.3%.	164
5.1	Cluster and stellar parameters.	171
5.2	Categories of parameters in the hierarchical model	177
5.3	Parameter distributions for Simulated Clusters	179
5.4	RMSE and coverage probabilities (CP) of 68.3% credible intervals (CI) for IFMR parameters from FB and case-by-case analyses.	180
5.5	RMSE and coverage probabilities (CP) of 68.3% credible intervals (CI) for population-level parameters from FB.	181
5.6	Point Estimates of IFMR parameters for 10 simulated clusters from both the cluster-specific and hierarchical fits	183
5.7	CPU Times of two methods in the IFMR project	185
5.8	Prior distributions, maximum V magnitudes and their references for five clusters	187
5.9	Estimates of parameters for five clusters from both the Hierarchical and Cluster-specific fits	191
5.10	Results from both the Hierarchical and Cluster-specific fits for WDs in Hyades	205

5.11	Results from the Hierarchical and Cluster-specific fits on WDs in M67	206
5.12	Results from the Hierarchical and Cluster-specific fits for WDs in NGC188, NGC2168 and NGC2477	207
5.13	Estimates of the average IFMR parameters with error bars from the hierarchical model and comparisons with other results	208
5.14	Estimates with error bars of cluster parameters (age, distance modulus, metallicity, absorption, IFMR constant and slope) for M67 under three cases.	214
B.1	Photometry for WDs in the Hyades.	251
B.2	Photometry for WDs in the M67.	252
B.3	Photometry for WDs in the NGC188, NGC2168 and NGC2477.	253

1 Introduction

This chapter covers the statistical techniques and astrophysical background in this research. Suppose we observe a sample of objects from a population of astronomical sources, for example, the photometry of 10 white dwarfs (WDs) from the Galactic halo, and we wish to estimate a particular parameter or a set of parameters for each object. We refer to these as the *object-level parameters*. By virtue of the population, there is a distribution of these parameters across the population of objects. This distribution can be described by another set of parameters that we refer to as the *population-level parameters*. Often we aim to estimate both the object-level parameters and population-level parameters. As we shall see, however, even if we are only interested in the object-level parameters, they can be better estimated if we also consider their population distribution.

In this thesis, we develop efficient computational algorithms to infer both the object-level and population-level parameters. We review statistical techniques and describe the astrophysical motivations in the remainder of this chapter.

1.1 Hierarchical Models

Hierarchical models, sometimes called multilevel models or mixed effects models, generalise ordinary regression methods, in which regression coefficients are themselves given a model. They can combine different units of data together in a study, leading

to a comprehensive understanding of the problem (Gelman, 2006a). They have been widely used in many fields, like medicine, biology, sociology, physics and so on. In medical research, there are a large number of clinical trials of a new drug on different people and it is important to combine all evidence to reach a conclusion. In biology, biologists have many sets of experiments performed under similar conditions. In a vote survey, we can have the poll from plenty of places and we want to merger them in a reasonable way to yield a good estimate. These kinds of problems are so common in both research and practice.

1.1.1 Background

Hierarchical modelling has numerous benefits, outperforming classical regression in predictive accuracy (Gelman, 2006a). Firstly, it uses all available units of data when inferring parameters for each particular unit (Efron, 1996), so it produces an estimation that borrows strength from all available datasets. Another advantage is that hierarchical fittings usually lead to estimators of reduced mean squared error (MSE) (e.g. Efron and Morris, 1973; Gelman, 2006a) compared with non-hierarchical ones. See Gelman (2006a), and Cressie et al. (2009) for more details about the strengths and limitations of hierarchical modelling.

Hierarchical modelling produces shrinkage estimates (Morris and Lysy, 2012), which shrink inward and have smaller MSE on average. Now we examine some basics about shrinkage estimators. James and Stein (1961) stunned the statistical world by showing that for the normal model

$$X_i \sim N(\mu_i, \sigma^2), i = 1, \dots, I, \text{ with } I > 3, \quad (1.1)$$

the maximum likelihood estimates (MLE) $\hat{\mu}_i^{ML} = X_i$ are dominated by $\hat{\mu}_i^{JS} = (1 -$

$B)X_i$ with $B = \frac{(I-2)\sigma^2}{\sum_{i=1}^I X_i^2}$ in terms of MSE (James and Stein, 1961; Morris and Lysy, 2012). As a matter of fact, the MSE of MLE is

$$E\left[\sum_{i=1}^I (\hat{\mu}_i^{ML} - \mu_i)^2 \middle| \mu_1, \dots, \mu_I\right] = I \cdot \sigma^2.$$

while for James-Stein estimator $\hat{\mu}_i^{JS}$,

$$\begin{aligned} & E\left[\sum_{i=1}^I (\hat{\mu}_i^{JS} - \mu_i)^2 \middle| \mu_1, \dots, \mu_I\right] \\ &= \sum_{i=1}^I E\left[(X_i - \mu_i)^2 + B^2 X_i^2 - 2BX_i(X_i - \mu_i)\right] \\ &= I\sigma^2 + (I-2)\sigma^2 E(B) - 2\sigma^2 E\sum_{i=1}^I \frac{\partial(BX_i)}{\partial X_i} \\ &= I\sigma^2 - E\frac{(I-2)^2\sigma^4}{\sum_{i=1}^I (X_i^2)} < I\sigma^2. \end{aligned} \tag{1.2}$$

$$\tag{1.3}$$

In Bayesian decision theory, we have sets $\Theta, \mathcal{X}, \mathcal{A}$, where Θ are the states of nature, $\mathcal{X}$ the possible observations and $\mathcal{A}$ the set of actions can be taken. An observation $\mathbf{x} \in \mathcal{X}$ is distributed as $F(\mathbf{x}|\theta)$ and therefore provides evidence about the state of nature $\theta \in \Theta$. A decision rule is a function $\delta : \mathcal{X} \rightarrow \mathcal{A}$, where for an observed data $\mathbf{x}$, we take the action $\delta(\mathbf{x}) \in \mathcal{A}$. We define a loss function $L : \mathcal{A} \times \Theta \rightarrow \mathbb{R}$ ($\mathbb{R}$ is the set of real number), which specifies the loss produced by taking action $a \in \mathcal{A}$ when the true state of nature is θ . Usually, we take the action $\delta(\mathbf{x})$ after observing data $\mathbf{x}$, so that the loss is $L(\delta(\mathbf{x}), \theta)$. The risk of rule δ is defined as

$$R(\delta, \theta) = E_{F(\mathbf{x}|\theta)}[L(\delta(\mathbf{x}), \theta)].$$

Suppose $\hat{\theta}$ and $\tilde{\theta}$ are two estimators (rules) of θ and they are functions of $\mathbf{x}$, then if $R(\hat{\theta}, \theta) \leq R(\tilde{\theta}, \theta)$ for all θ and $R(\hat{\theta}, \theta) < R(\tilde{\theta}, \theta)$ for some θ , we call $\tilde{\theta}$ is inadmissible.

Therefore, from Eq. (1.3), we can conclude that James-Stein method makes MLE inadmissible when $I \geq 3$.

Hierarchical modelling is closely related to James-Stein estimators (e.g. Efron and Morris, 1973; Efron and Morris, 1977). James-Stein estimation coincides with the posterior mean of μ_i derived from the empirical Bayes approach to fitting the hierarchical model in the following Equation (1.4)

$$\begin{aligned} X_i &\sim N(\mu_i, \sigma^2), i = 1, \dots, I, \text{ with } I > 3, \\ \mu_i &\sim N(0, \tau^2), \end{aligned} \tag{1.4}$$

where τ is the population standard deviation of μ_i s. As a matter of fact, given τ and X_i in the normal-normal model in (1.4), we derive the conditional mean of μ_i , i.e.,

$$E(\mu_i | X_i, \tau^2) = \frac{\frac{X_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}} = (1 - H)X_i, \tag{1.5}$$

with $H = \frac{\sigma^2}{\tau^2 + \sigma^2}$. Based on the model in Eq. (1.4), we have $X_i \sim N(0, \sigma^2 + \tau^2)$, so $\sum_{i=1}^I X_i^2 / (\sigma^2 + \tau^2) \sim \chi^2(I)$. By doing some calculation, $E\left(\frac{1}{\sum_{i=1}^I X_i^2 / (\sigma^2 + \tau^2)}\right) = \frac{1}{I-2}$, so we have

$$E\left(\frac{(I-2)\sigma^2}{\sum_{i=1}^I X_i^2}\right) = \frac{\sigma^2}{\sigma^2 + \tau^2}.$$

Consequently, we obtain an unbiased estimator of H , $\hat{H} = \frac{(I-2)\sigma^2}{\sum_{i=1}^I X_i^2}$, and replacing H in Eq. (1.5) with $\hat{H}$ yields the James-Stein estimators $\hat{\mu}_i^{\text{JS}}$. There are a few ways to derive similar estimators, such as finding the maximum likelihood estimate (MLE) or maximum a posterior (MAP) estimate of τ^2 and then estimating μ_i with its conditional mean in Eq. (1.5). We discuss empirical Bayes approach (Efron and Morris, 1972, 1973; Morris, 1983; Efron, 1996) in detail later.

1.1.2 Gaussian Linear Hierarchical Model

In this part, we review a commonly used hierarchical model — the Gaussian linear hierarchical model, also called Gaussian linear mixed effects model. Gaussian linear mixed effects model has the following structure

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_i); \quad (1.6)$$

$$\boldsymbol{\gamma}_i \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{D}), i = 1, 2, \dots, m, \quad (1.7)$$

where $\mathbf{Y}_i$ is a $N_i \times 1$ vector of observations, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are matrices of explanatory variables with dimensions $N_i \times p$ and $N_i \times q$ respectively; $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters which are called fixed effects; $\boldsymbol{\gamma}_i$ is a $q \times 1$ vector of unknown random effects and $\boldsymbol{\epsilon}_i$ is a $N_i \times 1$ vector of random errors with mean $\mathbf{0}$ and known covariance matrix $\boldsymbol{\Sigma}_i$. One main difference between fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{\gamma}_i$ lies in that $\boldsymbol{\beta}$ is the same for all $\mathbf{Y}_i$, but $\boldsymbol{\gamma}_i, i = 1, 2, \dots, m$ are from the same distribution. The total number of observations is $\sum_{i=1}^m N_i = N$. Generally, we assume

$$\text{Cov}(\boldsymbol{\gamma}_i, \boldsymbol{\epsilon}_j) = \mathbf{0} \quad \text{for all } i, j;$$

Model (1.6)-(1.7) can be expressed in matrix notations:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad (1.8)$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{bmatrix}_{N \times 1}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix}_{N \times p}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_m \end{bmatrix}_{N \times m \cdot q},$$

with

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix}_{m \cdot q \times 1} \sim N(\mathbf{0}_{m \cdot q \times 1}, \mathbf{I}_m \otimes \mathbf{D}), \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix}_{N \times 1} \sim N(\mathbf{0}_{N \times 1}, \boldsymbol{\Sigma}),$$

in which $\mathbf{I}_m$ is $m \times m$ identity matrix and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_m \end{bmatrix}_{N \times N}.$$

We derive the marginal distribution of $\mathbf{Y}$ via marginalising the random effects $\boldsymbol{\gamma}$

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\mathbf{D})), \tag{1.9}$$

in which $\mathbf{V}(\mathbf{D}) = \mathbf{Z}(\mathbf{I}_m \otimes \mathbf{D})\mathbf{Z}^T + \boldsymbol{\Sigma}$. In the following subsection, we will review some existing approaches to fitting parameters $\boldsymbol{\beta}$ and $\mathbf{D}$. For detailed discussions of hierarchical models, see Gelman and Hill (2007).

In the following sections, we cover a few approaches for fitting the Gaussian linear

mixed effects models. Broadly speaking, they can be categorised as frequentist and Bayesian methods.

1.1.3 Frequentist Estimation

In this section, we cover two frequentist methods for the Gaussian hierarchical model: maximum likelihood estimation and restricted maximum likelihood estimation.

Maximum Likelihood Estimation and Its Implementation

Maximum Likelihood Estimation (MLE) is one common approach to a wide range of statistical models. For the linear mixed effects model in (1.6)–(1.7), the likelihood for parameters $\boldsymbol{\beta}$ and $\mathbf{D}$ is

$$\begin{aligned} L(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{D}) &= \int \prod_{i=1}^m p(\mathbf{Y}_i|\boldsymbol{\gamma}_i, \boldsymbol{\beta}) p(\boldsymbol{\gamma}_i, \mathbf{D}) d\boldsymbol{\gamma}_1 \dots d\boldsymbol{\gamma}_m \\ &= \prod_{i=1}^m \frac{1}{\det(2\pi(\boldsymbol{\Sigma}_i + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T))} \exp \left\{ -\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta})^T \times \right. \\ &\quad \left. (\boldsymbol{\Sigma}_i + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i)^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) \right\} \end{aligned} \quad (1.10)$$

The MLE of $\boldsymbol{\beta}$ and $\mathbf{D}$ is the point estimate such that

$$(\hat{\boldsymbol{\beta}}_{ML}, \hat{\mathbf{D}}_{ML}) = \arg \max_{\boldsymbol{\beta}, \mathbf{D}} L(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{D}).$$

We show how to find the MLE for $\boldsymbol{\beta}$ and $\mathbf{D}$ via the expectation-maximisation (EM) algorithm. In the following derivations of different estimations, we might omit some constants which do not affect the computation. We treat random effects $\boldsymbol{\gamma}_i, i =$

1, 2, $\dots$, m as missing data, so the log-likelihood for parameters $\boldsymbol{\beta}$ and $\boldsymbol{D}$ is

$$\begin{aligned} L(\boldsymbol{Y}, \boldsymbol{\gamma} | \boldsymbol{\beta}, \boldsymbol{D}) = & -\frac{1}{2} \sum_{i=1}^m (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{Z}_i \boldsymbol{\gamma}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{Z}_i \boldsymbol{\gamma}_i) \\ & - \frac{m}{2} \ln \det(\boldsymbol{D}) - \frac{1}{2} \sum_{i=1}^m \boldsymbol{\gamma}_i^T \boldsymbol{D}^{-1} \boldsymbol{\gamma}_i. \end{aligned} \quad (1.11)$$

Given the data $\boldsymbol{Y}$ and t -th iteration of $\boldsymbol{\beta}$ and $\boldsymbol{D}$, we can derive the conditional density

$$\begin{aligned} p(\boldsymbol{\gamma}_i | \boldsymbol{Y}_i, \boldsymbol{\beta}^{(t)}, \boldsymbol{D}^{(t)}) & \propto p(\boldsymbol{Y}_i | \boldsymbol{\gamma}_i, \boldsymbol{\beta}^{(t)}) \times p(\boldsymbol{\gamma}_i | \boldsymbol{D}^{(t)}), \\ \boldsymbol{\mu}_i & = \left[(\boldsymbol{D}^{(t)})^{-1} + \boldsymbol{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Z}_i \right]^{-1} \boldsymbol{Z}_i \boldsymbol{\Sigma}_i (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}^{(t)}), \end{aligned} \quad (1.12)$$

$$\boldsymbol{\Phi}_i = \left[(\boldsymbol{D}^{(t)})^{-1} + \boldsymbol{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Z}_i \right]^{-1}, \quad (1.13)$$

$$\boldsymbol{\gamma}_i | \boldsymbol{Y}_i, \boldsymbol{\beta}^{(t)}, \boldsymbol{D}^{(t)} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Phi}_i). \quad (1.14)$$

From the EM algorithm, we update parameters

$$\boldsymbol{\beta}^{(t+1)} = \left(\sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{Z}_i \boldsymbol{\mu}_i), \quad (1.15)$$

$$\boldsymbol{D}^{(t+1)} = \frac{1}{m} \sum_{i=1}^m (\boldsymbol{\Phi}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T). \quad (1.16)$$

Upon the convergence of the EM algorithm, we obtain the point estimates of $\boldsymbol{\beta}$ and $\boldsymbol{D}$. Then based on the likelihood function in Eq. (1.10) we can compute the inverse information to obtain the variance of estimates of $\boldsymbol{\beta}$ and $\boldsymbol{D}$.

Restricted Maximum Likelihood (REML)

Restricted maximum likelihood (REML), first developed by Patterson and Thompson (1971), can be divided into two steps. In the first step, it generally assigns a flat prior

on $\boldsymbol{\beta}$, namely $p(\boldsymbol{\beta}) \propto 1$, and integrates $\boldsymbol{\beta}$ out. So the marginal likelihood for $\mathbf{D}$ is

$$L(\mathbf{Y}|\mathbf{D}) = \int L(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{D})p(\boldsymbol{\beta})d\boldsymbol{\beta}, \quad (1.17)$$

REML estimates $\mathbf{D}$ by maximizing $L(\mathbf{Y}|\mathbf{D})$ with respect to $\mathbf{D}$, namely

$$\hat{\mathbf{D}}_{REML} = \arg \max_{\mathbf{D}} L(\mathbf{Y}|\mathbf{D}).$$

Then we estimate $\boldsymbol{\beta}$ by maximizing

$$\hat{\boldsymbol{\beta}}_{REML} = \arg \max_{\boldsymbol{\beta}} L(\mathbf{Y}|\boldsymbol{\beta}, \hat{\mathbf{D}}_{REML}).$$

We show how to use Monte Carlo Expectation-Maximization (MCEM) algorithm to obtain the REML estimates.

Step 0 Initialization. Set initial value for $\mathbf{D}^{(0)} = \mathbf{D}^{init}$.

Step 1 Given the t -th iteration $\mathbf{D}^{(t)}$, we have the conditional densities:

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{D}^{(t)}, \mathbf{Y}, \boldsymbol{\gamma} &\sim N\left(\left(\sum_{i=1}^m \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^m \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{Z}_i \boldsymbol{\gamma}_i), \right. \\ &\quad \left. \left(\sum_{i=1}^m \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1}\right), \\ \boldsymbol{\gamma}_i|\mathbf{D}^{(t)}, \mathbf{Y}, \boldsymbol{\beta} &\sim N\left(\left((\mathbf{D}^{(t)})^{-1} + \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Z}_i\right)^{-1} \mathbf{Z}_i \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}), \right. \\ &\quad \left. \left((\mathbf{D}^{(t)})^{-1} + \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Z}_i\right)^{-1}\right). \end{aligned}$$

From these conditional densities, we can use a Gibbs sampler to simulate samples of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and denote them as $\boldsymbol{\beta}^{[1]}, \dots, \boldsymbol{\beta}^{[S]}$ and $\boldsymbol{\gamma}_i^{[s]}, i = 1, 2, \dots, m, s = 1, 2, \dots, S$, where S is the sample size.

Step 2 From the MCEM algorithm, we update

$$\mathbf{D}^{(t+1)} = \frac{1}{m \times S} \sum_{i=1}^m \sum_{s=1}^S \left(\gamma_i^{[s]} \gamma_i^{[s]T} \right).$$

Step 3 Repeat **Step 1** and **2** until the series $\{\mathbf{D}^{(t)}, t = 1, 2, \dots, \}$ converges.

We can obtain $\hat{\mathbf{D}}^{REML}$ from the MCEM algorithm, then we plug it in the marginal distribution of data in the Equation (1.9), so we have

$$\mathbf{Y}_i \sim N\left(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{Z}_i^T \hat{\mathbf{D}}^{REML} \mathbf{Z}_i + \boldsymbol{\Sigma}_i\right). \quad (1.18)$$

By maximizing the conditional likelihood in (1.18), we have the REML of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}^{REML} = \left(\sum_{i=1}^m \mathbf{X}_i^T (\boldsymbol{\Sigma}_i + \mathbf{Z}_i \hat{\mathbf{D}}^{REML} \mathbf{Z}_i^T)^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^m \mathbf{X}_i^T (\boldsymbol{\Sigma}_i + \mathbf{Z}_i \hat{\mathbf{D}}^{REML} \mathbf{Z}_i^T)^{-1} \mathbf{Y}_i.$$

The difference between MLE and REML lies in that the REML estimates $\mathbf{D}$ from the marginal likelihood function in Equation (1.17), but MLE estimate $\boldsymbol{\beta}$ and $\mathbf{D}$ simultaneously from the likelihood function in Equation (1.10). Although MLE is asymptotically unbiased, it is slightly biased in many cases. REML accounts for the degrees of freedom lost by estimating the fixed effects, and makes a less biased estimation of random effects variances. The estimates of $\mathbf{D}$ are invariant to the value of $\boldsymbol{\beta}$ and less sensitive to outliers in the data compared to ML estimates. To show the superiority of REML over ML, one basic example is the estimation of variance based on an independent and identically distributed random normal sample with unknown mean, namely, $Y_i \sim N(\mu, \sigma^2), i = 1, 2, \dots, n$. We can find the MLE: $\hat{\mu}_{ML} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ and $\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$. The REML estimates are: $\hat{\mu}_{REML} = \bar{Y}$ and $\hat{\sigma}_{REML}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. Clearly, $\hat{\sigma}_{REML}^2$ is unbiased and $\hat{\sigma}_{ML}^2$ is

asymptotically unbiased.

1.1.4 Bayesian Approaches

In this section we describe two Bayesian approaches: Fully Bayesian and Empirical Bayes.

Fully Bayesian Approach

The Fully Bayesian (FB) approach treats all parameters as random variables, obtaining the joint posterior distribution for them. To fit the hierarchical model (1.6)-(1.7), it requires a prior density $p(\boldsymbol{\beta}, \mathbf{D}) = p(\boldsymbol{\beta}|\mathbf{D})p(\mathbf{D})$. The standard prior distributions are $p(\boldsymbol{\beta}|\mathbf{D}) \propto 1$ and $p(\mathbf{D}) \sim \text{Inverse Wishart}(\boldsymbol{\Sigma}_0, n_0)$ ¹. So, the joint posterior density

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{D}|\mathbf{Y}) \propto p(\boldsymbol{\beta}, \mathbf{D})p(\boldsymbol{\gamma}|\boldsymbol{\beta}, \mathbf{D})p(\mathbf{Y}|\boldsymbol{\gamma}). \quad (1.19)$$

¹ The Inverse Wishart distribution is the conjugate prior for the covariance matrix of a multivariate normal distribution, having two parameters: scale matrix $\boldsymbol{\Psi}$ and degrees of freedom ν , and its probability density function is

$$p(\mathbf{X}|\boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{p\nu}{2}}\Gamma_p(\frac{\nu}{2})} |\mathbf{X}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}tr(\boldsymbol{\Psi}\mathbf{X}^{-1})},$$

in which Γ_p is the multivariate gamma function, and $tr(\cdot)$ is the trace function.

We can derive the conditional posterior densities for the linear mixed effects model,

$$\gamma_i | \mathbf{Y}, \mathbf{D}, \boldsymbol{\beta} \sim N \left(\left(\mathbf{D}^{-1} + \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Z}_i \right)^{-1} \mathbf{Z}_i \boldsymbol{\Sigma}_i (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}), \right. \quad (1.20)$$

$$\left. \left(\mathbf{D}^{-1} + \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Z}_i \right)^{-1} \right), \quad (1.21)$$

$$\boldsymbol{\beta} | \boldsymbol{\gamma} \sim N \left(\left(\sum_{i=1}^m \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^m \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{Z}_i \boldsymbol{\gamma}_i), \right. \quad (1.22)$$

$$\left. \left(\sum_{i=1}^m \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \right), \quad (1.23)$$

$$\mathbf{D} | \boldsymbol{\gamma} \sim \text{Inverse Wishart} \left(n_0 + m, \boldsymbol{\Sigma}_0 + \sum_{i=1}^m \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T \right). \quad (1.24)$$

A Gibbs sampler can be employed to draw samples from Equations (1.21)–(1.24). From these samples, we can obtain the mean, variance, credible interval for each parameter.

Empirical Bayes Approach

empirical Bayes (EB) is another type of important approaches to inferring hierarchical models (Efron and Morris, 1972, 1973; Morris, 1983; Efron, 1996; Carlin and Louis, 2000). The basic idea is to find a reasonable point estimate for a subset of parameters first and then derive the posterior density for other parameters conditional on the point estimate. In the first step, namely optimizing some parameters, commonly used ways are MLE, MAP and so on. For the linear mixed model (1.6)–(1.7), EB proceeds as follows.

Step 1. As in the MLE, we optimize parameters $\boldsymbol{\beta}$ and $\mathbf{D}$,

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{D}}) = \arg \max_{\boldsymbol{\beta}, \boldsymbol{\theta}} \int L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma} | \mathbf{Y}) d\boldsymbol{\gamma};$$

Step 2. Given $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{D}}$, find the conditional posterior distribution of $\boldsymbol{\gamma}$ by Bayes theorem, i.e,

$$\begin{aligned} p(\boldsymbol{\gamma}_i | \mathbf{Y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{D}}) &\propto p(\mathbf{Y}_i | \boldsymbol{\gamma}_i, \hat{\boldsymbol{\beta}}) p(\boldsymbol{\gamma}_i | \hat{\boldsymbol{D}}) \\ &\sim N\left(\left(\hat{\boldsymbol{D}}^{-1} + \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Z}_i\right)^{-1} \mathbf{Z}_i \boldsymbol{\Sigma}_i (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}), \left(\hat{\boldsymbol{D}}^{-1} + \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Z}_i\right)^{-1}\right) \end{aligned}$$

The main difference between the FB and EB methods is that FB takes the uncertainty of $\boldsymbol{\beta}$ and $\boldsymbol{D}$ into consideration, but EB does not (Kass and Steffey, 1989). EB is an approximation to the FB. Numerous studies have been done to compare the performances of these two approaches in different projects (Bernardinelli and Montomoli, 1992; Cui and George, 2008; Scott et al., 2010; Persaud et al., 2010).

1.2 Astrophysical Background

This section describes the astrophysical background for the projects in my research. We introduce the Large Magellanic Cloud (LMC), and stellar evolution.

Now we cover a few astronomy terminologies. Astronomy is a natural science that studies celestial objects and phenomena. It applies mathematics, physics, and chemistry, and attempts to explain the origin of those objects and phenomena and their evolution.

Astronomers can determine the luminosity of astronomical objects via telescopes. Luminosity is the total amount of energy emitted by a star, galaxy, or other astronomical object per unit time. In this research, the data of stars or star clusters are photometric magnitudes, a negative logarithmic transformation of luminosities, measured in various specific wavelengths or passbands, such as, the ultraviolet (U), blue (B) and visual (V) filters. The letters U, B, and V stand for ultraviolet, blue, and visual magnitudes.

Astronomers use two different definitions of magnitude: apparent magnitude and absolute magnitude. The apparent magnitude is the brightness of an object as it appears in the night sky from Earth, while the absolute magnitude describes the intrinsic brightness of an object as it would appear if it were placed at 10 parsecs distance from Earth. The brighter an object appears, the lower the value of its magnitude, with the brightest objects reaching negative values. The Sun has an apparent magnitude of -27 , the full moon -13 .

Generally astronomers study the distance modulus, age, metallicity, mass of a celestial object. The distance modulus is the difference between the apparent magnitude and absolute magnitude of a celestial object ($m - M$), and provides a measure of the distance to the object, r . The relationship between the distance modulus $m - M$ and distance in parsecs r is

$$m - M = 5 \log_{10}\left(\frac{r}{10}\right),$$

where m is the the apparent magnitude of the star and M is the absolute magnitude of the star. The age of a celestial object is generally in units like million year (Myr) or Giga year (Gyr). Metallicity is the fraction of mass of a star or other kind of astronomical object that is not in hydrogen or helium. In this thesis, we use the $\log_{10}$ scale of metallicity. Mass of astronomical objects are usually in solar mass, $M_{\odot}$.

1.2.1 Distance Modulus to the Large Magellanic Cloud

The Large Magellanic Cloud (LMC) is a satellite galaxy of the Milky Way. Its distance modulus is important for precisely determining extragalactic distance (De Grijs et al., 2014). Many researchers have attempted to estimate the distance modulus to the Large Magellanic Cloud (e.g. Pietrzyński et al., 2013; Inno et al., 2013; Muraveva

et al., 2014; Inno et al., 2016; Bhardwaj et al., 2016; Wagner-Kaiser et al., 2017), producing different estimates of the distance modulus based on diverse methods and observations.

We intend to propose a statistical method to systematically review different estimates of the distance modulus to the LMC and produce a comprehensive conclusion based on Bayesian hierarchical modelling. We perform a tentative hierarchical analysis on some published results and obtain our new results in Chapter 3.

1.2.2 Stellar Evolution

Stellar evolution is a branch of astronomy which focuses on how stars are formed and how they evolve over time. Based on the power source, the lifetime of a star is often divided into the following phases. Generally speaking, a star forms when interstellar matter collapses into a ball of hot plasma. When the mass of the ball is large enough, it usually ignites thermonuclear reaction that fuses hydrogen into helium. Then it will become a main-sequence (MS) star and the fusion of hydrogen into helium continues. Depending on its mass and composition, this phase can last for millions to billions of years. Stars with larger mass tend to be denser, hotter and faster to burn their hydrogen up. After the hydrogen in the core has been consumed the core collapses, causing the ignition of hydrogen in the area around the core. Simultaneously, the diameter of the star surges and its surface temperature decreases, becoming a red giant (RG) star. This period is much shorter, lasting only as long as one-tenth of the previous process. In this phase the temperature of its core continues to increase and it is possible to fuse helium into heavier elements, such as, carbon, oxygen, neon and so on. Then for a star weighing less than $8M_{\odot}$, its core stabilises as the outer layer blow away, forming a small hot dense core, called a white dwarf star. Nuclear fusion ceases in a white dwarf, and it gradually cools by radiating light.

Astronomers can observe stars through telescopes and obtain their photometric magnitudes over different filters. Stars in varying stages can be distinguished by their locations in the colour-magnitude diagram (CMD). Figure 1.1 shows the CMD for Hyades cluster, in which each dot is a star in the Hyades. The red and blue solid lines are the fitted main sequence/red giant (MS/RG) and white dwarf (WD) models, respectively. We talk about these stellar evolution models in detail later. Alongside the long solid red line are main-sequence or red giant stars. Stars nearby the solid blue curve are white dwarfs. Stars far away from these two lines are field stars, which are not members of Hyades.

For stars, the properties of main concern are their distance moduli to Earth, initial masses (their masses as zero-age main sequence stars), ages, metallicities. For simplicity, denote A , M , D , Z the base 10 logarithm of age, the solar mass, the distance modulus and the metallicity of a star. These parameters of a star are closely relevant to its photometric magnitudes as it evolves. More massive stars burn their hydrogen faster, thus leave the main-sequence earlier. Older stars tends to be fainter, while younger stars are brighter, emitting more energy. Also, more distant stars are dimmer. Metallicity measures the abundance of metals (in astronomy elements heavier than helium), which affects the star's emission. To characterise the evolution process of stars, astrophysicists have developed a number of mathematical models, such as main sequence red giant models, white dwarf cooling models, initial-final mass relation (IFMR) models, etc.

Before introducing astrophysical models for stars, we point out a couple of facts that influence our statistical methods. Firstly, between one third and one half of main sequence and red giant stars are binary or multiple systems (Jao et al., 2009; Sollima et al., 2010), which are hard to be decomposed into component stars. The luminosities of the component stars sum, leading to a difference from the expected magnitudes

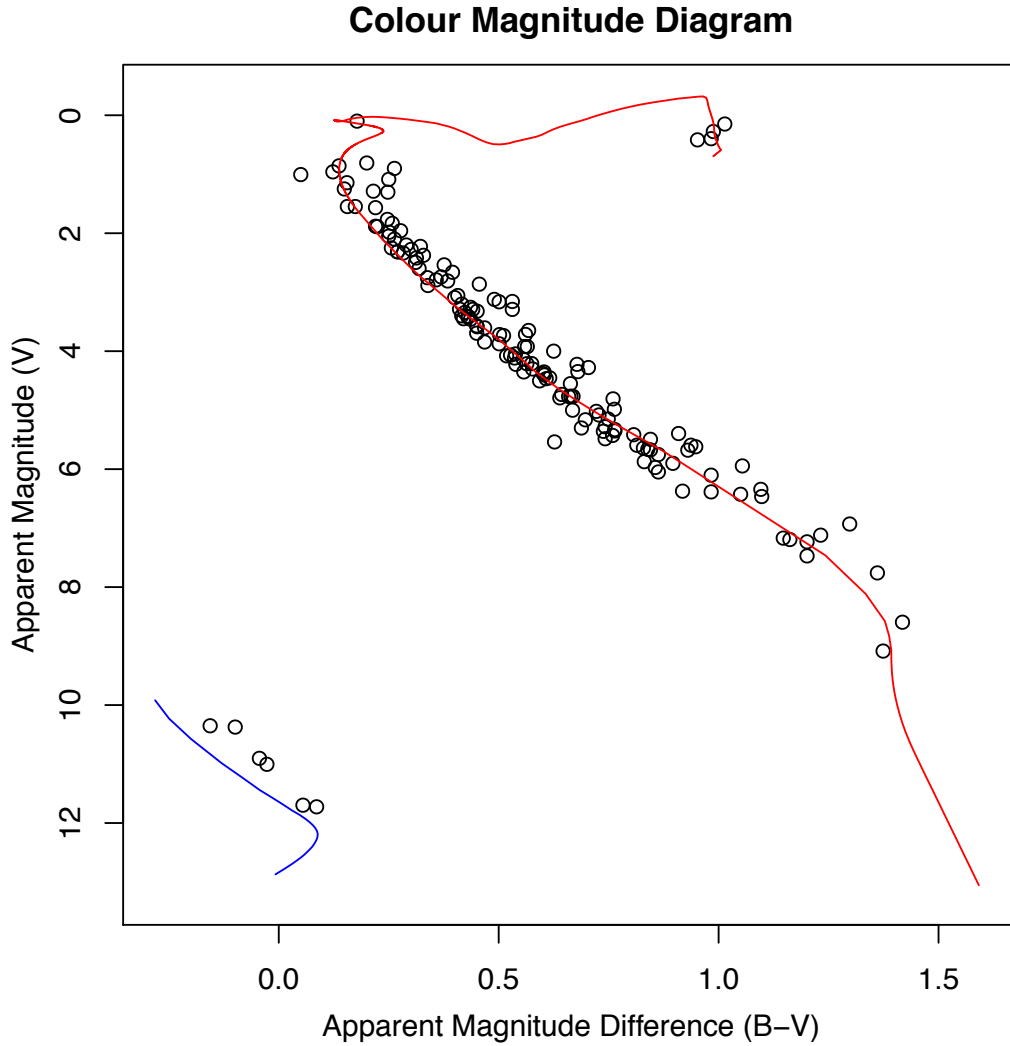


Figure 1.1: The CMD plot for Hyades. Each dot indicates one star in the Hyades cluster. The units on y-axis are magnitudes on V filter with lower values indicating brighter sources. The units on x-axis are difference of magnitudes between B and V filters. The solid red and blue lines are the fitted MS/RG and WD models. References: DeGennaro et al. (2009); van Dyk et al. (2009); Stein et al. (2013).

under the computer model with parameters set at the cluster values. Therefore, we treat all main sequence and red giant stars as binary systems. Secondly, because white dwarf + white dwarf binary systems are relatively rare (maybe 10% of the cluster white dwarfs (Stein et al., 2013)) all white dwarfs are modelled as single stars. Thirdly, for a star cluster consisting of numerous main sequence, red giant and white dwarf stars, their ages, distance moduli, metallicities are very similar. Only their masses differ. Lastly, stars that appear from the Earth’s observation point to be in or close to a star cluster but are not actually members of the cluster are called field stars. Even in the best studied cluster, some stars’ memberships are yet to be confirmed. In our fittings, every star can be either a cluster member or a field star. We set ζ to be the indicator of the membership of a star. If it is a cluster star, then $\zeta = 1$, otherwise $\zeta = 0$. For more details about statistical analysis of stellar evolution, refer to von Hippel et al. (2006); van Dyk et al. (2009); DeGennaro et al. (2009); O’Malley et al. (2013).

For convenience, we denote the observed magnitude in filter band j of star i as x_{ij} for $j = 1, \dots, N_i$ and $i = 1, \dots, I$, with I the number of stars in the dataset and N_i the number of magnitudes for the i -th star. The vector of observed magnitudes for star i , $\mathbf{X}_i = (x_{i1}, \dots, x_{iN_i})^T$, has known variance-covariance matrix Σ_i .

We denote $\Theta = (A, D, Z)$, the vector of cluster parameters. Given the initial mass (as a main sequence star) M_i of star i and cluster parameters Θ , the main sequence red giant (MS/RG) model predicts the expected magnitudes:

$$\mathbf{G}_{\text{MS/RG}}(M_i, \Theta).$$

Taking the binary system into consideration, from the astrophysical definitions, we

derive the expected magnitude for the j filter

$$\mu_{ij} = -2.5 \log_{10}(10^{-G_{\text{MS/RG},j}(M_i, \Theta)/2.5} + 10^{-G_{\text{MS/RG},j}(M_i R_i, \Theta)/2.5}) \quad (1.25)$$

where $G_{\text{MS/RG},j}(M_i \cdot R_i, \Theta)$ is the predicted magnitude in the j th filter for the secondary star in the binary system and $M_i \cdot R_i$ is the secondary mass. We let $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{iN_i})^T$ be the vector of predicted values of $\mathbf{X}_i$ if star i is a main sequence or red giant star. See Stein et al. (2013) for more details. For more about the MS/RG models, see Yi et al. (2001); Dotter et al. (2008).

For a white dwarf i , the computer model is more complicated. The MS/RG model computes how long the white dwarf lived as a main sequence red giant star, called the progenitor age,

$$\phi_{\text{progage},i} = F_{\text{MS/RG}}(Z, M_i), \quad (1.26)$$

where $F_{\text{MS/RG}}$ is the MS/RG computer model for white dwarfs. Then the star's age as a white dwarf can be computed, which is simply the total age A_i , less the progenitor age, ψ_{progage} . The WD cooling model calculates the effective temperature, ψ_{Teff} , and radius of the WD as a function of its age and white dwarf mass:

$$(\psi_{\text{Teff},i}, \psi_{\text{radius},i}) = \mathbf{F}_{\text{cooling}}(A_i - \psi_{\text{progage},i}, M_{\text{WD},i}), \quad (1.27)$$

where $A_i - \psi_{\text{progage},i}$ is the cooling age of star i and $M_{\text{WD},i}$ is the final mass. After $\psi_{\text{radius},i}$ is obtained, we can compute the base-10 logarithm of the gravity experienced at the surface of white dwarf i , denoted $\psi_{\log g}$. From Newton's law, the surface gravity is

$$\psi_{\log g} = \log_{10}(GM_{\text{WD},i}/\psi_{\text{radius},i}^2), \quad (1.28)$$

where G is Newton's gravitational constant.

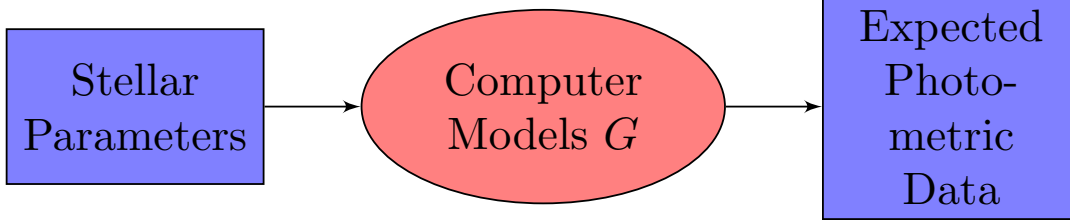


Figure 1.2: Stellar parameters and expected photometry data.

Finally, the WD atmosphere model returns the expected magnitudes for WD i ,

$$\boldsymbol{\mu}_i = \mathbf{F}_{\text{atmosphere}}(\psi_{T_{\text{eff}},i}, \psi_{\log g,i}). \quad (1.29)$$

The IFMR bridges the WD mass and its initial mass as a MS star. We consider the parametrised form for the IFMR,

$$M_{\text{WD},i} = f(M_i, \boldsymbol{\alpha}). \quad (1.30)$$

These models connect characteristics of a star to its photometric magnitudes. Astronomers can observe photometry data of stars via telescopes with measurement errors. The statistical model between photometry $\mathbf{X}$ and characteristics of a star is presented as follows:

$$E(\mathbf{X}|A, M, D, Z) = G(A, M, D, Z), \quad (1.31)$$

in which $G(\cdot)$ summarizes astrophysical models. In this study, computer models The role that computer models play in our research is summarised in Figure 1.2.

One notable attribute of astrophysical models in $G(\cdot)$ is that they are usually highly complicated, nonlinear and lack closed forms, which makes the statistical inference on A, M, D, Z extremely difficult. The joint marginal posterior distribution

of A, M, D of a serpent shape in figure 1.3 gives a glimpse of the complexity of $G(\cdot)$. It plots a Monte Carlo sample from the 3-dimensional posterior distribution of mass, distance modulus, and $\log_{10}(\text{Age})$ for a typical WD, obtained with BASE-9. We include this figure to show the complex shape of the posterior distribution and the non-linear correlations among the parameters. We did not show the specific values of the parameters, because they are irrelevant to the purpose of this figure. This is the candidate’s own work, and this WD, J0822+3903 (Dame et al., 2016), is included in hierarchical analysis in Chapter 4.

To tackle this challenge, a large number of researchers turn to Bayesian statistics by taking advantage of state-of-the-art computational tools, such as Markov chain Monte Carlo (MCMC). van Dyk et al. (2009) proposed principled statistical methods to incorporate intricate stellar evolution models in a Bayesian framework. In O’Malley et al. (2013), a self-consistent and robust Bayesian statistical approach is employed to derive the ages, distances, metallicities and masses of white dwarfs. The program package, BASE-9 (von Hippel et al., 2014), short for Bayesian Analysis of Stellar Evolution with 9 parameters, is a comprehensive and dynamic toolkit to fit the parameters of a cluster or a star given their or its photometry.

In astrophysics, we often aim to estimate a set of parameters for each of a number of objects in a population, which motivates the application of hierarchical models. Noticeably, these models have gained popularity in astronomy mainly for two reasons. Firstly, they provide an approach to combining multiple sources of information. For instance, Mandel et al. (2009) employed Bayesian hierarchical models to analyse the properties of Type Ia supernova light curves by using data from Peters Automated InfraRed Imaging TElescope and the literature. Secondly, they generally produce estimates with less uncertainty. By combining information from supernovae Ia lightcurves, March et al. (2014) and Shariff et al. (2016) illustrate how a hierarchi-

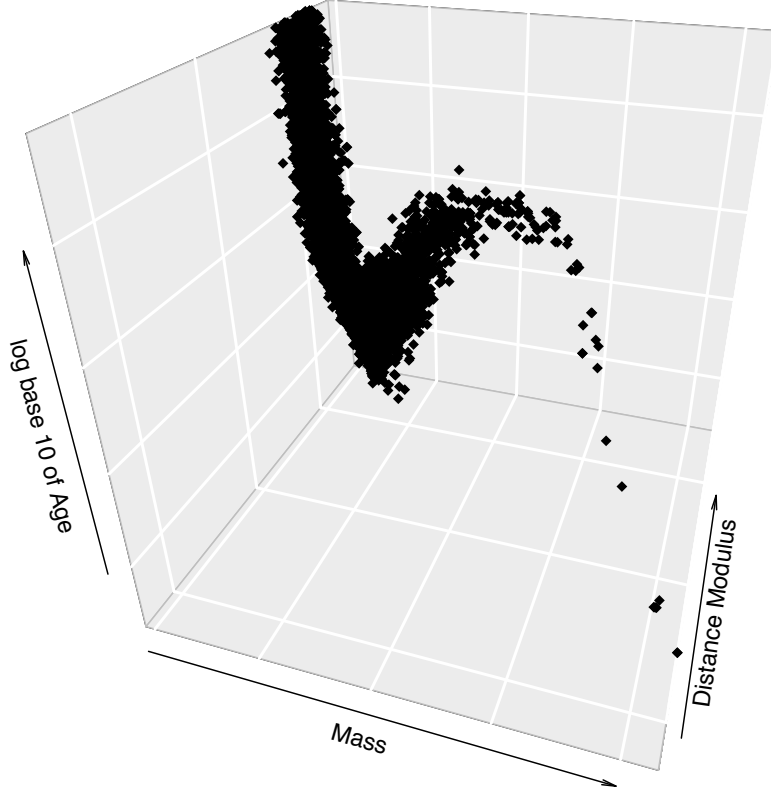


Figure 1.3: A WD’s 3-dimensional scatterplot of its marginal posterior distribution of mass, distance modulus, and $\log_{10}(\text{Age})$. Its shape is complicated, and the parameters have non-linear correlations. The specific values of the parameters are not included, because they are not relevant to the purpose of this figure. This WD, J0822+3903 (Dame et al., 2016), is included in analysis in Chapter 4.

cal model can improve the estimates of cosmological parameters. Similarly, Licquia and Newman (2015) obtained improved estimates of several global properties of the Milky Way by using a hierarchical model to combine previous measurements from the literature.

We focus on two projects in stellar evolution: hierarchical models of the age of a group of Galactic halo WDs and the IFMRs of a number of star clusters.

1.2.3 Hierarchical Modelling of the Age of Galactic Halo White Dwarfs

Age is one of the key information of an astronomical object, such as galaxy, star, cluster, etc. Our first project is to infer the age of our Galaxy via combining the age information in multiple white dwarfs. A galaxy like the Milky Way consists of three parts: (1) the disk, (2) the halo, and (3) the central bulge. These components are indicated schematically in the adjacent figure 1.4. The Galactic halo is a spherical volume of scattered stars, globular clusters surrounding a spiral galaxy. The halo white dwarfs are believed to be older than stars in the inner part of a galaxy, so they are of great importance to infer the formation and age of the galaxy.

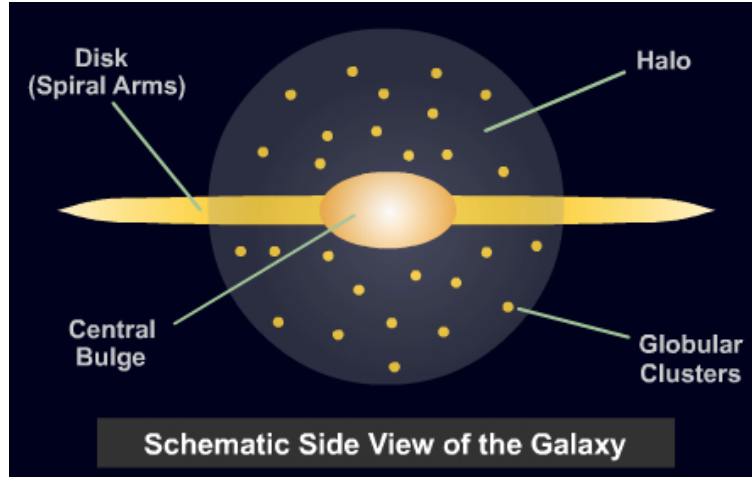


Figure 1.4: The figure of a Galaxy. Source: http://csep10.phys.utk.edu/OJTA2dev/ojta/course2/milkyway/components/components_t1.html-save.

In the first part, we investigate the age of a Galactic halo via a hierarchical model. Since the age of astronomical objects is usually huge, we take the base 10 logarithm of it. So when we talk about age, we sometimes means its logarithm. The data available are the photometry of a few white dwarfs. To make it clearer, denote A_i, D_i, Z_i, M_i the i -th star's age, distance modulus, metallicity, initial mass. The hierarchical model

is

$$(\mathbf{X}_i | A_i, D_i, Z_i, M_i) \sim N(G(\mathbf{A}_i, \mathbf{D}_i, \mathbf{Z}_i, \mathbf{M}_i), \Sigma_i), i = 1, 2, \dots, K; \quad (1.32)$$

$$A_i \sim N(\mu, \tau^2), \quad (1.33)$$

with Σ_i are known from telescopes, $G(\cdot)$ the computer-based white dwarf model. Also, for other parameters D_i, Z_i, M_i , astronomers can provide very good priors on them. The main objective is to make inferences about μ and τ^2 and A_i , $i = 1, \dots, K$.

1.2.4 Hierarchical Modelling of the Initial-Final Mass Relations for Star Clusters

Another problem in stellar evolution is to investigate the functional relationship between the initial and final mass of a white dwarf, the so-called initial-final mass relation (IFMR). The initial mass or zero-age main-sequence (ZAMS) mass of a star is the mass of a star when it starts to be main-sequence star. The final mass of a star is its mass when it evolves to be a WD, so the final mass is also called WD mass. There are quite a few estimated IFMRs in the literature, such as, Weidemann (2000); Williams et al. (2004); Kalirai et al. (2005, 2008); Salaris et al. (2009); Williams et al. (2009). IFMRs are very important when fitting stellar models. Taking different IFMRs may lead to significantly different results on the features of a cluster. From astrophysical theory, IFMRs are monotonically increasing functions, which can be approximated by locally linear segments. The masses for white dwarfs from the same star cluster are somewhat similar. So when fitting a dataset for one star cluster, we can deploy a linear IFMR for each cluster.

For better interpretation, we formulate the IFMR as follows.

$$M_{ij}^{WD} = b_i + a_i(M_{ij} - 3.0), \quad (1.34)$$

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N(\boldsymbol{\gamma}, \boldsymbol{\Sigma}), \quad (1.35)$$

where M_{ij}, M_{ij}^{WD} are the initial mass and WD mass of the j -th $j = 1, 2, \dots, N_i$ star in the i -th cluster $i = 1, 2, \dots, n$, $\epsilon_{ij} \sim N(0, \sigma^2)$. In Eq. (1.34), a_i is the slope of the linear IFMR for i -th cluster and b_i the WD mass when its ZAMS mass is $3.0M_{\odot}$.

1.3 Review of Statistical Computational Methods

Due to the complexity of the hierarchical models, we develop a number of statistical computational methods to fit them. Before presenting our algorithms, we cover several fundamental computational techniques for hierarchical models.

1.3.1 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a class of algorithm which can draw samples from probability density functions known up to a constant of proportionality. This idea originated from Metropolis et al. (1953), which proposed the well-known Metropolis Algorithm (MA).

Here we cover how MCMC algorithms work. Markov chain is one kind of stochastic process that possesses memoryless property. In a mathematical notation, Markov chain $\{X_n, n = 1, 2, \dots\}$ satisfies

$$P(X_{n+1}|X_1, \dots, X_n) = P(X_{n+1}|X_n), \quad (1.36)$$

which means that the distribution of X_{n+1} given all previous states only relies on X_n rather than all other preceded ones. Under regularity conditions, Markov chain has an invariant distribution satisfying

$$\pi(y)dy = \int P(dy|x)\pi(x)dx \quad (1.37)$$

where $P(dy|x)$ is the transition kernel, i.e., the probability of moving from x to the neighbourhood $(y, y + dy)$. After a Markov chain reaches its equilibrium, its trace contains all information of its invariant distribution. Therefore, it is reasonable to record the trace of a Markov chain in stationary state and make inferences on its invariant distribution based on its trace. This is the basic logic beneath MCMC methods.

MCMC methods turn the Markov chain theory around: the invariant distribution $\pi(x)$ is known (perhaps up to a constant multiple). The objective is to learn $\pi(x)$, such as, its mean, variance, modes and how it is distributed in the sample space. We can construct a Markov chain who converges to invariant distribution $\pi(x)$ and record the trace of this chain as a sample from $\pi(x)$. The major issue when implementing MCMC methods is how to find a transition kernel that satisfies Eq. (1.37). It can be proved that if a conditional probability density $p(y|x)$ satisfies the detailed balance condition

$$\pi(x)p(y|x) = \pi(y)p(x|y), \quad (1.38)$$

then $P(dy|x) = p(y|x)dy$ satisfies Eq. (1.37) (See Tierney, 1994).

Usually, it is hard to find a function $p(y|x)$ who right fits Eq. (1.38). Generally, we can construct $p(y|x)$ in the following way. Suppose we have a proposal distribution $q(y|x)$ which can generate candidates. If $\pi(x)q(y|x) = \pi(y)q(x|y)$, then we have obtained $p(x, y)$. But this is highly unlikely to happen. We might find, for example,

that for some x, y ,

$$\pi(x)q(y|x) > \pi(y)q(x|y),$$

which means that the chain moves from x to y more frequently than the other way around. To correct this problem, another function is introduced. Since the chain moves from y to x too rarely, we set acceptance probability $A(y, x) = 1$ and the probability of accepting movement from x to y is set to $A(x, y) = \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}$. For the case that $\pi(x)q(y|x) < \pi(y)q(x|y)$, we set the acceptance probability the other way around. For all cases, it holds that

$$A(x, y) = \min \left\{ \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1 \right\}.$$

It is trivial to verify that

$$\pi(x)q(y|x)A(x, y) = \pi(y)q(x|y)A(y, x).$$

So we can take $p(y|x) = q(y|x)A(x, y)$ and it satisfies the detailed balance condition.

Metropolis et al. (1953) proposed a symmetric proposal distribution, i.e., $p(x|y) = p(y|x)$, and the MCMC algorithm is known as MA. Subsequently in Hastings (1970), MA is generalised to permit proposal rules of any shape, i.e., Metropolis-Hastings Algorithm (MHA). One commonly seen challenge of MHA is that it might get stuck in one local mode of multimodal target densities, which causes the incomplete exploration of the whole sample space and hence leads to biased results.

To combat this situation, a large number of variants of MCMC algorithm have been developed by researchers to ensure the chain to traverse the sample space thoroughly. For example, parallel tempering (Geyer, 1991) runs multiple chains under various temperatures at the same time and allows two neighbouring chains to swap

samples. Multiple-Try Metropolis algorithm (Liu et al., 2000; Liu, 2008) makes numerous proposals in each iteration. To achieve the efficient mixing of Markov chain, Haario et al. (2001) develops an adaptive MCMC which updates the variance of the Gaussian proposal based on the sample variance of previous draws. Also, the theoretical justification of adaptive MCMC for random walk metropolis is violated when Gaussian proposal is varied in Haario et al. (2001). One implicit assumption of their adaptive Markov chain Monte Carlo (MCMC) is that the adaptation is applicable to every point in its state space, which might not be the best strategy in most practical settings. Having realised this, Craiu et al. (2009); Bai et al. (2011) suggests regional adaptation for proposal distribution in MCMC. Leman et al. (2009) proposes multi-set sampler, which traverses the whole sample space smoothly.

Another important kind of MCMC algorithm is Gibbs sampler (Geman and Geman, 1984), which is specifically designed for multivariate density functions. It samples iteratively from the conditional density for each variable given the newest updates for other variables.

The convergence of MCMC algorithms is a vital issue when they are applied in practice. Many researchers have attempted to address this problem (See, e.g. Gelman and Rubin, 1992; Cowles and Carlin, 1996; Brooks and Gelman, 1998; Nylander et al., 2007; Gelman and Shirley, 2011). Usually we use diagnostic tools to monitor the convergence of MCMC algorithms, such as, the time series plot of the samples, the autocorrelation plot, and Gelman-Rubin “shrink factor” (Gelman and Rubin, 1992).

1.3.2 Expectation-Maximisation algorithm

This subsection presents a brief description of expectation-maximization (EM) algorithm, why it works and its convergence property. The basic setting for the EM algorithm is that complete data $(\mathbf{X}, \mathbf{Y})$ is split into observed data $\mathbf{X}$ and missing

data $\mathbf{Y}$. Suppose the observed data $\mathbf{X} = (X_1, \dots, X_n)$ and the full likelihood function is $f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})$, with $\mathbf{Y}$ missing and $\boldsymbol{\theta}$ of interest. The ultimate goal is to make inferences on $\boldsymbol{\theta}$ based on observed data $\mathbf{X}$. Clearly, one would attempt to maximise the marginal likelihood of $\mathbf{X}$ by integrating out $\mathbf{Y}$, namely,

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \int f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) d\mathbf{Y}.$$

However, this way is usually intractable because the marginal distribution of $\mathbf{X}$ might require solving a complicated integral.

The EM algorithm, proposed by Dempster et al. (1977), is tailored to find the maximum likelihood estimate of $\boldsymbol{\theta}$ when one part of the data $\mathbf{Y}$ is missing. It works in the following way:

- Initialization: $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$;
- E-step: Given the t -th iteration $\boldsymbol{\theta}^{(t)}$, find the expectation:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E[\log f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(t)}];$$

- M-step: The $t + 1$ -th iteration

$$\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

The reason why EM works can be found in Dempster et al. (1977), and Wu (1983). Here is the basic sketch of the proof. From manipulations of conditional densities, we have

$$f(\mathbf{X}|\boldsymbol{\theta}) = f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})/f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}).$$

Taking logarithm and conditional expectation given $\boldsymbol{\theta}^{(t)}$ and $\mathbf{X}$ to both sides lead to:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log f(\mathbf{X}|\boldsymbol{\theta}) \\ &= E[\log f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}^{(t)}, \mathbf{X})|\boldsymbol{\theta}^{(t)}, \mathbf{X}] - E[\log f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(t)}] \\ &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \end{aligned}$$

By using Jensen inequality, for all $\boldsymbol{\theta}$ we have

$$\begin{aligned} H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E[\log f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) - \log f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(t)}] \\ &= -E\left[\log \frac{f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})}{f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}^{(t)})}\middle|\mathbf{X}, \boldsymbol{\theta}^{(t)}\right] \\ &\geq -\log E\left[\frac{f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})}{f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}^{(t)})}\middle|\mathbf{X}, \boldsymbol{\theta}^{(t)}\right] \\ &\geq 0. \end{aligned}$$

EM is very popular in many applied areas. But one notorious drawback of it is its slow convergence. As proved by Dempster et al. (1977), the speed matrix of EM algorithm is the matrix fraction of the observed information, $S = \mathcal{I}_{com}^{-1}\mathcal{I}_{obs}$, where

$$\mathcal{I}_{obs} = -\frac{\partial^2 \log p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \cdot \partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{MLE}}$$

is the observed data information matrix, and

$$\mathcal{I}_{com} = -E_{(X,Y)}\left\{\frac{\partial^2 \log p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \cdot \partial \boldsymbol{\theta}'} \middle| \mathbf{X}, \boldsymbol{\theta}\right\} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{MLE}}$$

is the complete data information matrix.

Exemple: Gaussian-Gaussian Hierarchical Model

For $j = 1, 2, \dots, J, i = 1, 2, \dots, N_j$,

$$\begin{aligned} Y_{ij} &= \mu_j + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2); \\ \mu_j &\sim N(\gamma, \tau^2), \end{aligned}$$

with $\mathbf{Y} = (Y_{11}, \dots, Y_{N_1 1}, \dots, Y_{N_J J})$ the observed data, the parameters of interest are $\theta = (\gamma, \tau^2, \sigma^2)$. $\boldsymbol{\mu} = (\mu_1, \dots, \mu_J)$ will be treated as latent variable or missing data. The likelihood function for the complete data $\mathbf{Y}$ and $\boldsymbol{\mu}$ is given by

$$\begin{aligned} p(\mathbf{Y}, \boldsymbol{\mu} | \gamma, \tau^2, \sigma^2) &= p(\mathbf{Y} | \boldsymbol{\mu}, \gamma, \tau^2, \sigma^2) p(\boldsymbol{\mu} | \gamma, \tau^2, \sigma^2) \\ &= \prod_{j=1}^J \prod_{i=1}^{N_j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (Y_{ij} - \mu_j)^2 \right\} \cdot \prod_{j=1}^J \frac{1}{\sqrt{2\pi\tau^2}} \exp \left\{ -\frac{1}{2\tau^2} (\mu_j - \gamma)^2 \right\}. \end{aligned}$$

By taking the logarithm, we have:

$$\begin{aligned} \log p(\mathbf{Y}, \boldsymbol{\mu} | \gamma, \tau^2, \sigma^2) &= \sum_{j=1}^J \sum_{i=1}^{N_j} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(Y_{ij} - \mu_j)^2}{2\sigma^2} \right] \\ &\quad + \sum_{j=1}^J \left[-\frac{1}{2} \log(2\pi\tau^2) - \frac{(\mu_j - \gamma)^2}{2\tau^2} \right], \end{aligned}$$

and the expected complete data log likelihood is

$$\begin{aligned}
Q(\gamma, \tau^2, \sigma^2 | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}) \\
= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^J \sum_{i=1}^{N_j} E[(Y_{ij} - \mu_j)^2 | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}] \\
- \frac{J}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \sum_{j=1}^J E[(\mu_j - \gamma)^2 | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}] \quad (1.39)
\end{aligned}$$

It can be derived that the conditional distribution of μ_j given observed data $\mathbf{Y}$ and $\gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}$, is

$$(\mu_j | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}) \sim N(w_j \gamma^{(t)} + (1 - w_j) \bar{Y}_{.j}, v_j^2), \quad (1.40)$$

with $w_j = \sigma^{2(t)} / (\sigma^{2(t)} + N_j \tau^{2(t)})$, $\bar{Y}_{.j} = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ij}$, $v_j^2 = w_j \tau^{2(t)}$.

The E-step consists of computing

$$\begin{aligned}
\hat{\mu}_j &= E[\mu_j | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}] = w_j \gamma^{(t)} + (1 - w_j) \bar{Y}_{.j}, \\
\hat{\mu}_j^2 &= E[\mu_j^2 | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}] = v_j^2 + \hat{\mu}_j^2; \\
E\left[\sum_{j=1}^J \mu_j | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}\right] &= \sum_{j=1}^J \hat{\mu}_j, \\
E\left[\sum_{j=1}^J \sum_{i=1}^{N_j} (Y_{ij} - \mu_j)^2 | \mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}\right] &= \sum_{j=1}^J \sum_{i=1}^{N_j} (Y_{ij} - \bar{Y}_{.j})^2 + \sum_{j=1}^J N_j [\bar{Y}_{.j}^2 + \hat{\mu}_j^2 - 2\bar{Y}_{.j} \hat{\mu}_j], \\
E\left[\sum_{j=1}^J (\mu_j - \gamma)^2\right] &= \sum_{j=1}^J [\hat{\mu}_j^2 + \gamma^2 - 2\gamma \hat{\mu}_j].
\end{aligned}$$

Then in the M-step, compute the value of γ , σ^2 and τ^2 that maximize Eq. (1.39).

Combining this with the E-step yields the EM iteration:

$$\begin{aligned}\gamma^{(t+1)} &= \frac{1}{J} \sum_{j=1}^J \left[w_j \gamma^{(t)} + (1 - w_j) \bar{Y}_{\cdot j} \right]; \\ \sigma^{2(t+1)} &= \frac{1}{N} \left\{ \sum_{j=1}^J \sum_{i=1}^{N_j} (Y_{ij} - \bar{Y}_{\cdot j})^2 + \sum_{j=1}^J N_j [w_j^2 (\gamma^{(t)} - \bar{Y}_{\cdot j})^2 + v_j^2] \right\}; \\ \tau^{2(t+1)} &= \frac{1}{J} \left\{ \sum_{j=1}^J [w_j \gamma^{(t)} + (1 - w_j) \bar{Y}_{\cdot j}]^2 + \sum_{j=1}^J v_j^2 \right\} - (\gamma^{(t+1)})^2.\end{aligned}$$

1.3.3 Monte Carlo Expectation-Maximisation Algorithm

In this section, we briefly introduce the MCEM algorithm. In last section, the EM algorithm can be used to find the maximum likelihood estimate for $\boldsymbol{\theta}$. In the E-step, it is necessary to find the expected value of the complete log-likelihood function conditional on observed data and the previous iteration, namely,

$$E[\log f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(t)}] = \int \log f(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta}) p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) d\mathbf{y}.$$

For distributions in the exponential family, this integral is usually solvable, whereas for some complicated distribution, it might be intractable. Therefore, EM algorithm might be not applicable in cases where there is no closed form in the E-step. A reasonable remedy would be to use Monte Carlo approximation as an alternative to the integral (Wei and Tanner, 1990), which means that

$$E[\log f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(t)}] \approx \frac{1}{S} \sum_{s=1}^S \log f(\mathbf{X}, \mathbf{Y}^{[t,s]}|\boldsymbol{\theta}),$$

with $\mathbf{Y}^{[t,s]} \sim p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^{(t)})$, $s = 1, \dots, S$.

In a nutshell, the MCEM works as follows. Initialize parameters first, $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$.

For $t = 1, 2, \dots$,

- MC step: Given the t -th iteration $\boldsymbol{\theta}^{(t)}$, sample $\mathbf{Y}^{[t,s]} \sim p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}^{(t)})$, for $s = 1, 2, \dots, S$;

- E step: Evaluate Monte Carlo expectation

$$Q_{MC}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \frac{1}{S} \sum_{s=1}^S [\log f(\mathbf{X}, \mathbf{Y}^{[t,s]}|\boldsymbol{\theta})];$$

- M step:

$$\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} Q_{MC}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

For the Gaussian-Gaussian hierarchical model, given the t -th update for $\boldsymbol{\theta}^{(t)} = (\gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)})$, draw a sample of size S for unobservable data $\mu_j^{[t,s]}, j = 1, 2, \dots, J$ from their conditional distributions in Equation 1.40. From ordinary EM algorithm, we have that

$$\begin{aligned} \gamma^{(t+1)} &= \frac{1}{J} \sum_{j=1}^J E(\mu_j|\mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}); \\ \tau^{2(t+1)} &= \frac{1}{J} \sum_{j=1}^J E[(\mu_j - \gamma^{(t+1)})^2|\mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}]; \\ \sigma^{2(t+1)} &= \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^{N_j} E[(Y_{ij} - \mu_j)^2|\mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}, \sigma^{2(t)}]. \end{aligned}$$

For MCEM, the expectation will be replaced by sample mean, namely,

$$\begin{aligned}\gamma^{(t+1)} &= \frac{1}{J \cdot S} \sum_{s=1}^S \sum_{j=1}^J \mu_j^{[t,s]}, \\ \tau^{2(t+1)} &= \frac{1}{J \cdot S} \sum_{j=1}^J \sum_{s=1}^S [(\mu_j^{[t,s]} - \gamma^{(t+1)})^2]; \\ \sigma^{2(t+1)} &= \frac{1}{N \cdot S} \sum_{s=1}^S \sum_{j=1}^J \sum_{i=1}^{N_j} E[(Y_{ij} - \mu_j^{[t,s]})^2].\end{aligned}$$

For the convergence of MCEM algorithm, see Chan and Ledolter (1995) and Fort et al. (2003) for more details. Generally we check the time series plot of the iterations of parameter estimates and see whether it has converged to a constant value.

1.4 Novel Statistical Algorithms Illustrated in Astrophysics

Fully modelling a population of objects within a hierarchical model requires substantial computational investment and often specialised computer code, especially for complicated problems such as the computer astrophysical model $G(\cdot)$. In this study, we develop novel, efficient statistical algorithms to conveniently obtain the improved estimates available under a Bayesian hierarchical model. By taking advantage of the existing code for case-by-case analyses, our methods simultaneously estimate parameters of the individual objects and parameters that describe their population distribution. Our methods specialise in the computational side of Bayesian hierarchical modelling and are known to produce estimators of parameters of the individual objects that are on average closer to their true values than estimators based on the case-by-case analyses. Our algorithms are helpful especially when the likelihood function of each object is highly complicated and computer-based and there is a black-box code available to perform case-by-case analyses. The new algorithms are elaborated

in Chapter 2.

We illustrate our techniques in three settings: 1.) meta-analysis of published results on the distance modulus to the LMC, 2.) pooling a number of Galactic halo WDs in a hierarchical model to obtain more precise estimates of their ages, and 3.) pooling multiple star clusters into a hierarchical model to study the IFMR. We must notify the readers that these applications in astrophysics are tentative and preliminary, aimed to present how the new statistical algorithms can be helpful to astronomy problems.

Our first example is introduced primarily to illustrate a different use of hierarchical models. Rather than a population of objects, we study a collection of estimates of a single parameter: the distance modulus to the LMC. In this setting, we use a hierarchical model to combine a “population” of 13 estimates of the distance to derive a single comprehensive analysis. We include this analysis to show how to combine case-by-case estimates into a Bayesian hierarchical model and produce shrinkage estimates in a simple but realistic setting. Our estimate of the distance modulus to the LMC is 18.525 ± 0.028 . The inferred result is subject to change, depending on the studies included in the hierarchical model. The detailed analysis of 13 published distance modulus to the LMC is shown in Chapter 3.

In the second application, we pool a number of Galactic halo WDs into a Bayesian hierarchical model and employ our efficient algorithms to fit the hierarchical model in both the EB and FB approaches. We perform a simulation study to illustrate the advantage of hierarchical fits over the commonly used case-by-case analyses. We find that approximately two thirds of the estimated WD ages are closer to their true ages under the hierarchical model. Using optical and near-infrared photometry of ten candidates halo WDs, we simultaneously estimate their individual ages and the mean age of these halo WDs; the latter is estimated to be $12.04^{+0.80}_{-0.75}$ Gyr. The standard

deviation of the ages of 10 WDs included in the hierarchical model is estimated to be 0.95 ± 0.48 Gyr. However, it is worth mentioning that these figures are preliminary and may vary, depending upon a number of factors such as astrophysical models and model assumptions. This project is a prototype for pooling large number of WDs in the Gaia data and estimating their ages. For details on this project, refer to Chapter 4.

In the third project, we propose a Bayesian hierarchical model for cluster IFMRs, show how this model can be fit using existing software, and use a suite of simulation studies to verify the statistical advantages of the resulting IFMR estimates. This work is also introductory, aimed at illustrating how our statistical methods can be employed to estimate IFMRs of star clusters. We apply the Bayesian hierarchical model to fit five clusters: the Hyades, M67, NGC 188, NGC 2168 and NGC 2477. We obtain the shrinkage estimates of IFMR parameters for these five clusters. We estimate the mean linear IFMR of these five clusters to be

$$M_{\text{final}} = (0.09 \pm 0.04)M_{\text{initial}} + (0.44 \pm 0.14)M_{\odot}.$$

In addition, for these five clusters, the standard deviations of their IFMR slope and intercept are estimated as $0.07^{+0.08}_{-0.07}$ and 0.26 ± 0.23 , respectively. All figures presented here are subject to change, depending upon the models and data we use. The detailed analysis of IFMRs is shown in Chapter 5.

In the ages of Galactic halo WDs projects, we treat the WDs as objects, whose $\log_{10}$ ages comes from the same population–Gaussian distribution. So we combine multiple WDs into one single comprehensive model. Similarly, in the IFMRs of star clusters, clusters are treated as objects and IFMRs of clusters are assumed from the same bivariate normal distribution. The power of Bayesian hierarchical modelling can

be conceived of as coming from 1) an informative assumption, which is that all the objects belong to a population with a particular distribution of the parameters of the objects across the population, and 2) that it otherwise is difficult to come up with a technique that can combine the individual results when they may have asymmetric posterior density functions. It is well known in the statistical world that Bayesian hierarchical modelling produces shrinkage estimates that on average have smaller MSE (Morris and Lysy, 2012). So using Bayesian hierarchical modelling to astrophysical problems can improve the precision of estimates. However, fitting Bayesian hierarchical models for complicated astrophysical problems is computationally challenging, which requires our novel statistical algorithms.

One of the primary benefits of our algorithms is that it takes advantages of the existing black-box code which is able to fit one object at a time. We only need to write wrapper code that calls the existing programs. This saves substantial human time investment that might otherwise be devoted to writing new code. Our algorithms make Bayesian hierarchical modelling tractable for complicated astrophysical problems. We see that our algorithms have great potential in many other areas.

In practice hierarchical models can be difficult to implement in complex situations that require sophisticated computer code to fit even a single object. The primary contribution of my thesis is to develop efficient algorithms for fitting hierarchical models in this situation, treating the computer code for fitting a single object as a black box that we do not open and do not need to alter. To give context to my methods, I develop them using several examples from astronomy. These examples are not meant to be definitive in terms of their astronomical results, but are meant to be testbeds for proofs of concept of my new algorithms. The black-box software in my thesis, called BASE-9, can perform case-by-case analysis and produce MCMC samples for stellar or cluster parameters (e.g., von Hippel et al., 2006; van Dyk et al., 2009;

O'Malley et al., 2013; Stein et al., 2013; von Hippel et al., 2014). However, this black-box contains complicated stellar evolution models that are unavailable in closed forms, which causes computational challenges when combining multiple objects such as white dwarfs (WDs) or star clusters in a hierarchical model. I employed my methods to three astronomy problems: the published distance moduli to the Large Magellanic Cloud (LMC), the ages of a group of Galactic halo WDs candidates, and the initial-final mass relations (IFMRs) of a few star clusters. I have done simulation studies under these three settings, and I find that the root mean squared error (RMSE) of hierarchical estimates for object-level parameters reduce by 54.3% for the LMC project, 75.0% for the WD project, and 82.0% for the IFMR project, compared with the estimates from case-by-case analyses. My two-stage sampler algorithm uses case-by-case samples from the black-box and it only takes an additional 10 minutes to produce the shrinkage estimates under the hierarchical models, only increasing the computing time by 3.3% and 0.5% for the WD and IFMR projects, respectively. Because the LMC project uses a Gaussian-Gaussian hierarchical model that is straightforward, its computing time is not important. The interval estimates produced by hierarchical models perform also well, with their actual coverage probabilities close to the nominal level.

In summary, I have developed efficient statistical algorithms that fit Bayesian hierarchical models by using case-by-case fits from a black-box code and produce estimates of good properties, costing a manageable amount of time. The three astrophysical projects are preliminary and tentative, but are not meant to give definitive answers for astronomical problems. I include the LMC project to show how to combine case-by-case fits into a hierarchical model in a simple but realistic setting. The WD hierarchical model is a prototype analysis for Gaia, which will measure exquisite distances to tens of thousands of WDs. The IFMR project is used to illustrate how to combine multiple clusters into a hierarchical model and improve estimates of IFMR

parameters when some of them has insufficient data to constrain the IFMR parameters.

2 Statistical Methodology and Computation

Bayesian hierarchical models combine multiple objects (i.e., units of data) and produce partially pooled (shrinkage) estimates that have better statistical properties such as smaller mean squared errors and better predictive accuracy (Gelman, 2006a; Morris and Lysy, 2012; Gelman et al., 2014; Katahira, 2016). In practice hierarchical models can be difficult to implement in complex situations that require sophisticated computer code (black-box) to fit even a single object. In this Chapter, we develop computationally efficient algorithms that fit Bayesian hierarchical models by taking advantage of the existing black-box code that we do not open and do not need to alter.

2.1 Introduction

In this section, we formulate our two-stage algorithms for fitting general hierarchical models and illustrate how they work. We consider a general hierarchical model, specifically

$$\begin{aligned} y_i &\stackrel{\text{indep}}{\sim} p(y_i|\theta_i, \xi_i), \\ \xi_i &\stackrel{\text{indep}}{\sim} p(\xi_i|\theta_i), \quad \text{for } i = 1, \dots, n, \\ \theta_i &\stackrel{\text{indep}}{\sim} p_H(\theta_i|\psi), \end{aligned} \tag{2.1}$$

Table 2.1: Parameters in the hierarchical model

Object-level parameters	
θ_i	can be univariate or multivariate, and has hierarchical modelling
ξ_i	can be univariate or multivariate, and has no hierarchical modelling
Population-level parameters	
ψ	can be univariate or multivariate

where θ_i and ξ_i are object-level parameters and ψ is the population-level parameter. Notice that only θ_i is the modelled hierarchy, each ξ_i has its own prior distribution. In the linear model setting, the $\xi_i, i = 1, \dots, n$ are called fixed effects and the $\theta_i, i = 1, \dots, n$ are called random effects. Parameters in the model 2.1 are summarised in Table 2.1. The ‘H’ in the subscript of the prior distribution for θ_i indicates its hierarchical nature. We again allow y_i, θ_i and ξ_i to be multivariate and set $\mathbf{y} = (y_1, \dots, y_n)$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$. Finally we assign a prior distribution to ψ as $\psi \sim p(\psi)$.

Although fitting the hierarchical model in (2.1) is our ultimate goal, we fit the case-by-case model as an intermediate step. For the case-by-case model, we suppose

$$y_i | \theta_i, \xi_i \stackrel{\text{indep}}{\sim} p(y_i | \theta_i, \xi_i) \text{ with } (\theta_i, \xi_i) \sim p_{\text{CBC}}(\theta_i, \xi_i)$$

for $i = 1, \dots, n$, where the subscript on the prior distribution indicates its case-by-case, non-hierarchical nature. We set $p_{\text{CBC}}(\theta_i, \xi_i) = p_{\text{CBC}}(\theta_i)p(\xi_i | \theta_i)$, where $p(\xi_i | \theta_i)$ is the prior distribution used in the hierarchical model in (2.1).

We assume that we can obtain a Monte Carlo sample from the case-by-case analysis

$$p_{\text{CBC}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \mathbf{y}) \propto \prod_i^n p(y_i | \theta_i, \xi_i) p_{\text{CBC}}(\theta_i, \xi_i) \quad (2.2)$$

using an existing black-box software package. In the next two sections, we introduce

the two-stage algorithms that obtain Fully Bayesian (FB) and empirical Bayes (EB) fits for the hierarchical model in Eq. (2.1). These algorithms allow researchers to fit hierarchical models easily, using a straightforward wrapper code along with the existing software for a case-by-case analysis. There is no need to develop a new software package to fit the hierarchical models from scratch. This saves significant human effort, especially when the likelihood function $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})$ is complicated.

2.2 Fully Bayesian Approach

The joint posterior distribution of all the parameters under the hierarchical model in Eq. (2.1) is

$$p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi | \mathbf{y}) \propto p(\psi) \prod_{i=1}^n p_{\text{H}}(\theta_i | \psi) p(\xi_i | \theta_i) p(y_i | \theta_i, \xi_i). \quad (2.3)$$

To take advantage of the case-by-case fits in (2.2), we establish a relationship between $p_{\text{CBC}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \mathbf{y})$ and $p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \psi, \mathbf{y})$ for any specified ψ , namely

$$p_{\text{CBC}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \mathbf{y}) \propto \frac{p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \psi, \mathbf{y}) p_{\text{CBC}}(\boldsymbol{\theta})}{p_{\text{H}}(\boldsymbol{\theta} | \psi)}. \quad (2.4)$$

To establish (2.4), first notice that

$$\begin{aligned} p_{\text{CBC}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \mathbf{y}) &= \frac{p(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\xi}) p_{\text{CBC}}(\boldsymbol{\theta}, \boldsymbol{\xi})}{p_{\text{CBC}}(\mathbf{y})} \\ &= \frac{p_{\text{CBC}}(\boldsymbol{\theta})}{p_{\text{CBC}}(\mathbf{y})} p(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\xi} | \boldsymbol{\theta}), \end{aligned} \quad (2.5)$$

where $p_{\text{CBC}}(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\xi}|\boldsymbol{\theta})p_{\text{CBC}}(\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\xi}$. Similarly, under the hierarchical model the conditional posterior distribution of $\boldsymbol{\theta}$ given ψ can be written

$$\begin{aligned} p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}|\psi, \mathbf{y}) &= \frac{p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\xi}|\boldsymbol{\theta})p_{\text{H}}(\boldsymbol{\theta}|\psi)}{p_{\text{H}}(\mathbf{y}|\psi)} \\ &= \frac{p_{\text{H}}(\boldsymbol{\theta}|\psi)}{p_{\text{H}}(\mathbf{y}|\psi)}p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\xi}|\boldsymbol{\theta}), \end{aligned} \quad (2.6)$$

where $p_{\text{H}}(\mathbf{y}|\psi) = \int p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\xi}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\psi)d\boldsymbol{\theta}d\boldsymbol{\xi}$. Finally solving (2.6) for $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\xi}|\boldsymbol{\theta})$, plugging into (2.5), and ignoring terms constant in $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ yields the relationship in (2.4). By marginalizing $(\theta_j, \xi_j), j \neq i$ on both sides of Eq. (2.4), it can be shown that

$$p_{\text{CBC}}(\theta_i, \xi_i|y_i) \propto \frac{p_{\text{H}}(\theta_i, \xi_i|\psi, y_i)p_{\text{CBC}}(\theta_i)}{p_{\text{H}}(\theta_i|\psi)}. \quad (2.7)$$

To obtain samples from the joint posterior distribution $p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi | \mathbf{y})$ in (2.3) under the hierarchical model, we employ a two-stage algorithm. In the first stage, for each object i we sample from the posterior distribution $p_{\text{CBC}}(\theta_i, \xi_i|y_i)$ under the case-by-case analysis. In the second stage, we apply a MH within Gibbs sampler to sample from $p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi | \mathbf{y})$, in which MH is used to update $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ by taking their case-by-case samples as proposals. This can be achieved using the following algorithm.

Two-stage FB Algorithm:

Stage 1: For each (θ_i, ξ_i) , obtain a MCMC sample from their joint case-by-case posterior distribution, and denote it $(\theta_i, \xi_i)^{[\text{CBC}, t]}, t = 1, 2, \dots, T$. Appropriate burn-in and convergence checks are required to get a reasonable chain.

Stage 2 Initialisation: Initialise each (θ_i, ξ_i) to randomly selected values from the case-by-case samples, i.e., $(\theta_i, \xi_i)^{[\text{TS}, 0]} = (\theta_i, \xi_i)^{[\text{CBC}, \tilde{r}_i]}$ with $\tilde{r}_i$, a random number between 1 and T .

Stage 2 Iteration: For $t = 1, \dots, T$

Stage 2a: Generate a draw $\psi^{[\text{TS},t]}$ from $p(\psi|\boldsymbol{\theta}^{[\text{TS},t-1]}) \propto p_{\text{H}}(\boldsymbol{\theta}^{[\text{TS},t-1]}|\psi)p(\psi)$.

Stage 2b: Given $\psi^{[\text{TS},t]}$, update each (θ_i, ξ_i) via a MHIS, where $p_{\text{H}}(\theta_i, \xi_i|y_i, \psi^{[\text{TS},t]})$ is the target conditional distribution and $p_{\text{CBC}}(\theta_i, \xi_i|y_i)$ is the proposal distribution. To sample from $p_{\text{CBC}}(\theta_i, \xi_i|y_i)$, we recycle the sample obtained in **Stage 1**, specifically, we generate random numbers $r_1, \dots, r_n$ between 1 and T , and set the new proposal to $(\theta_i, \xi_i)^{\text{prop}} = (\theta_i, \xi_i)^{[\text{CBC}, r_i]}, i = 1, \dots, n$.

Stage 2c: Compute the n MH acceptance probabilities using the formula (2.7),

$$\begin{aligned} R_i &= \frac{p_{\text{H}}((\theta_i, \xi_i)^{\text{prop}}|\psi^{[\text{TS},t]}, y_i)/p_{\text{CBC}}((\theta_i, \xi_i)^{\text{prop}}|y_i)}{p_{\text{H}}((\theta_i, \xi_i)^{[\text{TS},t-1]}|\psi^{[\text{TS},t]}, y_i)/p_{\text{CBC}}((\theta_i, \xi_i)^{[\text{TS},t-1]}|y_i)} \\ &= \frac{p(\theta_i^{\text{prop}}|\psi^{[\text{TS},t]})/p_{\text{CBC}}(\theta_i^{\text{prop}})}{p(\theta_i^{[\text{TS},t-1]}|\psi^{[\text{TS},t]})/p_{\text{CBC}}(\theta_i^{[\text{TS},t-1]})}, \end{aligned} \quad (2.8)$$

Stage 2d: For $i = 1, \dots, n$, set

$$(\theta_i, \xi_i)^{[\text{TS},t]} = \begin{cases} (\theta_i, \xi_i)^{\text{prop}} & \text{with probability } \min\{1, R_i\}, \\ (\theta_i, \xi_i)^{[\text{TS},t-1]} & \text{otherwise.} \end{cases}$$

The MH step (**Stage 2c** and **Stage 2d**) is the key to whether the two-stage algorithm works or not. In many practical problems, hierarchical models produce estimates $p_{\text{H}}(\theta_i|\mathbf{y})$ that shrink inwards, so generally the proposal distribution $p_{\text{CBC}}(\theta_i, \xi_i | y_i)$ in the MH step is over-dispersed relative to the target distribution $p_{\text{H}}(\theta_i, \xi_i | \psi, y_i)$. This justifies the effectiveness of our two-stage algorithms to a wide range of problems. Chapters 4 and 5 present the application of the two-stage algorithm to the hierarchical models for complicated astrophysical problems.

2.3 Empirical Bayes Approach

The EB approach fits the hierarchical model in Eq. (2.1) by first finding the MAP estimate of ψ , i.e., $\hat{\psi}^{\text{MAP}}$ and then inferring each θ_i, ξ_i via a case-by-case analysis using the prior $p(\theta_i | \hat{\psi}^{\text{MAP}})$. Computing $\hat{\psi}^{\text{MAP}}$, i.e.,

$$\hat{\psi}^{\text{MAP}} = \arg \max_{\psi} \int p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi | \mathbf{y}) d\boldsymbol{\theta} d\boldsymbol{\xi}, \quad (2.9)$$

is a challenge owing to the potentially high-dimensional integral. EM algorithm (e.g. Dempster et al., 1977; van Dyk and Meng, 2010) is often employed to find $\hat{\psi}^{\text{MAP}}$. Basically, it consists of two steps: E-step (Expectation) and M-step (Maximisation), and it proceeds in the following.

EM algorithm

Step 0: Initialize $\psi^{[0]}$.

Step 1: For $t = 1, \dots, T$

E-Step:

$$\begin{aligned} Q(\psi | \psi^{[t-1]}) &= \mathbb{E} \left[\log p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi | \mathbf{y}) | \psi^{[t-1]}, \mathbf{y} \right] \\ &= \int \log p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi | \mathbf{y}) p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \psi^{[t-1]}, \mathbf{y}) d\boldsymbol{\theta} d\boldsymbol{\xi}, \end{aligned} \quad (2.10)$$

where

$$p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \psi^{[t-1]}, \mathbf{y}) = \frac{p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi^{[t-1]} | \mathbf{y})}{\int p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi}, \psi^{[t-1]} | \mathbf{y}) d\boldsymbol{\theta} d\boldsymbol{\xi}}.$$

M-step:

$$\psi^{[t]} = \arg \max_{\psi} Q(\psi | \psi^{[t-1]}).$$

Step 2: Repeat E-step and M-step until $\psi^{[t]}, t = 1, 2, \dots, T$ converges.

EM algorithm is quite useful to a wide range of problems, but when the likelihood function $p(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\xi})$ is complicated or does not have a closed form, EM is not helpful. Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) provides an effective remedy, which replaces the **E-Step** by Monte Carlo integration. Empirical Bayes Gibbs Sampler (EBGS) (Casella, 2001) is one kind of MCEM algorithm to fit the hierarchical model in Eq. (2.1) via EB and proceeds as follows.

EBGS algorithm

Step 0: Initialise $\psi^{[\text{EBGS}, 0]}$ ¹.

Step 1: For $t = 1, \dots, T$

Step 1a: Given $\psi^{[\text{EBGS}, t-1]}$, generate a sample $(\boldsymbol{\theta}^{[s]}, \boldsymbol{\xi}^{[s]}), s = 1, \dots, S$ from the conditional posterior distribution $p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \psi^{[\text{EBGS}, t-1]}, \mathbf{y})$ via a Gibbs sampler.

Step 1b: Replace $Q(\psi | \psi^{[\text{EBGS}, t-1]})$ in Eq. (2.10) by a Monte Carlo approximation (with subscript “MC”),

$$Q_{\text{MC}}(\psi | \psi^{[\text{EBGS}, t-1]}) = \frac{1}{S} \sum_{s=1}^S \log p_{\text{H}}(\psi, \boldsymbol{\theta}^{[s]}, \boldsymbol{\xi}^{[s]} | \mathbf{y}). \quad (2.11)$$

¹We use “EBGS” to mark quantities from EBGS algorithm.

Step 1c:

$$\psi^{[\text{EBGS},t]} = \arg \max_{\psi} Q_{\text{MC}}(\psi | \psi^{[\text{EBGS},t-1]}).$$

Step 2: Upon convergence of $\psi^{[\text{EBGS},t]}$ to $\hat{\psi}^{\text{MAP}}$, produce a final sample from $p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \mathbf{y}, \hat{\psi}^{\text{MAP}})$.

We can also derive a two-stage algorithm that uses only the sample obtained under the case-by-case analysis in Eq. (2.2) and implements the EB approach. To do this, we can approximate the target log marginal posterior distribution in Eq. (2.10) via Importance Sampling (IS) or Sampling Importance Resampling (SIR) using the case-by-case output. Rewriting Eq. (2.4), we have

$$\begin{aligned} \frac{p_{\text{H}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \psi, \mathbf{y})}{p_{\text{CBC}}(\boldsymbol{\theta}, \boldsymbol{\xi} | \mathbf{y})} &\propto \frac{p_{\text{H}}(\boldsymbol{\theta} | \psi)}{p_{\text{CBC}}(\boldsymbol{\theta})} \\ &\propto \prod_{i=1}^n \frac{p_{\text{H}}(\theta_i | \psi)}{p_{\text{CBC}}(\theta_i)}. \end{aligned} \quad (2.12)$$

The two-stage algorithm for EB proceeds as follows.

Two-stage algorithm for EB

Step 0: Obtain a MCMC sample of $\boldsymbol{\theta}^{[\text{CBC},\ell]}, \boldsymbol{\xi}^{[\text{CBC},\ell]}, \ell = 1, \dots, L$ from the case-by-case analyses. Appropriate burn-in and convergence checks are required to get a reasonable chain.

Initialization Set $t = 0$ and initialise $\psi^{[\text{EBTS},0]}$ ².

Step 1: IS For $t = 1, \dots, T$,

Step 1a: Compute the unnormalised importance weights $\tilde{w}_{\ell} = \frac{p_{\text{H}}(\boldsymbol{\theta}^{[\text{CBC},\ell]} | \psi^{[\text{EBTS},t-1]})}{p_{\text{CBC}}(\boldsymbol{\theta}^{[\text{CBC},\ell]})}$, for $\ell = 1, \dots, L$ and normalise them via $w_{\ell} = \tilde{w}_{\ell} / \sum_{j=1}^L \tilde{w}_j$.

²We use “EBTS” to mark the EB two-stage algorithm.

Step 1b: Approximate $Q(\psi|\psi^{[\text{EBTS},t-1]})$ in Eq. (2.10) by

$$Q_{\text{EBTS}}(\psi|\psi^{[\text{EBTS},t-1]}) = \sum_{\ell=1}^L \log p_{\text{H}}(\psi, \boldsymbol{\theta}^{[\text{CBC},\ell]}, \boldsymbol{\xi}^{[\text{CBC},\ell]}|\mathbf{y}) \cdot w_{\ell}.$$

Step 1c: Update ψ by

$$\psi^{[\text{EBTS},t]} = \arg \max_{\psi} Q_{\text{EBTS}}(\psi|\psi^{[\text{EBTS},t-1]}).$$

Step 1': SIR For $t = 1, \dots, T$,

Step 1a': For each i , obtain a sample of size S by resampling $(\theta_i^{[\text{CBC},\ell]}, \xi_i^{[\text{CBC},\ell]})$

with probabilities proportional to $\tilde{w}_{[i,\ell]} = \frac{p(\theta_i^{[\text{CBC},\ell]}|\psi^{[\text{EBTS},t-1]})}{p_{\text{CBC}}(\theta_i^{[\text{CBC},\ell]})}$, $\ell = 1, \dots, L$.

Denote the resample by $(\boldsymbol{\theta}^{[\text{SIR},s]}, \boldsymbol{\xi}^{[\text{SIR},s]}), s = 1, \dots, S$.

Step 1b': Approximate $Q(\psi|\psi^{[\text{EBTS},t-1]})$ in Eq. (2.10) by

$$Q(\psi|\psi^{[\text{SIR},t-1]}) = \frac{1}{S} \sum_{s=1}^S \log p_{\text{H}}(\psi, \boldsymbol{\theta}^{[\text{SIR},s]}, \boldsymbol{\xi}^{[\text{SIR},s]}|\mathbf{y}). \quad (2.13)$$

Step 1c': Update ψ via

$$\psi^{[\text{EBTS},t]} = \arg \max_{\psi} Q(\psi|\psi^{[\text{SIR},t-1]}).$$

Step 2: Upon convergence of $\psi^{[\text{EBTS},t]}$ to $\hat{\psi}^{\text{MAP}}$, produce a final sample from $p(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}, \hat{\psi}^{\text{MAP}})$.

Either **Step 1** or **Step 1'** can be used, provided a sufficient large sample from the case-by-case analyses are obtained. Both the IS and SIR version of the two-stage algorithm work efficiently. However, they perform badly when the samples under case-by-case analyses fail to represent the full distribution. We recommend readers keeping alert while implementing the two-stage algorithms for EB. One approach to

check whether the algorithms work or not is to select multiple case-by-case samples and use them to run the two-stage algorithms and see whether point estimates from different case-by-case samples converge to the same values.

2.4 Prior Distributions

The focus of this thesis is to devise efficient algorithms to compute Bayesian hierarchical models when the likelihood function is complicated, and to apply new algorithms to tackle problems in Astronomy. Here, however, we summarise the general theory on prior distributions and justify the priors we use in this thesis.

A prior distribution, often simply prior, of a parameter is the probability distribution that quantifies one's beliefs about this parameter before some evidence is taken into account. For example, in a coin toss experiment, if one believes that the coin is fair, then the prior probability of it landing a head is a half.

Broadly speaking, there are three types of priors: subjective, objective and informative, and non-informative. A subjective prior expresses the experimenter's personal belief, which usually makes the research less objective. See Chapter 3 of Berger (2013) for a detailed discussion about subjective priors. The second type is objective and informative prior distributions that can be formulated by using historical data or published research results. For instance, in stellar evolution many research groups have attempted to infer the properties of a certain star cluster such as Hyades. These scientific results can be used as priors for the upcoming studies. To put it in another way, today's posterior is tomorrow's prior. In this research, we sometimes use objective and informative priors when we analyse astrophysical data.

The third kind, non-informative priors, receive most attention and many methods are developed to obtain them from different perspectives in the statistical literature. Non-informative priors are applied when not much information about parameters are

known beyond the data included in the model. Later we review a few common ways to obtain non-informative priors in Section 2.4.1.

In this research, all statistical methods are in Bayesian framework, so the selection of prior distributions is an integral part. This section consists of two parts: (i) the general ways to create prior distributions and (ii) specifically the prior distributions for variance parameters in hierarchical models.

2.4.1 General Methods to Create Non-informative Priors

In this section, we cover general basics to obtain non-informative priors. Other terms for a non-informative prior are reference prior, diffuse prior, objective prior and vague prior. Because no prior really contains no information, a non-informative prior can also be called weakly informative prior.

In general, a non-informative prior is one which is dominated by the likelihood function. Various criteria of non-informative priors have been proposed, such as, invariance (Jeffreys, 1961), maximum entropy (Jaynes, 1983), reference priors (Bernardo, 1979) and agreement with classical estimators (Tiao and Box, 1973; Meng and Zaslavsky, 2002). We discuss these principles in more details later.

The Jeffreys prior is invariant to reparametrization. Suppose that θ is the parameter of interest and its likelihood function is $p(Y|\theta)$. The Fisher information is

$$I(\theta) = -E_Y \left[\frac{d^2}{d\theta^2} \log p(Y|\theta) \right] = E_Y \left[\left(\frac{d \log p(Y|\theta)}{d\theta} \right)^2 \right].$$

Jeffreys prior for θ is defined as $p(\theta) \propto I(\theta)^{1/2}$. For any one-to-one and differentiable transformation $\theta = h(\phi)$, we have

$$I(\phi) = I(\theta) \left| \frac{d\theta}{d\phi} \right|^2.$$

If one reparametrizes the original model in terms of ϕ , then Jeffreys prior of ϕ will be

$$I(\phi)^{1/2} = I(\theta)^{1/2} \left| \frac{d\theta}{d\phi} \right|.$$

This is an informal justification of the invariance of Jeffreys priors to reparametrizations.

Another idea, proposed by Jaynes (1983), is the principle of maximum entropy. The foundation lies that the Shannon entropy

$$H(X) = E[-\log(p(X))]$$

of a probability distribution $p(x)$ measures the amount of information in the distribution $p(x)$. The larger the entropy, the less information is provided by the distribution. Thus, one can compute the entropies of a set of distributions and compare them, choosing the distribution with largest entropy as a non-informative prior. Usually uniform prior has the maximum entropy.

Bernardo (1979) develops the reference prior, which is to maximise the expected Kullback-Leibler divergence of the posterior distribution relative to the prior, i.e.,

$$p(\theta) = \arg \max_{p(\theta)} D_{\text{KL}}(p(\theta) || p(\theta|X)),$$

where $p(\theta)$ and $p(\theta|X)$ are the prior and posterior of θ , and $D_{\text{KL}}(\cdot || \cdot)$ is the Kullback-Leibler divergence, defined as

$$D_{\text{KL}}(p(\theta) || p(\theta|X)) = \int_{-\infty}^{\infty} p(\theta) \log \frac{p(\theta)}{p(\theta|X)}.$$

As to the pros and cons of different non-informative prior distributions, refer to

Kass and Wasserman (1996) for such discussion. Practical problems related to non-informative prior distributions include the requirement that the posterior distribution should be proper. Proper prior distributions always produce proper posterior, whereas improper priors may lead to improper posteriors. Many non-informative priors tend to be improper, so it is necessary to choose well-founded priors or to check whether the posterior distribution is proper.

2.4.2 Priors for Variance Parameters in Hierarchical Models

Here we focus on the priors for variance parameters in hierarchical models, which are directly relevant to research projects in this thesis. Because we always use a Gaussian model for the population, i.e., the Gaussian assumption of μ_i in Eq. 2.15, we concentrate on the prior distributions for τ or τ^2 or the population variance-covariance matrix.

Many studies on non-informative prior distributions have been done for the variance or standard deviation in hierarchical models (Daniels, 1999; Daniels and Kass, 1999; Barnard et al., 2000; Daniels and Kass, 2001; Gelman et al., 2005; Gelman, 2006a; Gustafson et al., 2006; Huang and Wand, 2013; Schuurman et al., 2016). Gelman (2006a) presents a clear summary of common choices of prior distributions for scale parameters in the hierarchical model and proposes a half-t family of conditionally conjugate priors for standard deviation τ , i.e.,

$$p(\tau) \propto \left(1 + \frac{1}{\nu} \left(\frac{\tau}{A}\right)^2\right)^{-(\nu+1)/2},$$

where ν and A are tuning parameters. Large but finite values of A yield weakly informative prior of τ . When $\nu = 1$ or $A \rightarrow \infty$, it degenerates to the uniform prior on τ . When $\nu = 1$, it half-t distribution becomes the proper half-Cauchy,

$$p(\tau) \propto (\tau^2 + A)^{-1}.$$

According to Gelman (2006a), when the number of objects I in the hierarchical model is greater than or equal to 5, the uniform prior on τ , $p(\tau) \propto 1$, is non-informative and generally will work. If $I < 5$, the uniform prior is likely to overestimate τ and leads to less shrinkage in the estimation of object-level parameters μ_i . If a proper prior on τ is required, approximating the uniform prior on τ by a uniform on a wide range is recommended.

The uniform prior on $\log \tau$, $p(\log \tau) \propto 1$, yields improper posterior distribution. So we cannot choose it as a prior. The conjugate prior for τ^2 , the inverse gamma distribution $\Gamma^{-1}(\epsilon, \epsilon)$ with small positive number ϵ to be non-informative, is easy to compute. The posterior distribution, however, is sensitive to the parameter ϵ in it (Gelman, 2006a; Browne et al., 2006). So inverse gamma is not suggested as an objective prior.

When the second level of the hierarchical model is multivariate Gaussian as in the IFMRs project in Chapter 5, we have to put a prior distribution on the variance-covariance matrix, $\mathbf{\Gamma}$. Huang and Wand (2013) generalises the half-t prior distribution family in Gelman (2006a) to multivariate version, i.e.,

$$\mathbf{\Gamma} \mid \lambda_1, \lambda_2 \sim \text{Inverse Wishart} \left(2\nu \begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{pmatrix}, \nu + 1 \right),$$

where Inverse Wishart ³ is prior distribution for the variance parameter $\mathbf{\Gamma}$, $\lambda_1, \lambda_2 \sim$

³ Inverse Wishart distribution is the conjugate prior for the covariance matrix of a multivariate normal distribution, having two parameters: scale matrix $\mathbf{\Psi}$ and degrees of freedom ν , and its probability density function is

$$p(\mathbf{X} \mid \mathbf{\Psi}, \nu) = \frac{|\mathbf{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{p\nu}{2}} \Gamma_p(\frac{\nu}{2})} |\mathbf{X}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2} \text{tr}(\mathbf{\Psi} \mathbf{X}^{-1})},$$

in which Γ_p is the multivariate gamma function, and $\text{tr}(\cdot)$ is the trace function.

Inverse Gamma($1/2, 1/\Lambda$)⁴. Huang and Wand (2013) suggests that $\nu = 2$ leads to marginal uniform distribution for correlations and arbitrarily large positive Λ leads to arbitrarily weakly informative prior distributions for standard deviations.

Though inverse Wishart distribution is conjugate to multivariate normal, it is informative when variances in scale matrix are small, so it is not recommended to use in the hierarchical analysis. For more details about non-informative priors for variance-covariance matrix, refer to Barnard et al. (2000); Chung et al. (2015); Schuurman et al. (2016).

In conclusion, for the one-dimensional scale parameters in the second level of hierarchical models such as τ in Eq. 2.15, the uniform $p(\tau) \propto 1$ is a well justified and non-informative prior distribution when the number of objects I is no less than 5. When $I < 5$, we follow the suggestion of Gelman (2006a), take a half-t prior distribution for τ . As to the multivariate case, i.e., the variance-covariance matrix $\mathbf{\Gamma}$ in Chapter 5, the hierarchical prior distribution proposed by Huang and Wand (2013) is a well-studied and non-informative prior. Because this thesis concentrates on efficient computing algorithms for fitting intricate Bayesian hierarchical models, we just choose well-founded non-informative priors for the scale parameters in the second level of hierarchical models.

⁴Inverse Gamma is the reciprocal of of Gamma distribution, with two parameters: shape α and rate β , and its density function is

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right),$$

where Γ is the Gamma function.

2.5 Numerical Validation of Gaussian-Gaussian Hierarchical Model

In this section, we perform a simulation study to demonstrate various approaches for fitting a Gaussian-Gaussian hierarchical model in Eq. (2.14)–(2.15). In a Gaussian-Gaussian model,

$$Y_i \sim N(\mu_i, \sigma^2); \quad (2.14)$$

$$\mu_i \sim N(\gamma, \tau^2), i = 1, 2, \dots, I, \quad (2.15)$$

$Y_i, i = 1, 2, \dots, I$ are the observed data, $\boldsymbol{\mu} = \{\mu_i, i = 1, 2, \dots, I\}$ are object-level parameters, and γ , and τ^2 are population-level parameters. σ^2 is a preset and known variance parameter. The aim of this simulation study is twofold, first to verify that the two-stage sampler works effectively, second to compare the MSE of estimates from different approaches.

2.5.1 Data Simulation

We simulate data from the Gaussian-Gaussian model in Eq.s (2.14)–(2.15). Without loss of generality, set the number of objects $I = 10$, population-level parameters $\gamma_{\text{True}} = 0.5$ and $\tau_{\text{True}} = 0.2$. Generate true values of object-level parameters $\mu_{\text{True},i}, i = 1, \dots, I$ via $\mu_{\text{True},i} \sim N(\gamma_{\text{True}}, \tau_{\text{True}}^2)$. To investigate how the ratio σ^2/τ^2 affects mean squared errors (MSEs) of different estimators, we vary the value of σ^2 , such that $\sigma_k^2 = \tau_{\text{True}}^2 \cdot 10^{r_k}$, in which $r_k = \log_{10} \frac{\sigma_k^2}{\tau_{\text{True}}^2}$, $k = 1, \dots, K$, taking an equally spaced sequence from -3 to 3 of length $K = 20$. Subsequently, for each k we simulate the

Table 2.2: Notations in the simulation study

Type	Parameters	True values in simulation
Population-level Parameters	γ, τ	$\gamma_{\text{True}} = 0.5, \tau_{\text{True}} = 0.2$
Object-level Parameters	$\mu_i, i = 1, \dots, I$	$\mu_{\text{True},i} \sim N(\gamma_{\text{True}}, \tau_{\text{True}}^2)$
Data	$Y_i, i = 1, \dots, I$	$Y_i \sim N(\mu_{\text{True},i}, \sigma_k^2)$

observed data for each object

$$Y_i \sim N(\mu_{\text{True},i}, \sigma_k^2), i = 1, \dots, I.$$

Table 2.2 presents various notations for parameters and data in the simulation study

After simulating data $Y_i, i = 1, \dots, I$, we employ four algorithms to obtain point estimates of object-level parameters $\mu_i, i = 1, \dots, I$ and population-level parameters γ and τ^2 . For each σ_k^2 , we repeat the whole process – data generation and estimating parameters – 50 times to compute the MSEs and coverage probabilities of estimates from different algorithms.

The the following subsections cover the four algorithms of fitting the hierarchical model. These four algorithms are the case-by-case, fully Bayesian with Gibbs sampler, fully Bayesian with two-stage sampler, and the empirical Bayes approach.

2.5.2 Case-by-case Approach

The case-by-case approach does not pool multiple objects into one single model in Eq. (2.14) and it infers object-level parameters individually by using one unit of data. It employs the first level in the hierarchical model, i.e., Eq. (2.14), to fit each object-level parameter μ_i with data in the i -th group, i.e., Y_i . And σ is assumed to be known. So the maximum likelihood estimate (MLE) for μ_i is

$$\hat{\mu}_i^{\text{ML}} = Y_i,$$

and the variance of this estimate is σ^2 . In the Bayesian perspective, the posterior distribution of μ_i from the case-by-case analysis is $p(\mu_i|Y_i) = N(\mu_i|Y_i, \sigma^2)$.

The case-by-case approach does not use the second level model in Eq. (2.15), because it infers object-level parameters μ_i based on the i -th unit of data Y_i only. Therefore, it fails to yield estimates of the population-level parameters γ and τ .

2.5.3 Fully Bayesian Approach

Two algorithms are employed to fit the hierarchical model in Eq.s (2.14)–(2.15) in the fully Bayesian approach. Because a prior distribution on γ and τ is required, we always take $p(\gamma, \tau) \propto p(\gamma|\tau)p(\tau)$ in this section and $p(\gamma|\tau) \propto 1$. As discussed in Section 2.4, we assign a uniform prior on the standard deviation τ , namely, $p(\tau) \propto 1$, or equivalently, $p(\tau^2) \propto (\tau^2)^{-\frac{1}{2}}$. Therefore, the joint posterior density for all parameters is

$$p(\gamma, \tau^2, \mu_1, \dots, \mu_I | \mathbf{Y}) \propto p(\tau^2) \prod_i^I p(\mu_i | \gamma, \tau^2) p(Y_i | \mu_i). \quad (2.16)$$

We use Gibbs sampler to fit model in Eq. (2.14)–(2.15), and it proceeds as follows.

Step 1: Initialise $\boldsymbol{\mu}^{(0)} = \mathbf{Y} = (Y_1, \dots, Y_I)$;

Step 2: For $t = 1, \dots, T$, sample

Step 2a: $\tau^{2(t)} \sim \text{Inverse Gamma}(\frac{I}{2}-1, \sum_{i=1}^I (\mu_i^{(t-1)} - \bar{\mu}^{(t-1)})^2/2)$, where $\bar{\mu}^{(t-1)} = \frac{1}{I} \sum_{i=1}^I \mu_i^{(t-1)}$, in which Inverse Gamma(α, β) stands for the inverse gamma distribution with probability density function $p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$.

Step 2b: $\gamma^{(t)} \sim N(\bar{\mu}^{(t-1)}, \tau^{2(t)}/I)$;⁵

Step 2c: $\mu_i^{(t)} \sim N\left(\frac{\tau^{2(t)}Y_i + \sigma^2\gamma^{(t)}}{\tau^{2(t)} + \sigma^2}, \frac{\tau^{2(t)}\sigma^2}{\tau^{2(t)} + \sigma^2}\right)$, for $i = 1, \dots, I$.

⁵Here the first and second parameters in normal distributions are the mean and variance.

The two-stage sampler can also be employed to sample from the joint posterior density in Eq. (2.16). Details of how it proceeds are presented in the next section. When point estimates of parameters are required, we use the posterior sample means.

2.5.4 Two-stage Sampler for Fully Bayesian

The two-stage algorithm consists of two parts. The first stage draws samples for each μ_i separately under the case-by-case analyses. From the Bayesian perspective, the posterior distribution of μ_i from the case-by-case analysis is $p_{\text{CBC}}(\mu_i|Y_i) = N(\mu_i|Y_i, \sigma^2)^6$. The second stage uses a Metropolis-Hastings (MH) within Gibbs sampler. Given $\boldsymbol{\mu}$, it updates γ and τ^2 just as in the Gibbs sampler, but given γ and τ^2 it updates $\boldsymbol{\mu}$ via a Metropolis-Hastings independence sampler (MHIS) that randomly selects one draw from the case-by-case samples as the proposal and accepts it according to the regular MH acceptance probability.

Stage 1: For each i obtain a sufficient MCMC sample under the case-by-case analysis, $\mu_i^{[\text{CBC}, t]} \sim N(Y_i, \sigma^2)$ for $t = 1, \dots, T$.

Stage 2 Initialization: For each i , set $\mu_i^{[\text{TS}, 0]}$ to a randomly selected value from its case-by-case samples and we use superscripts “TS” and “CBC” to mark samples from the two-stage algorithm and the case-by-case fits, respectively.

Stage 2 Iterations: For $t = 1, \dots, T$,

Stage 2a: Sample $\tau^{2[\text{TS}, t]} \sim \text{Inverse Gamma}(\frac{I}{2}-1, \sum_{i=1}^I (\mu_i^{[\text{TS}, t-1]} - \bar{\mu}^{[\text{TS}, t-1]})^2 / 2)$
with $\bar{\mu}^{[\text{TS}, t-1]} = \frac{1}{I} \sum_{i=1}^I \mu_i^{[\text{TS}, t-1]}$.

Stage 2b: Sample $\gamma^{[\text{TS}, t]} \sim N(\bar{\mu}^{[\text{TS}, t-1]}, \tau^{2[\text{TS}, t]} / I)$;

⁶We use $N(x|a, \sigma^2)$ to denote the normal density function with mean a and variance σ^2 .

Stage 2c: Given $\gamma^{[\text{TS},t]}$ and $\tau^{2[\text{TS},t]}$, use MHIS to update each μ_i from its conditional distribution

$$\begin{aligned} p_H(\mu_i | \gamma^{[\text{TS},t]}, \tau^{2[\text{TS},t]}, Y_i) &\propto p(\mu_i | \gamma^{[\text{TS},t]}, \tau^{2[\text{TS},t]}) p(Y_i | \mu_i) \\ &\propto p(\mu_i | \gamma^{[\text{TS},t]}, \tau^{2[\text{TS},t]}) p_{\text{CBC}}(\mu_i | Y_i). \end{aligned}$$

For each i , we use the posterior distribution under the case-by-case analysis, $p_{\text{CBC}}(\mu_i | Y_i)$ as the proposal distribution. This simply involves randomly selecting μ_i^* from the case-by-case MCMC sample obtained in **Stage 1**.

Stage 2d: Compute the I MH acceptance probabilities,

$$\begin{aligned} R_i &= \frac{p(\mu_i^* | Y_i, \gamma^{[\text{TS},t]}, \tau^{2[\text{TS},t]}) p_{\text{CBC}}(\mu_i^{[\text{TS},t-1]} | Y_i)}{p(\mu_i^{[\text{TS},t-1]} | Y_i, \gamma^{[\text{TS},t]}, \tau^{2[\text{TS},t]}) p_{\text{CBC}}(\mu_i^* | Y_i)} \\ &= \frac{N(\mu_i^* | \gamma^{[\text{TS},t]}, \tau^{2[\text{TS},t]})}{N(\mu_i^{[\text{TS},t-1]} | \gamma^{[\text{TS},t]}, \tau^{2[\text{TS},t]})}; \end{aligned}$$

Stage 2e: For $i = 1, \dots, n$, independently set

$$\mu_i^{[\text{TS},t]} = \begin{cases} \mu_i^* & \text{with probability } \min(1, R_i) \\ \mu_i^{[\text{TS},t-1]} & \text{otherwise.} \end{cases}$$

2.5.5 Empirical Bayes Approach

To fit the Gaussian-Gaussian hierarchical model in Eq.s (2.14)–(2.15) via the empirical Bayes approach, we find the maximum a posterior (MAP) estimates of γ , and τ , i.e., $\hat{\gamma}^{\text{MAP}}$ and $\hat{\tau}^{\text{MAP}}$, then we use $N(\hat{\gamma}^{\text{MAP}}, \hat{\tau}^{\text{MAP}})$ as a prior distribution and derive posterior distributions for μ_i . As in the fully Bayesian analysis we take the uniform prior for τ , i.e., $p(\tau) \propto 1$. When obtaining the MAP estimates of γ and τ , we use the

marginal distribution of Y_i , i.e., $Y_i \sim N(\gamma, \tau^2 + \sigma^2)$. Then we can find the closed-form solution for $\hat{\gamma}^{\text{MAP}} = \bar{Y}$ and $\hat{\tau}^{\text{MAP}} = \sqrt{\max\{\frac{1}{I} \sum_i (Y_i - \bar{Y})^2 - \sigma^2, 0\}}$. Actually the MAP estimates of γ and τ coincide with MLE of them due to the use of uniform prior on τ .

2.5.6 Result of Simulation Study

This section reports the findings from the simulation study. In this thesis, we compare and evaluate estimators from different methods or algorithms via two criteria: root mean squared error (RMSE) of point estimates and nominal versus actual coverage property of interval estimates in repeated sampling. The coverage of an interval is a frequency property. By definition an interval that covers the true value of the parameter, say, 68.3% of the time on repeated simulation/sampling of the data is a 68.3% confidence interval (Neyman, 1937). Of course, a Bayesian credible interval with probability 68.3% may not cover the parameter 68.3% of the time on repeated samples (Gelman, 2008). If it does, it is indeed a 68.3% confidence interval. A few researchers attempted to study prior distributions that leads to the agreement of credible intervals and their frequentist coverage probabilities. For more details, see Datta and Ghosh (1995); Datta (1996); Levine and Casella (2003).

The four methods in this section – case-by-case, EB, FB-Gibbs, and FB with TS sampler – all produce MCMC samples for the object-level parameters $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_I\}$. For short, we denote FB with Gibbs sampler and TS sampler as FB-Gibbs, FB-TS, respectively. The case-by-case, approach does not yield estimates of population-level parameters γ and τ . EB leads to point (MAP) estimates of γ and τ , while both FB methods provide MCMC samples of γ and τ . When computing the RMSEs of point estimates, we utilise posterior means from MCMC samples, and when computing coverage of interval estimates we use the 68.3% credible intervals from posterior

distributions by finding the 15.9% and 84.1% quantiles from the MCMC samples. For EB, we only have point estimates of γ and τ , no interval estimates. We make plots to show the RMSEs and tabulate RMSEs and coverage probabilities of estimates under all four methods.

Here Fig. 2.1 shows the performance of four different approaches in terms of RMSEs. Fig. 2.1 consists of three plots, the top-left showing RMSEs of four estimators of $\boldsymbol{\mu}$ in log scale versus $\log_{10}(\sigma^2/\tau^2)$, the top-right displaying RMSEs of γ from four methods in log scale versus $\log_{10}(\sigma^2/\tau^2)$, and the bottom-left exhibiting RMSEs of τ^2 in log scale versus $\log_{10}(\sigma^2/\tau^2)$. In these plots, the blue solid, red dashed, black dotted and purple dash-dotted lines represent outcomes from the case-by-case, EB, FB-Gibbs and FB-TS, respectively. Throughout these three plots, the black dotted (FB with Gibbs sampler) and purple dash-dotted (FB with TS sampler) lines always coincide, although there are very small divergences in RMSE of γ due to Monte Carlo errors. The consistency of their performances implies that our new TS sampler works nicely for the Gaussian-Gaussian hierarchical model.

The log RMSEs of both the object-level parameter $\boldsymbol{\mu}$ and the population-level parameters γ and τ increase as σ^2/τ^2 grows. For γ , there are no differences between the EB and FB methods. However, estimates of τ^2 produced by FB methods (Gibbs and TS samplers) have larger MSEs than EB regardless of σ^2/τ^2 . This can be explained by the fact that the marginal posterior distribution of τ^2 has a right long tail because the true value of $\tau_{\text{True}}^2 = 0.04$ is very close to 0 and the parameter space of τ^2 is asymmetric and restricted to be positive (Gelman, 2006a). So when we compute the posterior mean of τ^2 under the FB, it would be greater than the posterior mode, which is the estimate of τ^2 under the EB. The rise of RMSEs of population-level parameters γ and τ with the increase of variance ratio σ^2/τ^2 , means that both EB and FB will experience difficulty in accurately estimating γ and τ when σ^2/τ^2 is huge (say, over

100). Because we use the prior $p(\tau) \propto 1$, so the EB estimates of γ and τ are the same as MLEs of them, which further implies that the unbiased MLE will also have difficulty (Browne et al., 2006).

As to the object-level parameters, of our most concern, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_I)$, the log scale RMSEs from all four approaches increase almost linearly with $\log_{10}(\sigma^2/\tau^2)$. In other words, RMSEs from four methods and σ^2/τ^2 have a power-law relationship with positive exponents. There is no difference between the case-by-case and the rest three hierarchical methods when $\log_{10}(\sigma^2/\tau^2) \leq -1$, i.e., $\sigma^2 \leq \tau^2/10$. When $\log_{10}(\sigma^2/\tau^2) > -1$, however, the case-by-case RMSE increases with significantly larger exponent than the hierarchical methods. The EB has slightly smaller increase rate than the FB methods. This is mainly caused by the overestimation of τ under FB approaches, which is the consequence of long-tailed posterior distribution of τ^2 .

Table 2.3 presents the RMSEs of estimates of both the object-level parameter $\boldsymbol{\mu}$ and the population-level parameters γ and τ based on four methods: EB, FB-Gibbs, FB-TS, and case-by-case approach. Just as shown in the Fig. 2.1 above, EB estimates of $\boldsymbol{\mu}$ have the smallest RMSEs and case-by-case estimates have the largest RMSEs. FB estimates of $\boldsymbol{\mu}$ have slightly larger RMSEs than those of EB. The key findings in this table are 1.) as σ^2/τ^2 increases, the ratio of case-by-case RSMES of $\boldsymbol{\mu}$ to those from hierarchical fits (EB and FB) monotonically increases; 2.) when $\sigma^2/\tau^2 = 12.743$, hierarchical (EB and FB) RMSEs (about 1.0) is nearly half of that from case-by-case fits (around 2.2).

Table 2.4 displays the actual coverage probabilities (CP) and their 68.3% confidence intervals (CI) for credible intervals with nominal coverage probability 68.3% for the object-level parameters $\boldsymbol{\mu}$ from aforementioned four methods under a few settings. The 68.3% CIs of coverage probabilities in this section are obtained via Clopper-Pearson method (Clopper and Pearson, 1934). From this table, the case-by-

Table 2.3: The RMSE of the various estimates of μ , γ and τ under numerous simulated settings. For μ , its RMSE is the root of average MSE of $\mu_1, \dots, \mu_I$.

Simulation Settings	EB			FB-Gibbs			FB-TS			Case-by-case		
	μ	γ	τ	μ	γ	τ	μ	γ	τ	μ	γ	τ
$\sigma^2/\tau^2 = 0.001$	0.02	0.047	0.018	0.02	0.047	0.031	0.02	0.048	0.031	0.02	n.a.	n.a.
$\sigma^2/\tau^2 = 0.018$	0.085	0.052	0.024	0.085	0.052	0.044	0.086	0.051	0.043	0.086	n.a.	n.a.
$\sigma^2/\tau^2 = 0.162$	0.247	0.066	0.022	0.247	0.066	0.041	0.247	0.067	0.041	0.26	n.a.	n.a.
$\sigma^2/\tau^2 = 1.438$	0.569	0.103	0.047	0.561	0.104	0.08	0.569	0.109	0.083	0.748	n.a.	n.a.
$\sigma^2/\tau^2 = 12.743$	0.963	0.187	0.169	1.03	0.196	0.374	1.066	0.204	0.383	2.178	n.a.	n.a.

case credible intervals have the best coverage property, always close to the nominal level 68.3%. The FB approaches with Gibbs and TS samplers are consistent in terms of coverage property of their credible intervals, fluctuating around the nominal level 68.3% when $\sigma^2/\tau^2 \leq 1.438$. But when $\sigma^2/\tau^2 = 12.743$, the actual coverage probabilities of FB-led credible intervals (about 80%) significantly exceed their nominal level, 68.3%. This is caused by the overestimation of τ . EB has the worst coverage property, coverage probabilities of credible intervals of μ decreasing with the growth of σ^2/τ^2 . When $\sigma^2/\tau^2 = 12.743$, the actual coverage of EB intervals is about 35%, significantly under the nominal level. The difficulty of hierarchical methods in coverage property is due to two reasons, 1.) the small number of objects, $I = 10$, and 2.) the large variance ratio σ^2/τ^2 . Other research have also reported and discussed this problem, refer to Browne and Goldstein (2002); Browne et al. (2006) for more details.

Table 2.5 displays the actual coverage probabilities (CP) and their 68.3% confidence intervals (CI) for credible intervals with nominal coverage probability 68.3% for the population-level parameters γ and τ from FB approaches. This table does not include the case-by-case and EB methods, because they fail to provide interval estimates of γ and τ . EB can produce interval estimates of γ and τ , but it requires further effort and it is out of the focus of this research. Based on figures in this table, the FB approaches with Gibbs and TS samplers are consistent in terms of actual coverage property of their credible intervals, though there are some differences due to Monte Carlo errors. When $\sigma^2/\tau^2 = 12.743$, the difference of actual coverage probabilities of Gibbs and TS samplers appears, which will vanish if we run the MCMC chain long enough. The actual coverage probabilities of credible interval estimates of γ tend to exceed their nominal levels (68.3%). As discussed above, this is caused by the overestimation of τ^2 due to its asymmetric parameter space. The actual coverage of τ is quite close to its nominal level 68.3% under the first three settings. FB

Table 2.4: The average coverage probabilities (CP) and their 68.3% confidence intervals (CI) of various credible intervals (with nominal coverage probability 68.3%) of $\boldsymbol{\mu}$ (averaged over $\mu_1, \dots, \mu_I$) under numerous simulated settings.

Simulation Settings	EB		FB-Gibbs		FB-TS		Case-by-case	
	CP	68.3% CI	CP	68.3% CI	CP	68.3% CI	CP	68.3% CI
$\sigma^2/\tau^2 = 0.001$	0.682	(0.66, 0.703)	0.692	(0.67, 0.713)	0.682	(0.66, 0.703)	0.69	(0.668, 0.711)
$\sigma^2/\tau^2 = 0.018$	0.644	(0.621, 0.666)	0.648	(0.625, 0.67)	0.652	(0.629, 0.674)	0.662	(0.64, 0.684)
$\sigma^2/\tau^2 = 0.162$	0.658	(0.635, 0.68)	0.694	(0.672, 0.715)	0.682	(0.66, 0.703)	0.718	(0.696, 0.739)
$\sigma^2/\tau^2 = 1.438$	0.468	(0.445, 0.491)	0.658	(0.635, 0.68)	0.664	(0.642, 0.686)	0.688	(0.666, 0.709)
$\sigma^2/\tau^2 = 12.743$	0.348	(0.326, 0.371)	0.816	(0.797, 0.834)	0.786	(0.766, 0.805)	0.684	(0.662, 0.705)

Table 2.5: The actual Coverage Probability (CP) and their 68.3% Confidence Interval (CI) of FB estimates of population-level parameters γ and τ under numerous simulated settings.

Settings	FB-Gibbs		FB-TS	
	CP γ (68.3% CI)	CP τ (68.3% CI)	CP γ (68.3% CI)	CP τ (68.3% CI)
$\sigma^2/\tau^2 = 0.001$	0.88 (0.815, 0.927)	0.68 (0.601, 0.751)	0.88 (0.815, 0.927)	0.68 (0.601, 0.751)
$\sigma^2/\tau^2 = 0.018$	0.84 (0.771, 0.893)	0.66 (0.58, 0.733)	0.82 (0.749, 0.876)	0.64 (0.56, 0.714)
$\sigma^2/\tau^2 = 0.162$	0.66 (0.58, 0.733)	0.66 (0.58, 0.733)	0.64 (0.56, 0.714)	0.66 (0.58, 0.733)
$\sigma^2/\tau^2 = 1.438$	0.66 (0.58, 0.733)	0.76 (0.684, 0.824)	0.66 (0.58, 0.733)	0.76 (0.684, 0.824)
$\sigma^2/\tau^2 = 12.743$	0.86 (0.793, 0.91)	0.84 (0.771, 0.893)	0.8 (0.727, 0.859)	0.7 (0.621, 0.77)

approaches experience difficulty in last two cases, which may be caused by smaller number of objects, $I = 10$, in the hierarchical model. This problem is also discussed in Browne et al. (2006).

Table 2.6 presents the computing times for algorithms in the simulation study. Throughout my whole thesis, all timings are carried out on a Ubuntu linux server that has 64 AMD Opteron 2.5 GHz processors. I wrote code in R programming language (R Core Team, 2017) to undertake computations in my thesis and when parallel computing is required, I used the “doParallel” package. On other computer systems or programming languages, the relative CPU times of different methods should be similar. Because Gaussian-Gaussian hierarchical model is very simple and closed-form solutions are available for the case-by-case and EB methods, so these two methods take very short time. Gibbs and TS samplers cost longer time to simulate samples from the joint posterior distribution. The Gaussian-Gaussian hierarchical model is not computationally challenging and we include it just to compare estimates from

Table 2.6: CPU Times of different methods in this simulation study

Algorithms	Time
Case-by-case fit	About 0.07 seconds for 10000 draws
Hierarchical methods	
EB	About 0.07 seconds for 10000 draws
FB-Gibbs	About 0.44 seconds for 10000 draws
FB-TS	About 0.67 seconds for 10000 draws

hierarchical and case-by-case approaches. In summary, the case-by-case and EB hierarchical fits cost the least time, about 0.07 seconds for 10000 draws. The FB-Gibbs and FB-TS spends longer time, about 0.44 and 0.67 seconds, respectively.

Here is the summary of previous analysis. 1. In terms of RMSEs of object-level parameters $\boldsymbol{\mu}$, the EB and FB point estimates perform similar, having smallest MSE, and the case-by-case method has the largest RMSE. 2. In terms of coverage property, the case-by-case credible intervals perform the best. The FB credible intervals over-covered for $\boldsymbol{\mu}$ and γ due to the overestimation of τ . The EB credible intervals significantly under-covered and are worst of these approaches. 3. FB method has its unique benefits, because it accounts for uncertainties of both object-level and population-level parameters and naturally provides the whole posterior distribution for all parameters. For EB, we can obtain the uncertainty of estimates via Fisher information, but it takes more efforts. The case-by-case estimate does not pool data in a hierarchical model, so fail to provide information of the population. 4. The FB samples from the TS and Gibbs samplers are consistent in terms of RMSEs and coverage property of credible intervals. In total, readers are recommended to employ FB approaches, because its point estimates have smaller RMSEs, its interval estimates have good coverage properties. Also, our two-stage algorithm is consistent with the Gibbs sampler for the FB method, which is further verified by quantile-quantile (QQ) plots in Fig. Fig. 2.2 and 2.3.

Fig. 2.2 and 2.3 display the QQ plots of samples from the Gibbs sampler (y axes) and our two-stage sampler (x axes) for both the object-level (μ) and population-level (γ and τ) parameters. These QQ plots lie quite close to the 45° red line, although there is some deviation from the red line in several plots. These deviations are caused by Monte Carlo errors, accounting for only a tiny part (less than 1%) of the distributions. So we conclude that our two-stage sampler works nicely with the Gaussian-Gaussian hierarchical model.

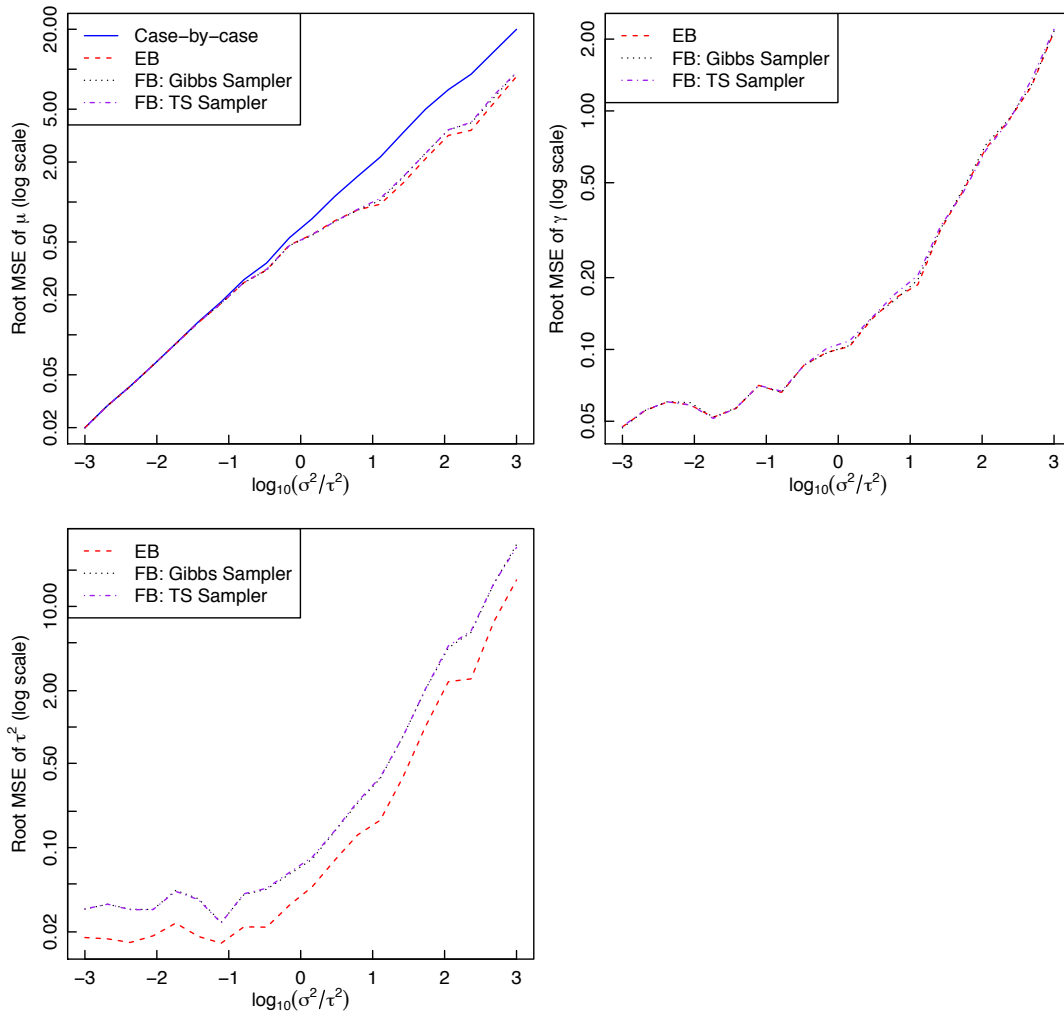


Figure 2.1: Comparisons of the RMSEs for μ, γ and τ from different approaches of fitting the Gaussian-Gaussian model.

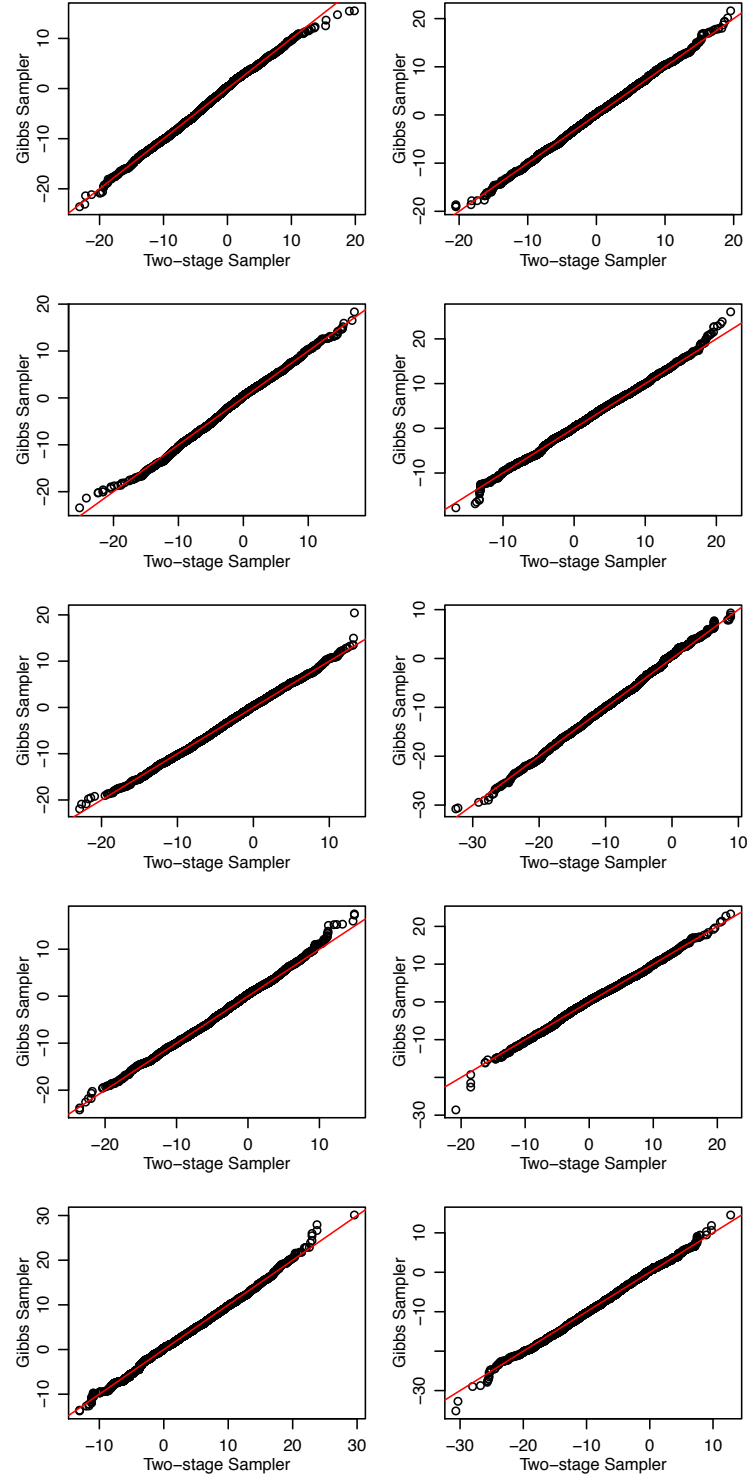


Figure 2.2: Quantile-quantile plots for the two-stage sampler versus Gibbs sampler of $\mu_i, i = 1, 2, \dots, 10$ in the Gaussian-Gaussian model.

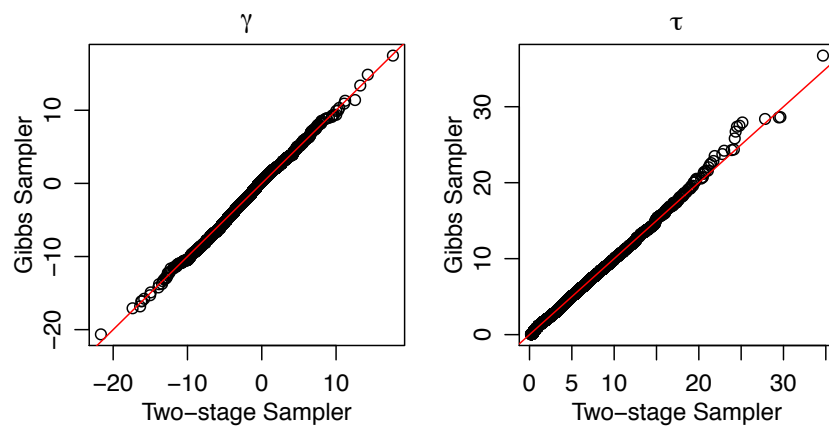


Figure 2.3: Quantile-quantile plots for the two-stage sampler versus Gibbs sampler of γ, τ in the Gaussian-Gaussian model.

3 Distance Modulus to the Large Magellanic Cloud

In astronomy, astronomers use the cosmic distance ladder (also known as the extragalactic distance scale) to determine the distances to celestial objects. The Large Magellanic Cloud (LMC) is a satellite galaxy of the Milky Way, and its distance modulus is important to determining the distance to other celestial objects further from the Milky Way (e.g. Pietrzyński et al., 2013; De Grijs et al., 2014). Many researchers have estimated the distance modulus to the LMC (e.g. Laney and Stobie, 1994; Clementini et al., 2003; Pietrzyński et al., 2013; Inno et al., 2013; Muraveva et al., 2014; Inno et al., 2016; Bhardwaj et al., 2016; Wagner-Kaiser et al., 2017), and produced different estimates of the distance modulus based on diverse methods and observations.

The distance to the LMC has been determined using a variety of standard candles, such as, Cepheid variables, Red Clump stars, RR Lyrae stars, Miras variables, supernova 1987A, and eclipsing binaries.

A Cepheid variable is a type of giant star that pulsates radially, varying in both diameter and surface temperature and producing changes in brightness with a well-defined stable period and amplitude. The strong direct relationship between a Cepheid variable's luminosity and pulsation period makes Cepheids as important

indicators of cosmic benchmarks for scaling galactic and extragalactic distances (e.g., Feast and Catchpole, 1997; Laney and Stobie, 1994; Di Benedetto, 1997; Gieren et al., 1998; Groenewegen, 2000).

An eclipsing binary star is a binary star in which the orbital plane of the two stars lies so nearly in the line of sight of the observer that the components undergo mutual eclipses. They can also be used to directly measure the distances to external galaxies (e.g., Fitzpatrick et al., 2002).

The red clump is a clustering of red giants in the Hertzsprung–Russell diagram at around 5,000 K and absolute magnitude (M_V) +0.5, slightly hotter than most red-giant-branch stars of the same luminosity. It is visible as a denser region of the red giant branch or a bulge towards hotter temperatures. They can be used as candles to determine the distance to the LMC (e.g., Romaniello et al., 2000; Pietrzyński and Gieren, 2002; Sarajedini et al., 2002; Clementini et al., 2003).

Mira variables are a class of pulsating variable stars characterized by very red colours, pulsation periods longer than 100 days, and amplitudes greater than one magnitude in the infrared and 2.5 magnitude at visual wavelengths. They are red giants in the very late stages of stellar evolution, on the asymptotic giant branch, that will become white dwarfs within a few million years. They can be used to determine the distance to the LMC (e.g., van Leeuwen et al., 1997; Feast, 2005).

SN 1987A was a supernova in the outskirts of the Tarantula Nebula in the LMC. It was easily visible to the naked eye from the Southern Hemisphere. Astronomers have employed it to measure the distance modulus to the LMC (e.g., Panagia, 1998).

Numerous attempts have been made to summarise the published distance moduli (e.g., Alves, 2004; De Grijs et al., 2014; Crandall and Ratra, 2015). One commonly used way to summarise the population of published distance moduli is to compute the precision-weighted average as the population mean, and compute the weighted mean

standard deviation as its uncertainty (Crandall and Ratra, 2015). In this chapter, we illustrate how Bayesian hierarchical models combine case-by-case fits (published distance moduli) and produce a comprehensive analysis. This analysis is just a testbed of our statistical techniques, and the astronomical results from our analysis are not definitive. The aim of this chapter is to show readers how Bayesian hierarchical models work in a simple but realistic example.

3.1 Introduction

We employ the statistical techniques to 13 published distance moduli in Table 3.1 compiled by Clementini et al. (2003) to test how our statistical methods work. In the meantime, we obtain preliminary results about the published distance moduli population. These distance moduli results are subject to change, depending upon the data included. The target of this chapter is to illustrate how to use Bayesian hierarchical models to study the population of published results on a particular problem, not to yield a scientifically sound estimate of the distance modulus to the LMC.

Besides statistical errors, the various distance estimates may be subject to systematic errors. We aim to estimate the magnitude of these systematic errors. If we further assume that the systematic errors tend to average out among the various estimators, we can obtain a better comprehensive estimator of the distance modulus. Let μ_i be the best estimate of the distance modulus that could be obtained with method i , i.e., with an arbitrarily large dataset. Because of systematic errors, μ_i does not equal the true distance modulus, but is free of statistical error.

In the following sections, we treat published results as case-by-case fits and combine them in a hierarchical model to yield a comprehensive analysis.

Table 3.1: Published Distance Indicators

Method	Dist. Mod. $D_i \pm \sigma_i$	References
Cepheids: trig. paral.	18.70 ± 0.16	Feast and Catchpole (1997)
Cepheids: MS fitting	18.55 ± 0.06	Laney and Stobie (1994)
Cepheids: B-W	18.55 ± 0.10	Di Benedetto (1997); Gieren et al. (1998)
Cepheids: P/L relation	18.575 ± 0.2	Groenewegen (2000)
Eclipsing binaries	18.4 ± 0.1	Fitzpatrick et al. (2002)
Clump	18.42 ± 0.07	Clementini et al. (2003)
Clump	18.45 ± 0.07	Clementini et al. (2003)
Clump	18.59 ± 0.09	Romaniello et al. (2000)
Clump	18.471 ± 0.12	Pietrzyński and Gieren (2002)
Clump	18.54 ± 0.10	Sarajedini et al. (2002)
Miras	18.54 ± 0.18	van Leeuwen et al. (1997)
Miras	18.54 ± 0.14	Feast (2005)
SN 1987a	18.54 ± 0.05	Panagia (1998)

3.2 Hierarchical Analysis

Consider the statistical model,

$$D_i \sim N(\mu_i, \sigma_i^2), \quad i = 1, \dots, 13, \quad (3.1)$$

$$\mu_i \sim N(\gamma, \tau^2), \quad (3.2)$$

where D_i is the actual estimated distance modulus based on the method/dataset i including statistical error, σ_i is the known standard deviation of the statistical error, γ is the true distance modulus of the LMC, and τ is the standard deviation of the systematic errors of the various estimates. In this research, we treat each published result as an object, then μ_i , the statistically error-free distance modulus to the LMC from study i , is the object-level parameter. The true distance modulus to the LMC, γ , and standard deviation of systematic errors from different measurements, τ , are population-level parameters. A summary of parameters is presented in Table 3.2.

Table 3.2: Parameters in the hierarchical model

Object-level parameters	
μ_i	statistical-error-free distance modulus to the LMC via method i
Population-level parameters	
γ	true distance modulus to the LMC
τ	standard deviation of systematic errors from different methods

Eq. (3.2) specifies our assumption that the systematic errors tend to average out. We denote $\mathbf{D} = (D_1, \dots, D_{13})$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{13})$.

We take both the EB and FB approach to fitting the hierarchical model in Eq. (3.1)–(3.2). Our Bayesian approaches (EB and FB) require a prior distribution for γ and τ . It is a standard practice to factorise the joint prior, $p(\gamma, \tau) \propto p(\gamma|\tau)p(\tau)$, and set $p(\gamma|\tau) \propto 1$. As discussed in Section 2.4, we assign a non-informative prior on the standard deviation τ , namely, $p(\tau) \propto 1$, or equivalently, $p(\tau^2) \propto (\tau^2)^{-\frac{1}{2}}$.

Because this hierarchical model is straightforward, we take the Gibbs sampler to implement the FB method. For simplicity, we call this approach FB with Gibbs sampler, or in short form, FB-Gibbs. We cover the detailed way to implement the EB approach. This involves first estimating the population-level parameters γ and τ and then plugging these estimates in Eq. (3.2) and using it as the prior distribution for each μ_i . Finally the individual μ_i are estimated with their posterior expectations, $E(\mu_i|\mathbf{D}, \hat{\gamma}, \hat{\tau}^2)$ and their posterior standard deviations, $SD(\mu_i|\mathbf{D}, \hat{\gamma}, \hat{\tau}^2)$ are used as 1σ error bars.

We estimate γ and τ by maximising their joint posterior density,

$$\begin{aligned}
 p(\gamma, \tau|\mathbf{D}) &\propto \int p(\tau, \gamma, \boldsymbol{\mu}|\mathbf{D}) d\boldsymbol{\mu} \\
 &= p(\gamma, \tau) \prod_{i=1}^{13} \int p(D_i|\mu_i) p(\mu_i|\gamma, \tau) d\mu_i.
 \end{aligned} \tag{3.3}$$

The values of γ and τ that maximise Eq. (3.3) are known as maximum a posterior (MAP) estimates. For any τ , Eq. (3.3) is maximised with respect to γ by

$$\hat{\gamma}(\tau) = \frac{\sum_{i=1}^{13} D_i / (\tau^2 + \sigma_i^2)}{\sum_{i=1}^{13} 1 / (\tau^2 + \sigma_i^2)}, \quad (3.4)$$

where $\hat{\gamma}(\tau)$ is a function of τ . The profile posterior density of τ is obtained by evaluating Eq. (3.3) at $\hat{\gamma}(\tau)$ and τ , i.e., $p(\hat{\gamma}(\tau), \tau | \mathbf{D})$. The global maximiser of the profile posterior distribution is the MAP estimate of τ and must be obtained numerically.

As shown in the left panel of Fig. 3.1, the posterior density of τ monotonically decreases from its peak near 0, which means that the MAP estimator of τ is 0, a poor summary of the profile posterior. This is because 0 is the lower boundary of the possible values of τ . A better estimate can be obtained using a logarithmic transformation of the population standard deviation, specifically, $\xi = \ln \tau$. The joint posterior of γ and ξ can be expressed as

$$p(\gamma, \xi | \mathbf{D}) = p(\gamma, \tau | \mathbf{D}) \left| \frac{d\tau}{d\xi} \right|.$$

The profile posterior of ξ is plotted in the right panel of Fig. 3.1, is more symmetric, and is better summarised by its mode. After having estimated ξ with its MAP estimate, we compute $\hat{\tau} = \exp(\hat{\xi})$ and $\hat{\gamma} = \gamma(\hat{\tau})$. See Park et al. (2008) for a discussion of transforming parameters to achieve approximate symmetry in the case of mode-based estimates. This method is a variant of the common EB approach, and we call it EB with logarithmic transformation, or in short form, EB-log.

We use EM algorithm to find the MAP estimates of γ, ξ and τ , and Fig. 3.2 exhibits the iterations of γ and τ . It can be seen that both γ and τ stabilise after 50 iterations.

Plugging $\hat{\gamma}$ and $\hat{\tau}$ into the prior for μ_i given in Eq. (3.2), we can compute the

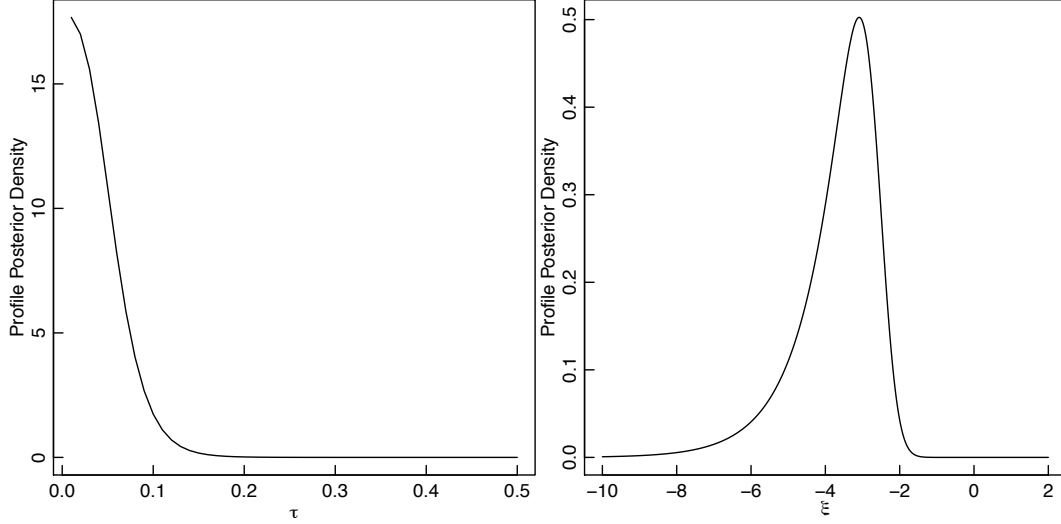


Figure 3.1: The profile posterior distribution of τ (left panel) and of $\xi = \ln \tau$ (right panel). The modal estimate of τ is zero and is less representative of the distribution than is the modal estimate of ξ .

posterior distribution of each μ_i as

$$p(\boldsymbol{\mu}|\mathbf{D}, \hat{\gamma}, \hat{\tau}^2) \propto \prod_{i=1}^{13} p(D_i|\mu_i)p(\mu_i|\hat{\gamma}, \hat{\tau}).$$

3.3 Simulation Study

In this section, we perform a simulation study to investigate the performance of the hierarchical model in Eq. (3.1)–(3.2) and compare three different statistical methods: Empirical Bayes with log transformation (EB-log), Fully Bayesian (FB) and case-by-case analysis. Chapter 2 covered EB, FB and case-by-case methods. Here, EB-log is a variant of EB and refers to Section 3.2 for its details. Basically we use a logarithm transformation on the population-level parameter τ to avoid the case that the maximum a posterior (MAP) estimate of τ to be 0.

Based on their fitted values with the data in Table 3.1, we set the true values for

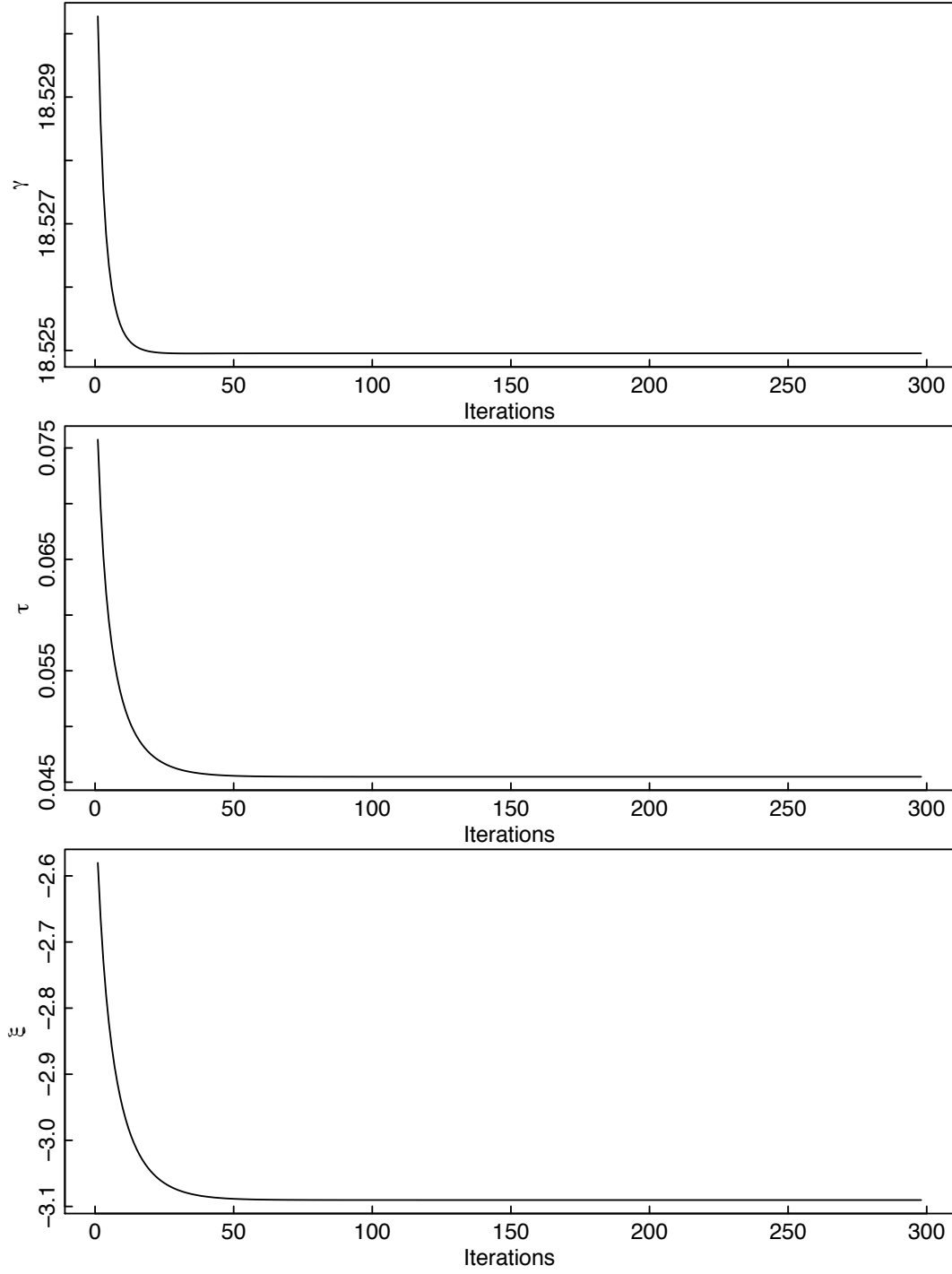


Figure 3.2: The EM iterations of γ , τ and ξ .

population-level parameters $\gamma = 18.525$ and $\tau = 0.045$. Then we simulate 25 sets of true values for $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{13})$ from $\mu_i \sim N(\gamma, \tau^2)$. For clarity, we denote the j -th replicate as $\boldsymbol{\mu}_j = (\mu_{ij}, i = 1, \dots, 13), j = 1, \dots, 25$. For each replicate of $\boldsymbol{\mu}$ we generate 100 datasets $\mathbf{D}$ with known σ_i from Table 3.1. Specifically, for the j -th replicate $\boldsymbol{\mu}_j$, we repeat the simulation of $D_{ij} \sim N(\mu_{ij}, \sigma_i^2)$ 100 times. So in total, we simulate 25×100 datasets and based on each simulated dataset $\mathbf{D}$ we utilise three methods – EB-log, FB and Case-by-case – to estimate both the object-level and population-level parameters $\boldsymbol{\mu}, \gamma$ and τ .

We compare estimates in terms of two criteria: i.) bias, standard deviation and root mean squared error (RMSE) of point estimates; 2.) actual versus nominal coverage probabilities of interval estimates. When we require point estimates, we utilise posterior means from MCMC samples, and for interval estimates we use the 68.3% credible intervals from posterior distributions by finding the 15.8% and 84.1% quantiles from the MCMC samples.

Fig. 3.3 shows the biases (top-left panel), standard deviations (top-right), RMSEs (bottom-left) of estimates of μ_{ij} versus simulated values of μ_{ij} , where estimates obtained from the case-by-case, FB and EB-log are in black plus signs, blue circles and green triangles, respectively. The top-left panel presents the biases of estimates of μ_{ij} versus simulated μ_{ij} . Black plus signs from the case-by-case analyses always oscillate around 0 and almost all of them fall in the range from -0.05 to 0.05 . By contrast, the biases from FB (blue circles) and EB-log (green triangles) present a similar pattern in that the biases of estimates of μ_{ij} are linearly decreasing in μ_{ij} . This is due to the fact that hierarchical methods such as EB-log and FB produce shrinkage estimates, objects (μ_{ij}) at both ends of the distribution are shrunk slightly to the middle. So the small values of μ_{ij} in the left side tends to be overestimated, and the large values of μ_{ij} in the right side are underestimated. The top-right panel

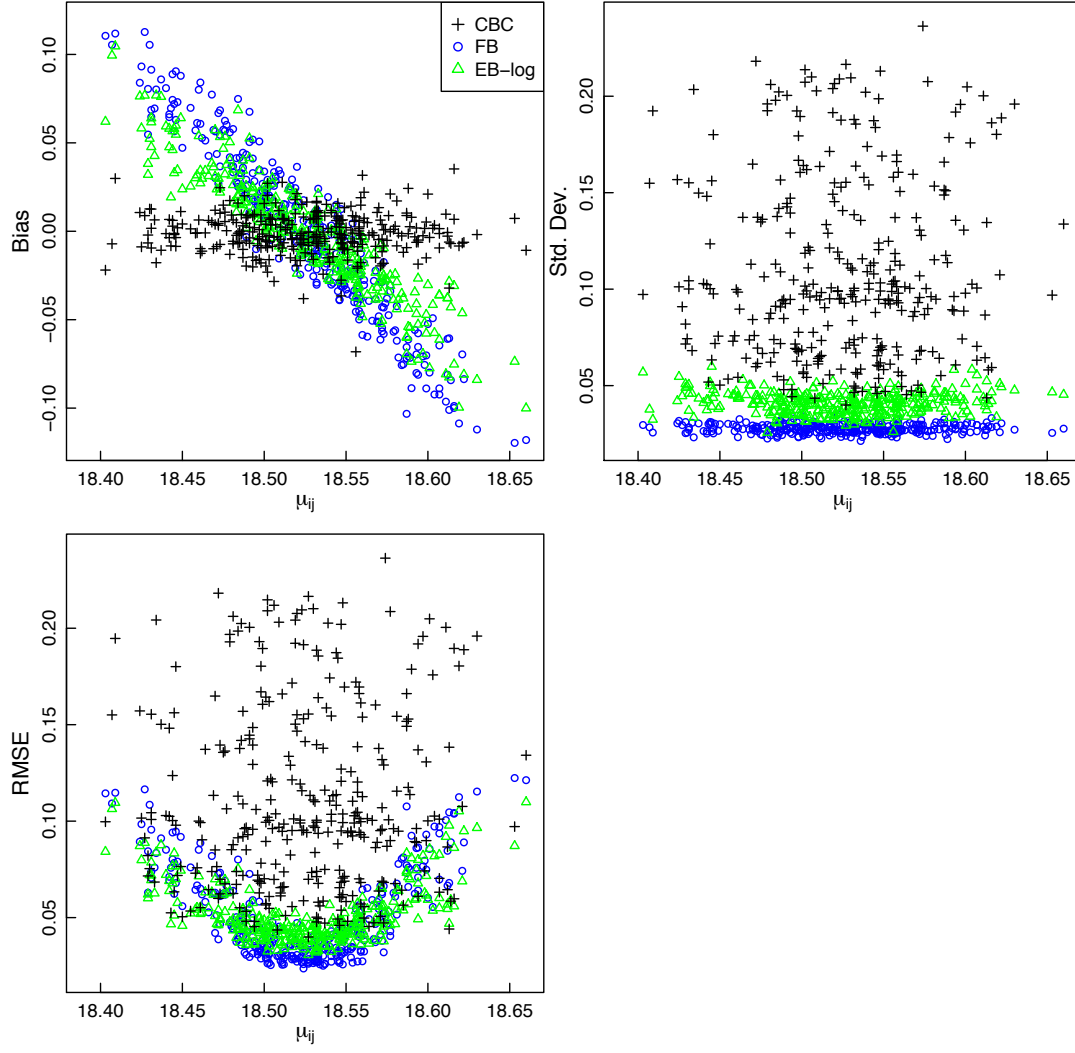


Figure 3.3: The biases, standard deviations, and RMSEs of point estimates of μ_{ij} versus true values of μ_{ij} . Estimates are obtained from three approaches: the case-by-case (black plus signs), FB-Gibbs (blue circles) and EB-log (green triangles).

Table 3.3: The average bias, standard deviation (SD) and RMSE of the various estimates of μ , γ and τ . Each figure has a multiplication factor 10^{-2} and we omit it to make the numbers more legible.

Items	EB-log			FB-Gibbs			Case-by-case		
	Bias	S.D.	RMSE	Bias	S.D.	RMSE	Bias	S.D.	RMSE
μ	-0.14	4.08	5.12	-0.14	2.77	5.03	-0.09	10.99	11.05
γ	-0.09	2.5	2.81	-0.09	2.52	2.83	n.a.	n.a.	n.a.
τ	1.77	2.18	2.82	1.72	2.53	3.09	n.a.	n.a.	n.a.

shows the standard deviations of point estimates of μ_{ij} , and case-by-case estimates have the largest standard deviations, followed by EB-log. The FB estimates have the smallest standard deviations, i.e., the largest precision. The bottom-left panel displays the RMSEs of three methods. The case-by-case estimates have the largest RMSEs, i.e., least accuracy. By contrast, hierarchical methods have much smaller RMSEs. Hierarchical methods trade biases for variance reductions and finally they attain smaller RMSEs.

Table 3.3 presents the average bias, standard deviation and RMSE of three estimates of μ , γ and τ . The average bias of μ from the case-by-case method is -0.09 , comparable to EB-log and FB, all at -0.14 . The average standard deviation of μ from the FB-Gibbs is the smallest (2.77), followed by EB-log (4.08), and the case-by-case method has the largest mean standard deviation (10.99), about three times that of FB-Gibbs. In terms of average RMSE, the FB-Gibbs performs the best (5.03), followed by the EB-log (5.12), and the case-by-case is worst (11.05), about twice the hierarchical methods.

As to the population-level parameters, the case-by-case method does not have the second-level model in Eq. 3.2, so it fails to provide estimates of γ and τ . The EB-log and FB-Gibbs methods behave similarly in terms of the bias, variance and MSE of γ and τ , though there is some difference between the standard deviations of τ from the

Table 3.4: The actual coverage probabilities (CP) and their 68.3% confidence intervals (CI) of the various credible interval estimates of $\boldsymbol{\mu}$, γ and τ . The nominal coverage probability is 68.3%. We obtain these results by pooling estimates from all simulated datasets. For $\boldsymbol{\mu}$, its coverage probability is the average over $\mu_1, \dots, 13$.

Items	EB-log		FB-Gibbs		Case-by-case	
	CP	68.3% CI	CP	68.3% CI	CP	68.3% CI
$\boldsymbol{\mu}$	0.645	(0.642, 0.647)	0.675	(0.673, 0.678)	0.685	(0.683, 0.688)
γ	n.a	n.a.	0.83	(0.826, 0.834)	n.a.	n.a.
τ	n.a	n.a.	0.82	(0.816, 0.824)	n.a.	n.a.

EB-log and FB-Gibbs. Also, both the EB-log and FB-Gibbs slightly overestimate the population standard deviation τ . For FB-Gibbs, the overestimation of τ is caused by its asymmetric parameter space, i.e., τ is limited to the positive real numbers (Gelman, 2006b).

Table 3.4 displays the actual coverage probabilities (CP) and their 68.3% confidence intervals (CI) for credible intervals with nominal coverage probability 68.3% for the object-level parameters $\boldsymbol{\mu}$ and the population-level parameters γ and τ . The 68.3% CIs of actual coverage probabilities in this section are obtained via Clopper-Pearson method (Clopper and Pearson, 1934). From this table, the case-by-case credible intervals of $\boldsymbol{\mu}$ exhibit a slightly better coverage property (68.5%), close to the nominal level 68.3%, than the FB-Gibbs estimates of $\boldsymbol{\mu}$ (67.5%). The EB-log estimates of $\boldsymbol{\mu}$ has its actual coverage, 64.5%, slightly lower than the nominal level. Because the EB-log and Case-by-case methods do not produce interval estimates for population-level parameters γ and τ , we only have actual coverage probabilities for γ and τ from the FB-Gibbs, which are 83% and 82%, respectively. The actual coverage probabilities of FB-Gibbs interval estimates of γ and τ exceed the nominal level, which is caused by the small number of objects, $I = 13$, in the hierarchical model. Other research have also reported and discussed this properly, see Browne and Goldstein (2002); Browne

et al. (2006) for more details.

In summary, we have a few comments about the EB-log, FB-Gibbs and case-by-case methods. 1.) The FB-Gibbs method has the smallest RMSE of $\boldsymbol{\mu}$, about half of the case-by-case, which has the largest one. The RMSE of $\boldsymbol{\mu}$ from EB-log is slightly greater than FB-Gibbs, but they are comparable. This is also the case for the standard deviation of estimates of $\boldsymbol{\mu}$. 2.) The average coverage probabilities for interval estimates of $\boldsymbol{\mu}$ from three methods are comparable, all close to the nominal level. 3.) The case-by-case analysis has no estimates for population-level parameters γ and τ , but the hierarchical approaches have estimates of γ and τ and their performances are similar. 4.) FB-Gibbs is the only approach that produces interval estimates for population-level parameters, although its actual coverage exceeds the nominal level due to the small number of objects in the hierarchical model.

Because this project is based on a Gaussian-Gaussian hierarchical model, the computing time is the same as the simulation study in Chapter 2. Refer to Table 2.6 for the computing time that each algorithm takes in the simulation study. The computing time required by the EB-log is the same as EB in Table 2.6. The CPU times for the FB-Gibbs and the case-by-case analysis are also the same as the times shown in Table 2.6.

3.4 Data Analysis

In this section, we present the results from hierarchical analyses via both the EB-log and FB-Gibbs approach by pooling the case-by-case results in Table 3.1. The target of this section is to show readers how Bayesian hierarchical models work in a simple but realistic example.

Firstly, we focus on the object-level parameter $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{13})$, which are the statistical error-free values of the distance modulus to the LMC from 13 studies.

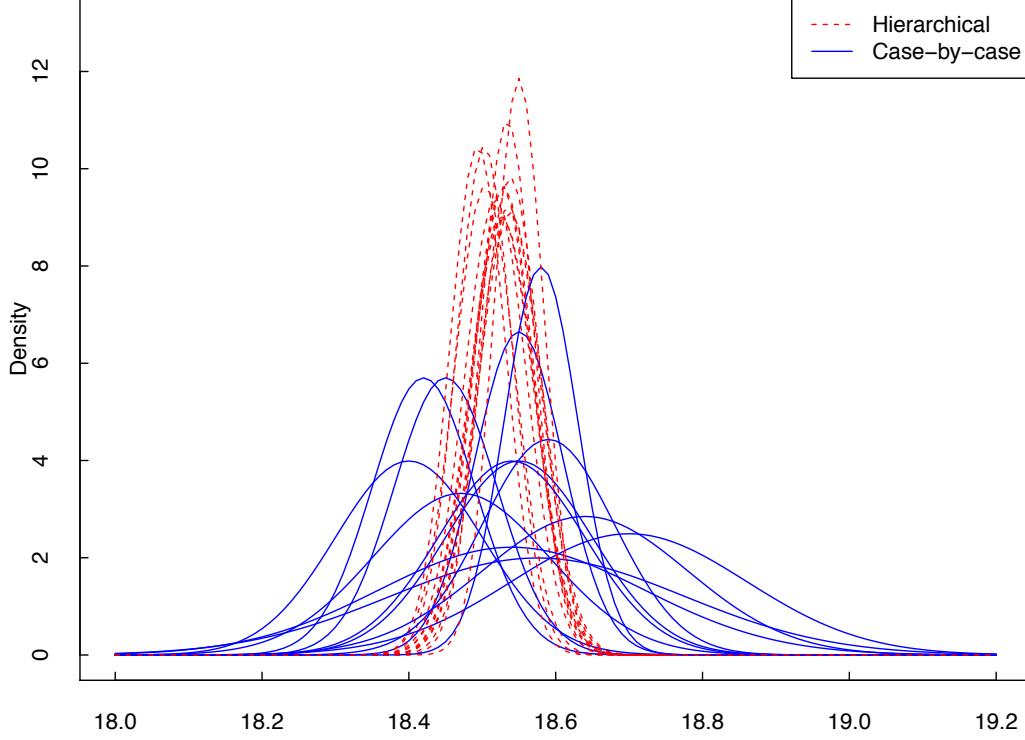


Figure 3.4: Posterior density functions of the LMC distance modulus based on hierarchical fitting (Red dashed lines) and based on the case-by-case analysis (Blue solid lines). Those based on hierarchical fitting shrink towards the centre relative to those based on the case-by-case analysis.

Fig. 3.4 shows the EB-log and case-by-case posterior distributions of μ_i . The hierarchical results (dashed red lines) are shrunk toward the centre relative to the case-by-case results (blue solid lines). Fig. 3.5 also demonstrates this shrinkage effect that point estimates (posterior means) from the EB-log fit tend to be closer to each other than those from case-by-case analyses.

In Fig. 3.4, the case-by-case density functions of μ_i range from 18.0 to 19.2, whereas the hierarchical posterior density functions are more precise, ranging from 18.3 to 18.7. This is an example of the shrinkage of the case-by-case fits towards their average that occurs when fitting a hierarchical model. We can also see this effect in the

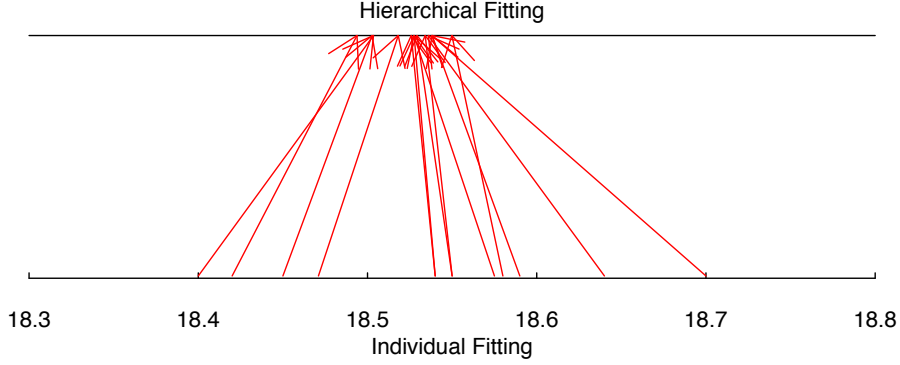


Figure 3.5: The point estimates of distance moduli from the hierarchical (upper line) and case-by-case (individual in the lower line) fittings. Each arrow connects the point estimate from one certain case-by-case analysis with its hierarchical counterpart.

mathematical formula of posterior means,

$$E(\mu_i|\tau, \mathbf{D}) = \frac{\hat{\gamma}(\tau)/\tau^2 + D_i/\sigma_i^2}{1/\tau^2 + 1/\sigma_i^2}, \quad (3.5)$$

which are weighted averages of the case-by-case estimates, D_i , and the combined MAP estimate $\hat{\gamma}$ of the mean distance modulus, given in Eq. (3.4).

Fig. 3.6 displays the point estimates and their 68.3% credible intervals of the statistically error-free distance modulus to the LMC based on 13 measurements from both the case-by-case and EB-log, FB-Gibbs fits. The triangles and solid lines are estimates from the EB-log fit, the squares and dotted lines from the FB-Gibbs fit, and the dots and dashed lines are based on the case-by-case fits published in the 13 studies. We observe from Fig. 3.6 that the 68.3% credible intervals from the hierarchical methods (EB-log and FB-Gibbs) are shorter than their counterparts from the case-by-case fits. The solid triangles, dots, and squares are posterior means from the case-by-case fits.

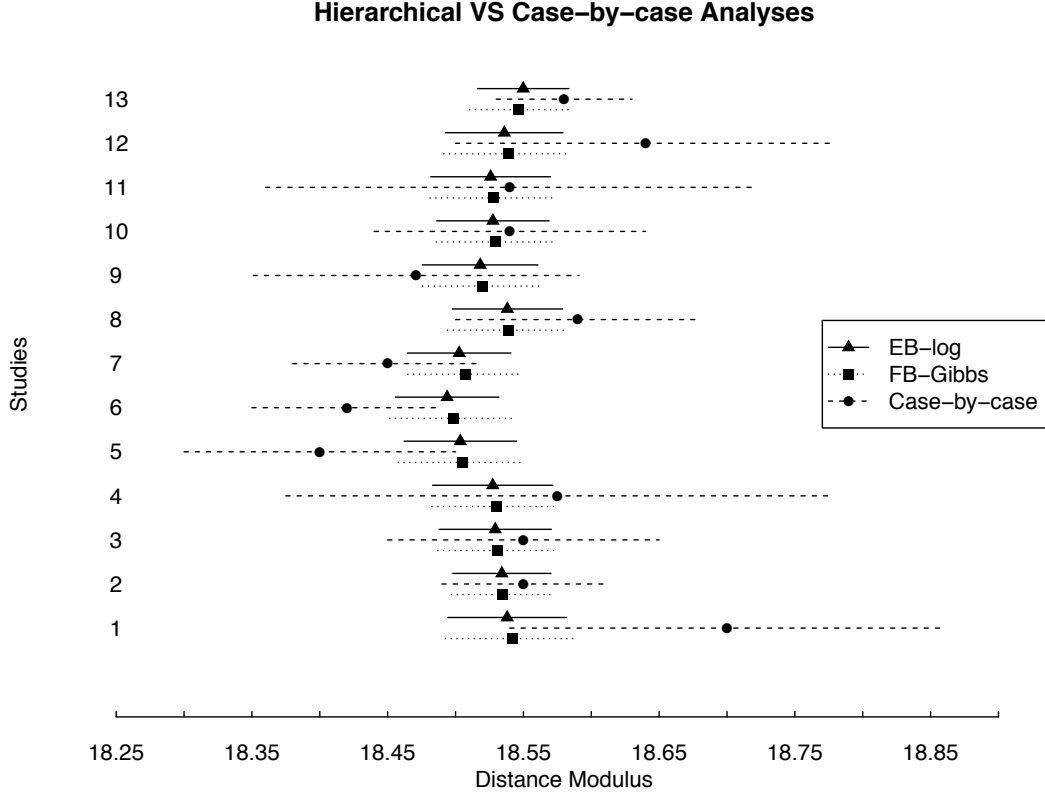


Figure 3.6: The point estimates and 68.3% credible intervals for the distance moduli to LMC from the EB-log, FB-Gibbs and case-by-case analyses.

EB-log, case-by-case and FB-Gibbs, respectively. The solid, dashed and dotted lines are 68.3% credible intervals from these three methods. All figures of estimates are shown in Table 3.5.

Based on our simulation study in Section 3.3, we can understand how these estimates behave. From figures in Table 3.3, the EB-log estimates of μ have the smallest MSE (4.282×10^{-3}), about a quarter of that of the case-by-case (1.44×10^{-2}). FB-Gibbs estimates of μ have slightly larger MSE (5.227×10^{-3}) than the EB-log, which is approximately a third of that from case-by-case analysis. So in terms of accuracy, EB-log is the best, followed by FB-Gibbs, and case-by-case analysis is worst. From the same table, the case-by-case yields estimates of μ that have the largest variance

Table 3.5: Point and 68.3% credible interval estimates of the object-level parameter $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{13})$ for 13 included studies from the EB-log, FB-Gibbs and case-by-case analyses.

Study	EB-log		FB		Case-by-case	
	Mean	Cred. Interv.	Mean	Cred. Interv.	Mean	Cred. Interv.
1	18.54	(18.49, 18.58)	18.54	(18.5, 18.58)	18.7	(18.54, 18.86)
2	18.53	(18.5, 18.57)	18.53	(18.5, 18.57)	18.55	(18.49, 18.61)
3	18.53	(18.49, 18.57)	18.53	(18.5, 18.57)	18.55	(18.45, 18.65)
4	18.53	(18.48, 18.57)	18.52	(18.49, 18.56)	18.57	(18.38, 18.77)
5	18.5	(18.46, 18.54)	18.51	(18.47, 18.54)	18.4	(18.3, 18.5)
6	18.49	(18.46, 18.53)	18.5	(18.46, 18.54)	18.42	(18.35, 18.49)
7	18.5	(18.46, 18.54)	18.51	(18.48, 18.54)	18.45	(18.38, 18.52)
8	18.54	(18.5, 18.58)	18.53	(18.5, 18.57)	18.59	(18.5, 18.68)
9	18.52	(18.48, 18.56)	18.52	(18.49, 18.56)	18.47	(18.35, 18.59)
10	18.53	(18.49, 18.57)	18.53	(18.5, 18.56)	18.54	(18.44, 18.64)
11	18.53	(18.48, 18.57)	18.53	(18.49, 18.57)	18.54	(18.36, 18.72)
12	18.54	(18.49, 18.58)	18.53	(18.5, 18.57)	18.64	(18.5, 18.78)
13	18.55	(18.52, 18.58)	18.54	(18.51, 18.58)	18.58	(18.53, 18.63)

(on average, 1.427×10^{-2}) among all three methods, about 15 and 7 times the variance from FB-Gibbs and EB-log, respectively. So in terms of precision of $\boldsymbol{\mu}$, the FB-Gibbs performs the best, and case-by-case method is also the worst.

From Table 3.4, the case-by-case credible intervals of $\boldsymbol{\mu}$ have the best coverage property, always close to the nominal level 68.3%. The FB-Gibbs and EB-log 68.3% credible interval estimates of $\boldsymbol{\mu}$ have actual coverage probability, about 56.1% and 57.3%, respectively.

For illustration, in Fig. 3.7 we plot the posterior expectation $E(\mu_i|\tau, \mathbf{D})$ of each of the best estimate of the distance moduli from each method as a function of τ as 13 coloured lines. The black solid line is the MAP estimate $\hat{\gamma}(\tau)$ plotted as a function of τ . When τ is close to zero, the conditional posterior means of each μ_i shrink toward the overall weighted mean $\hat{\gamma}(0) = \frac{\sum D_i/\sigma_i^2}{1/\sigma_i^2} = 18.526$. This weighted mean is used by others (De Grijs et al., 2014) to summarise multiple published results.

Table 3.6: Estimates for population-level parameters: the true distance modulus (γ) and standard deviation (τ) of systematic errors from 13 different studies.

Items	EB-log	FB-Gibbs
Estimates for the true distance modulus γ	18.525	18.525 ± 0.028
Estimates for the standard deviation of different studies τ	0.045	0.033 ± 0.032

The $\tau = 0$ case corresponds to no systematic error and relatively large statistical error. As τ becomes larger, the conditional posterior means approach the case-by-case estimators of the distance modulus marked by plus signs at the far right in Fig. 3.7. The red dashed vertical line indicates our estimate of τ and intersects the coloured curves at the hierarchical estimates of each μ_i . Fig. 3.7 shows how the hierarchical fit reduces to the case-by-case analyses as the variance of the systematic errors goes to infinity. We include Fig. 3.7 to illustrate the “shrinkage” of the estimates produced with hierarchical models, but such a plot is not needed to obtain the final fit.

Here we present the statistical inference of population-level parameters: the true distance modulus to the LMC, γ , and the standard deviation of systematic errors of 13 published studies, τ . From the FB-Gibbs, we obtain the point and interval estimates of γ and τ . However, we only have point estimates of them from the EB-log. Table 3.6 displays estimates of population-level parameters from EB-log and FB-Gibbs fits.

Table 3.6 presents the point estimates of γ and τ from the EB-log, the point estimates with one standard deviation error bar (or 68.3% credible interval) of γ and τ from the FB-Gibbs fit. From Table 3.6, the point estimates of γ from the EB-log and FB-Gibbs coincide at 18.525. The 68.3% credible interval of γ from FB-Gibbs is 18.525 ± 0.028 . The EB-log estimate of τ is larger than that of FB-Gibbs. This is caused by the logarithmic transformation of τ used by EB-log. The FB-Gibbs point estimate of τ is 0.033 with standard deviation 0.032, which means that 95%

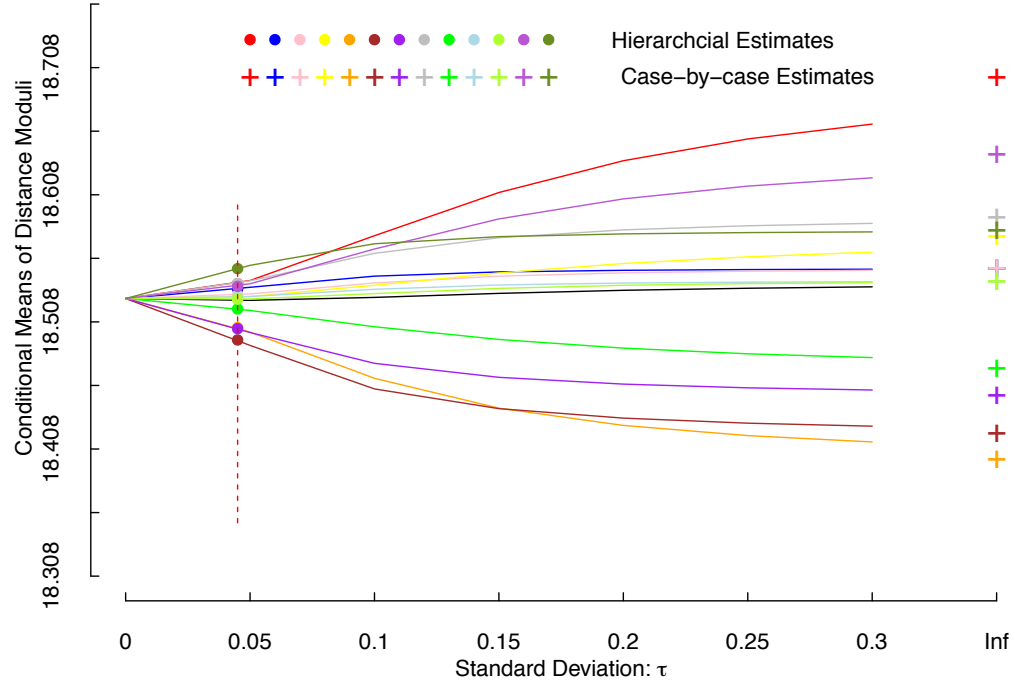


Figure 3.7: Conditional posterior means of distance modulus $E(\mu_i|\tau, \mathbf{D})$, as a function of the standard deviation of the systematic errors τ . The red dashed line indicates our estimate $\hat{\tau} = 0.045$. The 13 solid round dots of different colours show the hierarchical estimates of the distance modulus to the LMC using various methods, whereas the plus signs are estimates from the case-by-case analyses. (Plus signs and round dots of the same colour correspond to the same published result; see Table 3.1.)

credible interval of τ contains 0, i.e., τ is not significantly greater than 0. We have three comments on it. 1.) This verifies the fact that the posterior of τ peaks at 0 without the logarithmic transformation, as shown in Fig. 3.1. 2.) When $\tau = 0$, it means the statistical-error-free distance moduli from all 13 methods are equal to the true distance modulus, i.e., all these methods have no systematic errors, which is quite unrealistic. 3.) When $\tau = 0$, the hierarchical estimates of γ reduces to the commonly used weighted mean $\frac{\sum_i y_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$ and standard deviation $\sqrt{(\sum_i 1/\sigma_i^2)^{-1}}$ (See Podariu et al., 2001; De Grijs et al., 2014). The computing time of EB-log, FB-Gibbs and case-by-case fits are the same as EB, FB-Gibbs and case-by-case, respectively, in the simulation study in Chapter 2. Refer to Table 2.6 for specific numbers.

Based on the simulation study results in Table 3.4, the FB-Gibbs credible interval estimates of population-level parameters γ and τ are over-covering, and their coverage probabilities on γ and τ are about 80% and 76%, respectively.

This project aims to propose statistical methods to review published results on a certain problem. Our analysis of the published distance moduli to the LMC is preliminary and tentative. It serves as a test to check whether our statistical methods make sense. From the simulation study, we show the benefits of the Bayesian hierarchical methods: smaller MSEs and reliable estimates of population-level parameters. Numerous attempts have been made to combine published results on the distance modulus to the LMC. After compiling 233 separate distance modulus published between 1990 and 2003, De Grijs et al. (2014) suggests a slightly adjusted canonical distance modulus of $(m - M)_0 = 18.49 \pm 0.09$, which is obtained as the weighted average of all included results. Their result has a wider error bar (0.09) than ours (0.028), and it is understandable because they include 233 published distance moduli, which vary on a broad range. However, in our pilot analysis we only use 13 results compiled by Clementini et al. (2003). Our approaches produce completely data-driven

outcomes, so if different data are included in the analysis, our outcome will change. The published outcomes included in my analysis were published between 1994 and 2003, only a subset of those included in De Grijs et al. (2014), so we might use our method to analyse the data compiled by De Grijs et al. (2014) in the future.

3.5 Concluding Remarks

Distance modulus to the LMC is important in astronomy, so large number of attempts have been made to determine it (e.g., Di Benedetto, 1997; Fitzpatrick et al., 2002; Pietrzyński et al., 2013; Inno et al., 2013). A number of researchers have attempted to review the published distance moduli to the LMC (e.g., Clementini et al., 2003; Alves, 2004; De Grijs et al., 2014).

This chapter is a tentative analysis and aims to show how Bayesian hierarchical models combine case-by-case fits (published results) and produce shrinkage estimates in a simple but realistic setting. In this chapter, we investigate the use of a Bayesian hierarchical model to review 13 published distance moduli to the LMC and obtain statistical inferences of the population of published results. The idea is to use the published distance moduli and their errors as observed data and Gaussian errors for the statistical-error-free distance moduli from their methods, and then their statistical-error-free distance moduli are assumed to be a Gaussian distribution centred at the true distance modulus to the LMC and its standard deviation represents the magnitude of systematic errors of these methods. Of course, here we have made a quite strong assumption that systematic errors of different studies follow a Gaussian distribution with mean 0.

In this chapter, we focus on the statistical side of this research. We employ the EB-log and FB-Gibbs to fit the hierarchical model, show how hierarchical model works, perform a simulation study to compare the performance of hierarchical and

case-by-case methods. Based on the simulation study, we see the benefits of hierarchical fits over the case-by-case in terms of smaller MSEs and variances. Also, the credible intervals from hierarchical analyses tend to have some difficulty in attaining the nominal coverage probability, which is caused by the small number of objects.

As an experimental and preliminary test, we apply our statistical techniques to 13 published distance moduli to the LMC, compiled by Clementini et al. (2003). Using the hierarchical methods, we obtain the statistical inference of the population of 13 published distance moduli, i.e., the point and interval estimates of population mean γ and standard deviation τ of systematic errors of different methods. This analysis is very tentative in terms of small size of distance moduli included and the strong Gaussian assumption on systematic errors. However, this research does provide a new way to combine published distance moduli to the LMC and point estimate of γ can be an alternative to the weighted average.

In the future, we might use the 233 distance moduli collected by De Grijs et al. (2014) to provide a hierarchical analysis and obtain a statistical inference of the population of published distance moduli. Also, we may take account of the effect of different tracers, for example, the Cepheids, RR Lyrae, red clump stars, etc.

4 Estimating the Ages of a Group of Halo White Dwarfs

This research aims to develop efficient computational algorithms to fit Bayesian hierarchical models when the likelihood function is computer-based and not available in closed-form. In this chapter, we pool 10 White Dwarfs (WDs) into a hierarchical model and employ our statistical techniques to obtain estimates of ages of WDs that are expected to be closer to their true ages than fitting WDs one-by-one.

This chapter focuses on the statistical side of the problem, not on the astronomy side. We therefore do not investigate the underlying stellar evolution models. Instead we explore new algorithms to tackle computational challenges when pooling WDs into a Bayesian hierarchical model. Though the astronomy side of this research is preliminary, both in the WD sample size and the assumed age distribution, our technique lays the ground work for large-scale studies using data from the Gaia mission.

Suppose we have n WDs from the Galactic halo. We denote the l observed photometric magnitudes for WD i by $\mathbf{X}_i$ and the known variance-covariance matrix of their measurement errors by Σ_i . We aim to use $\mathbf{X}_i$ to estimate the age, distance modulus, metallicity, and zero-age main sequence (ZAMS) mass of WD i . Our WD model is specified in terms of the $\log_{10}(\text{age})$, distance modulus, metallicity, ZAMS mass of WDs and we denote these parameters by A_i, D_i, Z_i and M_i for $i = 1, \dots, n$

respectively. Because we are primarily interested in WD ages, we group the other stellar parameters into $\Theta_i = (D_i, Z_i, M_i)$. Finally to simplify notation, we write $\mathbf{X} = (\mathbf{X}_i, \dots, \mathbf{X}_n)$, $A = (A_1, \dots, A_n)$, and $\Theta = (\Theta_1, \dots, \Theta_n)$. Here we review a case-by-case analysis method for WDs and develop convenient approaches to obtain the hierarchical modelling fits with better statistical properties.

4.1 Case-by-case Analysis

The public-domain Bayesian software suite, Bayesian Analysis of Stellar Evolution with 9 parameters (BASE-9), allows one to precisely estimate cluster parameters based on photometry (von Hippel et al., 2006; DeGennaro et al., 2009; van Dyk et al., 2009). Researchers have applied BASE-9 to key open clusters (DeGennaro et al., 2009; Jeffery et al., 2011; Hills et al., 2015), extended BASE-9 to study mass loss from the main sequence through the white dwarf stage, the so-called Initial Final Mass Relation (IFMR) (Stein et al., 2013), and have demonstrated that BASE-9 can derive the complex posterior age distributions for individual field white dwarf stars (O’Malley et al., 2013).

In this Chapter we focus on fitting the parameters of individual WD stars. The statistical model underlying BASE-9 relates a WD’s photometry to its parameters,

$$\mathbf{X}_i | A_i, \Theta_i \sim N_l \left(G(A_i, \Theta_i), \Sigma_i \right), \quad (4.1)$$

where, N_l represents a l -variate Gaussian distribution, $G(\cdot)$ represents the underlying astrophysical models that predicts a star’s photometric magnitudes as a function of its parameters. Specifically G combines models for the main sequence through red giant branch (e.g. Dotter et al., 2008) and the subsequent white dwarf evolution (e.g. Bergeron et al., 1995; Montgomery et al., 1999).

The Bayesian approach employed by BASE-9 requires a joint prior density on (A_i, Θ_i) for each WD. This prior is factored into

$$p(A_i, \Theta_i) = p(A_i|\mu_{A_i}, \sigma_{A_i})p(D_i|\mu_{D_i}, \sigma_{D_i})p(Z_i|\mu_{Z_i}, \sigma_{Z_i})p(M_i), \quad (4.2)$$

where $p(A_i|\mu_{A_i}, \sigma_{A_i}^2)$, $p(D_i|\mu_{D_i}, \sigma_{D_i}^2)$, $p(Z_i|\mu_{Z_i}, \sigma_{Z_i}^2)$ are normal densities each with its own prior mean (i.e., μ_{A_i} , μ_{D_i} , and μ_{Z_i}) and standard deviations (i.e., σ_{A_i} , σ_{D_i} , and σ_{Z_i}). When possible, these prior distributions are specified using recent external studies. The prior on the mass M_i is specified as the Initial Mass Function (IMF) taken from Miller and Scalo (1979), i.e., $\log_{10}(M_i) \sim N(-1.02, 0.67729^2)$. For a more detailed discussion of prior distribution in this research, refer to Section 4.2.1.

BASE-9 deploys an MCMC sampler to separately obtain a Monte Carlo sample from each of the WD's joint posterior distributions,

$$p(A_i, \Theta_i|\mathbf{X}_i) \propto p(\mathbf{X}_i|A_i, \Theta_i)p(A_i|\mu_{A_i}, \sigma_{A_i}) \times \\ p(D_i|\mu_{D_i}, \sigma_{D_i})p(Z_i|\mu_{Z_i}, \sigma_{Z_i})p(M_i). \quad (4.3)$$

In this manner, we can obtain case-by-case fits of A_i and Θ_i for each WD using BASE-9.

4.2 Hierarchical Modelling of a Group of White Dwarfs

As shown in Section 2.5.6, hierarchical models can outperform case-by-case analyses in terms of MSE. To take advantage of this, we embed the model in Eq. (4.1) into a hierarchical model for a sample of halo WDs. Because halo WDs have similar ages, we model the A_i via

$$A_i \sim N(\gamma, \tau^2). \quad (4.4)$$

In the hierarchical model A_i, D_i, Z_i , and M_i are the object-level parameters, while γ and τ are population-level parameters, the mean and standard deviation of the $\log_{10}$ ages of WDs in the galactic halo. For a more detailed description of parameters, see Table 4.1.

Here are some comments about the Gaussian assumption on $\log_{10}(\text{age})$ in Eq. (4.4). 1.) The true distribution of $\log_{10}(\text{age})$ for Galactic halo WDs is unknown. To use the Bayesian hierarchical method, we have to assume a distribution on A_i . The assumption of a common population incorporating an age constraint is the source of the statistical shrinkage that we will demonstrate. So we start off with a simple and commonly used Gaussian distribution. 2.) The main contributions of this research are efficient algorithms that fit a Bayesian hierarchical model of WDs when the case-by-case toolbox (in our case, the BASE-9) is available. Our algorithms will work when the population distribution of A_i becomes more complicated, though this requires a larger sample size than 10 WDs. 3.) From the astrophysics point of view, this assumption is idealised and it means that the star formation rate increases and decreases in a Gaussian manner. But in reality, halo WDs may arise from different accretion events, and more flexible distributions such as mixture Gaussians may be more reasonable to model their ages. My statistical techniques will work for flexible population distributions of ages. For now, with 10 halo WD candidates, a Gaussian distribution is appropriate because complicated distributions have more parameters, hence require larger data sets.

For the details on prior distributions of all parameters, refer to Section 4.2.1.

4.2.1 Prior Distributions

In this section, we discuss the prior distributions for parameters in both the case-by-case and hierarchical models. We aim to achieve objective results and also take

Table 4.1: Parameters in the hierarchical model

Object-level parameters	
A_i	$\log_{10}$ age of the WD i . This is the only parameter that is hierarchically modelled
D_i	distance modulus of WD i
Z_i	$\log_{10}$ solar normalised metallicity of WD i
M_i	mass of WD i
Population-level parameters	
γ	population mean $\log_{10}$ age for a group of WDs
τ	population standard deviation of $\log_{10}$ age for a group of WDs

advantage of best available information about these stars.

Table 4.2 displays the prior distributions used in this chapter. In the case-by-case analyses, we set uniform distribution on $\log_{10}(\text{age})$ of each WD to ensure the objectivity of the fits. The upper limit is the age of universe, 13.82 Gyr, i.e., 10.14 on the log base 10 scale. For the lower limit, we take 8.4 on the log base 10 scale, which is 251.19 Myr and an impossibly young age for halo WDs. Based on the photometric data of WDs, we have performed a sensitivity analysis on the lower limit of the uniform prior on A_i . Refer to Section 4.4.1 for more details. In our mathematical notations, sometimes we use $p(A_i|\mu_{A_i}, \sigma_{A_i})$ to denote the Gaussian prior on A_i with mean μ_{A_i} and standard deviation σ_{A_i} . This is because BASE-9 allows Gaussian prior distributions on all parameters. If a uniform distribution on A_i is required, we just set $\sigma_{A_i} = \infty$, then the Gaussian prior will be a uniform distribution. In the hierarchical model, we have Gaussian prior distribution $A_i \sim N(\gamma, \tau^2)$, where the γ and τ are unknown population-level parameters. We cover the prior distribution of γ and τ later in this section.

As to the metallicity parameter, Z_i , we usually employ very broad Gaussian prior distributions. For example, we usually set $\mu_{Z_i} = -1.5, \sigma_{Z_i} = 0.5$ to make it consistent for halo WDs. For the distance modulus, D_i , we use the best available knowledge

Table 4.2: Prior distributions for both object-level and population-level parameters

Parameters	Case-by-case Model	Hierarchical Model
A_i	$A_i \sim Unif(8.4, 10.14)$	$A_i \sim N(\gamma, \tau^2)$
D_i	$D_i \sim N(\mu_{D_i}, \sigma_{D_i}^2)$	Same as the left
Z_i	$Z_i \sim N(\mu_{Z_i}, \sigma_{Z_i}^2)$	Same as the left
M_i	$\log_{10}(M_i) \sim N(-1.02, 0.67729^2)$ truncated to range $0.1M_{\odot}$ and $8.0M_{\odot}$	Same as the left
γ	n.a.	$p(\gamma \tau) \propto 1$
τ	n.a.	$p(\tau) \propto 1$

from recent external studies about each WD. If no information available, we set a broad Gaussian prior on distance.

The prior on the mass M_i is chosen to be a version of initial mass function (IMF) taken from Miller and Scalo (1979), i.e., $\log_{10}(M_i) \sim N(-1.02, 0.67729^2)$. We obtain numerical constants from Table 7 and density function form in Eq. 30 in Miller and Scalo (1979). We truncate the initial mass of a WD to the range $0.1M_{\odot}$ and $8.0M_{\odot}$. The lower truncation is due to the fact that an initial mass of less than about $0.1M_{\odot}$ is not sufficient to initiate the fusing of hydrogen into helium necessary to form a star. The upper truncation is because the star clusters we study are sufficiently old that any stars with an initial mass above $8.0M_{\odot}$ would have used up their nuclear fuel long ago and become a neutron star or black hole, and thus would not be included in our observed data (van Dyk et al., 2009).

For the population-level parameters, γ and τ , we always take $p(\gamma, \tau) \propto p(\gamma|\tau)p(\tau)$ and $p(\gamma|\tau) \propto 1$. As discussed in Section 2.4, we follow the suggestion of Gelman (2006a) and assign a uniform prior on the standard deviation τ , namely, $p(\tau) \propto 1$, or equivalently, $p(\tau^2) \propto (\tau^2)^{-\frac{1}{2}}$.

4.2.2 Empirical Bayes Method

As in the Chapter 2 and 3, we can employ an EB approach to fit the hierarchical model in Eq. (4.1) and Eq. (4.4). One of the main benefits of our EB and FB routines is that they take advantage of the existing BASE-9 package, thus saving substantial time in code development. The details of our algorithms appear later. The joint posterior for parameters in the hierarchical model is

$$p(\gamma, \tau, A, \boldsymbol{\Theta} | \mathbf{X}) \propto p(\gamma, \tau) \prod_{i=1}^n p(\mathbf{X}_i | A_i, \boldsymbol{\Theta}_i) p(A_i | \gamma, \tau) \times \\ p(D_i | \mu_{D_i}, \sigma_{D_i}) p(Z_i | \mu_{Z_i}, \sigma_{Z_i}) p(M_i). \quad (4.5)$$

To avoid estimating τ as zero when its profile posterior distribution is highly skewed, we again implement EB on a transformed parameter. Here we consider two possibilities $\xi = \log \tau$ and $\delta = 1/\tau$. We denote the two resulting algorithms and their corresponding fits as Empirical Bayes with log transformation (EB-log) and Empirical Bayes with inverse transformation (EB-inv). EB alone refers to EB implemented without transforming τ . We compare the statistical properties of EB, EB-log, and EB-inv numerically in Section 4.3. For EB-log the joint posterior distribution can be written

$$p(\gamma, \xi, A, \boldsymbol{\Theta} | \mathbf{X}) \propto p(\gamma, \tau, A, \boldsymbol{\Theta} | \mathbf{X}) \cdot \frac{d\tau}{d\xi}. \quad (4.6)$$

The EB-log estimates are computed as follows.

Step 1: We obtain the MAP estimates for γ, ξ ,

$$(\hat{\gamma}, \hat{\xi}) = \max_{\gamma, \xi} \int \cdots \int p(\gamma, \xi, A, \boldsymbol{\Theta} | \mathbf{X}) dA d\boldsymbol{\Theta}, \quad (4.7)$$

i.e., via a Monte Carlo Expectation-Maximization (MCEM)-type algorithm

(Wei and Tanner, 1990). The Expectation-Maximization (EM) algorithm (e.g. Dempster et al., 1977) involves iteratively evaluating an integral over A and Θ , which MCEM accomplishes via Monte Carlo integration. Our MCEM-type algorithm iteratively calls BASE-9 to generate a Monte Carlo sample from $p(\Theta_i, A_i | \gamma^{(t)}, \xi^{(t)}, \mathbf{X}_i)$ for $i = 1, \dots, n$, where $\gamma^{(t)}$ and $\xi^{(t)}$ are the t -th iteration of γ and ξ within the MCEM algorithm. See the Appendix for the detailed algorithm.

Step 2: We compute $\hat{\tau} = \exp(\hat{\xi})$ and for each i use $p(A_i | \hat{\gamma}, \hat{\tau})$ as the prior distribution for A_i , when deriving the joint posterior distribution for A_i and Θ_i

$$p(A_i, \Theta_i | \mathbf{X}_i, \hat{\gamma}, \hat{\tau}) \propto p(\mathbf{X}_i | A_i, \Theta_i) p(A_i | \hat{\gamma}, \hat{\tau}) \times p(D_i | \mu_{D_i}, \sigma_{D_i}) p(Z_i | \mu_{Z_i}, \sigma_{Z_i}) p(M_i). \quad (4.8)$$

Step 3: Finally we use BASE-9 to obtain a Monte Carlo sample from $p(A_i, \Theta_i | \mathbf{X}_i, \hat{\gamma}, \hat{\tau})$, for each WD.

EB-inv proceeds in the same way, except $\frac{d\tau}{d\xi}$ is replaced by $\frac{d\tau}{d\delta}$ in Eq. (4.6), ξ is replaced by δ in the Eq. (4.7), and $\hat{\tau} = 1/\hat{\delta}$ in Step 2. For EB, set $\tau = \xi$ and $\frac{d\tau}{d\xi} = 1$, then all steps are the same as EB-log.

4.2.3 Fully Bayesian Method

The FB approach seeks to obtain a Monte Carlo sample from the joint posterior distribution in Eq. (4.5). In principle, this could be accomplished via the following Gibbs sampler:

Step 0: Initialize $(A_1^{(0)}, \dots, A_n^{(0)})$ at their fitted values under the case-by-case analyses in Eq. (4.3).

For $t = 1, 2, \dots$, run Step 1 and Step 2 iteratively.

Step 1: Sample the t -th iteration $\gamma^{(t)}$ and $\tau^{(t)}$ from their conditional posterior distribution

$$p(\gamma, \tau | A_1^{(t-1)}, \dots, A_n^{(t-1)}) \propto p(A_1^{(t-1)}, \dots, A_n^{(t-1)} | \gamma, \tau) \times p(\gamma, \tau). \quad (4.9)$$

Step 2: For each i , sample the t -th iteration $A_i^{(t)}$ and $\Theta_i^{(t)}$ from their conditional posterior $p(A_i, \Theta_i | \mathbf{X}_i, \gamma^{(t)}, \tau^{(t)})$ in Eq. (4.8). After an appropriate burnin period, $\{A_1^{(t)}, \Theta_1^{(t)}, \dots, A_n^{(t)}, \Theta_n^{(t)}, \gamma^{(t)}, \tau^{(t)}, t = 1, 2, \dots\}$ forms a correlated Monte Carlo sample from the posterior distribution.

Unfortunately, directly implementing Step 2 of this Gibbs sampler is very time-consuming, taking approximate 6 minutes for each star¹. To avoid iteratively calling BASE-9 within the Gibbs sampler, we consider the new sampler in Chapter 2 that replaces Step 2 of the Gibbs sampler with a Metropolis-Hastings (MH) step. The MH step uses the posterior distribution of (A_i, Θ_i) obtained under the n case-by-case analyses as its proposed distribution. Specifically, we run BASE-9 n times, once for each WD, to obtain a Monte Carlo sample from $\prod_{i=1}^n p(A_i, \Theta_i | \mathbf{X}_i)$, where the distributions inside the product are given in Eq. (4.3). To update $\mathbf{A}$ and $\mathbf{\Theta}$, during the Gibbs iteration we randomly select one realisation of this Monte Carlo sample and use a Metropolis-Hastings acceptance rule with target distribution $\prod_{i=1}^n p(A_i, \Theta_i | \mathbf{X}_i, \gamma^{(t)}, \tau^{(t)})$. To distinguish this new algorithm from the Gibbs sampler, we denote its sampled values with a tilde, $\{\tilde{A}_1^{(s)}, \tilde{\Theta}_1^{(s)}, \dots, \tilde{A}_n^{(s)}, \tilde{\Theta}_n^{(s)}, \tilde{\gamma}^{(s)}, \tilde{\tau}^{(s)}, s =$

¹We coded these steps in the R programming language and ran our code on a linux server with one core for each star. It took about 6 min to obtain one reliable draw from the posterior distribution. This is because BASE-9 must be reinitialised and retuned for each value of the prior distribution, as determined by $\gamma^{(t)}$ and $\tau^{(t)}$.

$1, 2, \dots\}$. It proceeds as follows.

Step 0a: For each WD run BASE-9 to obtain a Monte Carlo sample of $p(A_i, \Theta_i | \mathbf{X}_i)$ under the case-by-case analysis. Thin each chain to obtain an essentially independent Monte Carlo sample and label it $\{A_1^{(t)}, \Theta_1^{(t)}, \dots, A_n^{(t)}, \Theta_n^{(t)}, t = 1, 2, \dots, t_{MC}\}$.

Step 0b: Initialize each WD age at $\tilde{A}_i^{(1)} = A_i^{(1)}$ and the other parameters at $\tilde{\Theta}_i^{(1)} = \Theta_i^{(1)}$.

For $s = 1, 2, \dots$, we run Step 1 and Step 2 iteratively.

Step 1: Sample $\tilde{\gamma}^{(s)}$ and $\tilde{\tau}^{(s)}$ from $p(\gamma, \tau | \tilde{A}_1^{(s)}, \dots, \tilde{A}_n^{(s)})$.

Step 2: Randomly generate n integers on the range 1 to t_{MC} , $r_1, \dots, r_n$. For each i , set $A_i^* = A_i^{(r_i)}$, $\Theta_i^* = \Theta_i^{(r_i)}$ as the new proposal and accept $\tilde{A}^{(s+1)} = A_i^*$, $\tilde{\Theta}^{(s+1)} = \Theta_i^*$ with probability $\alpha = \min\{1, \frac{p(A_i^* | \tilde{\gamma}^{(s)}, \tilde{\tau}^{(s)})/p(A_i^* | \mu_{A_i}, \sigma_{A_i})}{p(\tilde{A}_i^{(s)} | \tilde{\gamma}^{(s)}, \tilde{\tau}^{(s)})/p(\tilde{A}_i^{(s)} | \mu_{A_i}, \sigma_{A_i})}\}$. Otherwise, set $\tilde{A}^{(s+1)} = \tilde{A}^{(s)}$, $\tilde{\Theta}^{(s+1)} = \tilde{\Theta}^{(s)}$.

Steps 1 and 2 are iterated until a sufficiently large Monte Carlo sample is obtained. The two-stage sampler is much faster than the Gibbs sampler. If a good sample from the case-by-case analysis is available, then it only takes a few minutes to obtain a Monte Carlo sample from the FB posterior distribution for the hierarchical model in Eq. (4.1) and Eq. (4.4). Unfortunately, two-stage sampler can be inefficient when dealing with hierarchical models on multi-dimensional parameters. In our analysis here only WD age is treated hierarchically, so the hierarchical part of the model is univariate. Fortunately, for this WD project, two-stage sampler is very efficient.

4.3 Simulation Study

To illustrate the performance of the various estimators of the object-level WD ages and the population-level parameters γ and τ , we perform a set of simulation stud-

Table 4.3: Distributions used in the simulation study.

Parameters	Distributions
$\log_{10}(\text{age})$	$N(\gamma = 10.09, \tau^2)$ truncated between 9.7 to 10.14
Distance modulus	$N(4.0, 2.5^2)$ truncated to interval (2.7, 5.7)
Metallicity	$N(-1.5, 0.02^2)$
Mass	$\log_{10}(\text{Mass}) \sim N(-1.02, 0.67729^2)$ truncated to (0.8, 3.0)

ies. Because the relative advantage of the shrinkage estimates compared with the case-by-case estimates depends both on the precision of the case-by-case estimates and the degree of heterogeneity of the object-level parameters, we repeat the simulation study under five settings, each with different values of observation error matrix Σ and population variance of $\log_{10}(\text{age})$, i.e., τ^2 . We simulate the parameters $\{A_i, D_i, T_i, M_i, i = 1, 2, \dots, N_1\}$ for each group of WDs from the distributions in Table 4.3, where $\gamma = 10.09$ (12.30 Gyr) is the population mean and τ varies among the simulation settings given in Table . For consistency with the data analyses in Section 4.4, we simulate u, g, r, i, z, J, H, K magnitudes for all WDs. Using BASE-9 for each setting, we simulate $N_2 = 25$ replicate datasets, each composed of $N_1 = 10$ halo WDs. For each WD in every group, we generate its $\log_{10}(\text{Age})$, distance modulus, metallicity and mass from distributions in Table 4.3, where τ is given in Table . The particular values and truncations in Table 4.3 and are chosen because they reflect plausible values for actual halo WDs.

We compute the empirical standard error for each magnitude by averaging the errors from the observed halo WDs in Section 4.4, and we denote by Σ_0 the variance matrix of observed magnitudes, i.e., the square of empirical standard errors for 8

magnitudes: u, g, r, i, z, J, H, K ,

$$\Sigma_0 = \begin{pmatrix} 0.304^2 & & & & & & & \\ & 0.092^2 & & & & & & \\ & & 0.027^2 & & & & & \\ & & & 0.026^2 & & & & \\ & & & & 0.068^2 & & & \\ & & & & & 0.062^2 & & \\ & & & & & & 0.086^2 & \\ & & & & & & & 0.083^2 \end{pmatrix}.$$

For simplicity in each setting all stars share the same diagonal observation variance, that is each $\Sigma_i = \Sigma, i = 1, 2, \dots, N_1$. The observation error variances for five simulation settings are described in terms of Σ_0 in Table . In the entire simulation study, we employ the Dotter et al. (2008) WD precursor models, Renedo et al. (2010) WD interior models, and WD atmospheres of Bergeron et al. (1995).

Subsequently we recover parameters with multiple approaches: EB, EB-log, EB-inv, FB and the case-by-case analysis. The case-by-case analysis in Eq. 4.1 fits WDs one-at-a-time, and infers a WD’s parameters based on its photometry. EB-type (EB, EB-log, EB-inv) methods and FB, however, based on the hierarchical model in Eq.s 4.1–4.4, fit all WDs simultaneously, yielding shrinkage estimates of object-level parameters. We specify non-informative broad prior distributions on each Θ_i , namely $D_i \sim N(4.0, 2.4^2)$, $Z_i \sim N(-1.5, 0.5^2)$ and IMF prior on M_i in Table 4.2. The case-by-case analyses require a prior distribution on each A_i and we use $A_i \sim Unif(8.4, 10.14)$ in Table 4.2. The hierarchical model uses Eq. (4.4) and requires priors on γ and τ , we again use the non-informative prior in Table 4.2. We compare the case-by-case estimates with shrinkage estimates based on the hierarchical model. Results from case-by-case analyses (in Eq. (4.1)) are indicted with a superscript I (for “individual”) and those from hierarchical analyses (in Eq. (4.1) and Eq. (4.4)) are indicted with

an H .

We compare estimates in terms of two criteria: i.) mean squared error (MSE) of point estimates; 2.) actual versus nominal coverage probabilities of interval estimates. When we require point estimates, we utilise posterior means from MCMC samples, and for interval estimates we use the 68.3% credible intervals from posterior distributions by finding the 15.8% and 84.1% quantiles from the MCMC samples.

We denote $\log_{10}(\text{age})$ of the i -th simulated WD in the j -th replicate group by A_{ij} . Using both the case-by-case and hierarchical analyses, we obtain MCMC samples of the $A_{ij}, i = 1, 2, \dots, N_1, j = 1, 2, \dots, N_2$. We estimate A_{ij} by taking the MCMC sample mean and denote the estimates based on case-by-case and hierarchical analyses by $\hat{A}_{ij}^I$ and $\hat{A}_{ij}^H$, respectively. We compute the absolute error² of each estimator $\hat{A}_{ij}$

$$\text{error}(\hat{A}_{ij}) = |\hat{A}_{ij} - A_{ij}|.$$

In our simulation study, we are mainly concerned with the difference between absolute errors from shrinkage and case-by-case estimates

$$\begin{aligned} \text{Diff}(A_{ij}) &= \text{error}(\hat{A}_{ij}^I) - \text{error}(\hat{A}_{ij}^H) \\ &= |\hat{A}_{ij}^I - A_{ij}| - |\hat{A}_{ij}^H - A_{ij}|, \end{aligned}$$

which compares the prediction accuracy of two methods. If $\text{Diff}(A_{ij}) \geq 0$, then the absolute deviation of the case-by-case estimate of star A_{ij} is greater than that of the shrinkage estimate.

Fig. 4.1 compares the performance of shrinkage estimates under simulated Setting I ($\tau = 0.05, \Sigma = \Sigma_0$). The corresponding summary plots for the other simulation set-

²We use this term to refer to the absolute value of the error

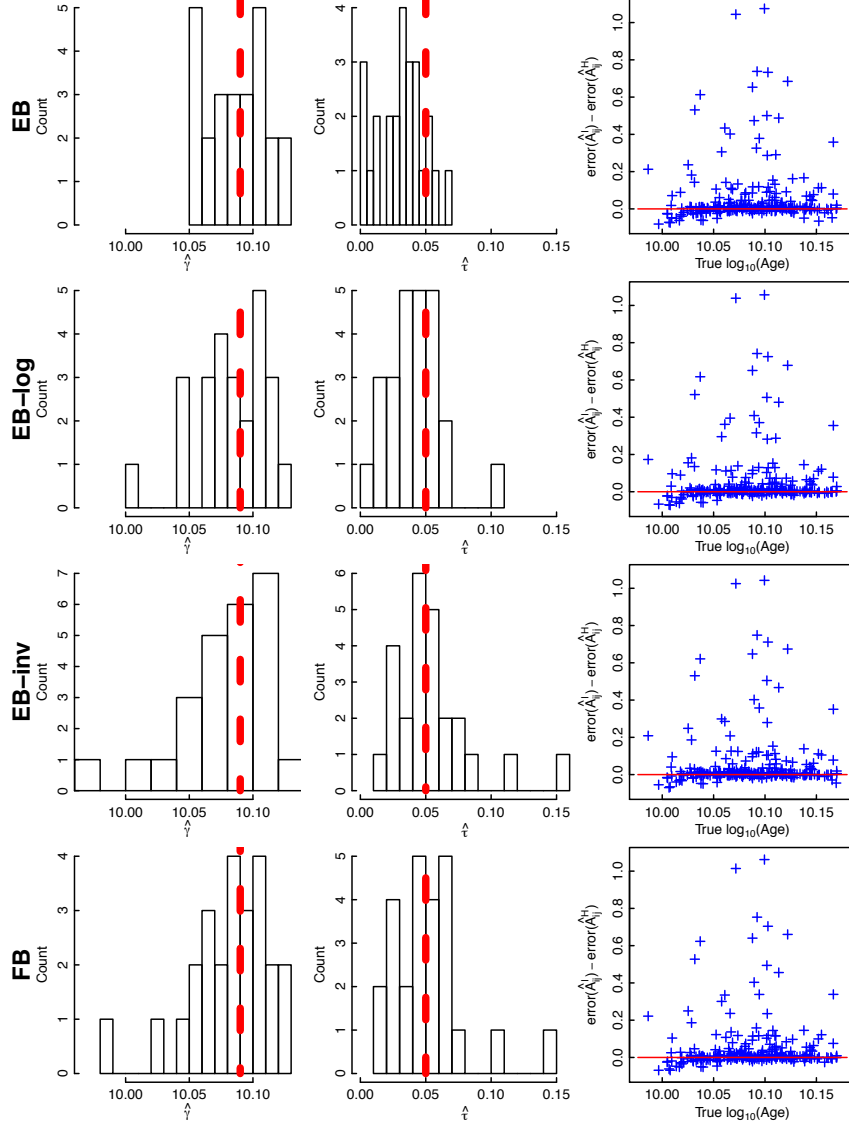


Figure 4.1: Results for Simulated Setting I, with $\tau = 0.05$ (standard deviation of $\log_{10}(\text{age})$) and $\Sigma = \Sigma_0$ (photometric error variances). The rows correspond to various fitting methods: EB, EB-log, EB-inv and FB, respectively. Each row includes: i) a histogram of the fitted values of $\hat{\gamma}$, across the $N_2 = 25$ simulation replicates, ii) a histogram of $\hat{\tau}$, and iii) a comparison of the difference in absolute errors in the fitted WD ages obtained with the case-by-case analysis and with each of the hierarchical methods. The thick dashed red lines in each histogram indicate the true values of γ or τ . In the right-most column, positive vertical values indicate larger error for the case-by-case fit, and the thin horizontal lines correspond to equal errors for the two methods.

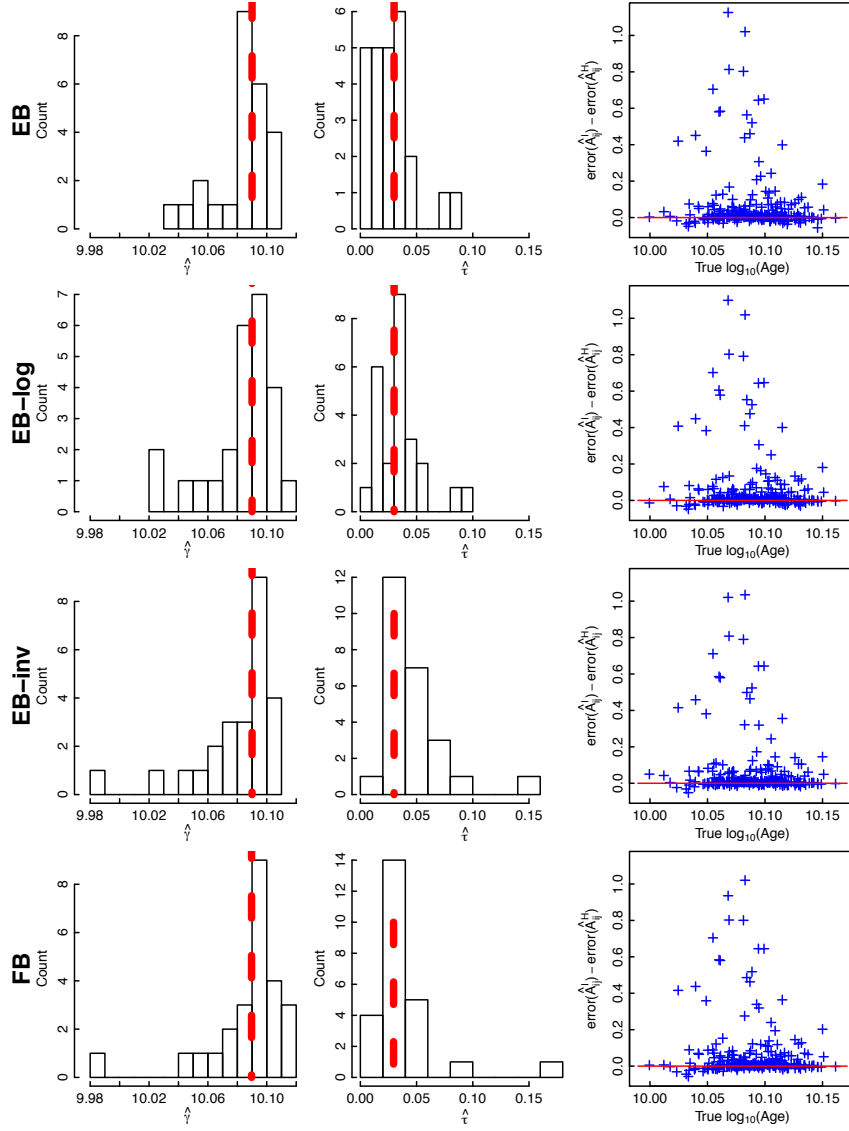


Figure 4.2: Similar to Fig. 4.1, but for the case $\tau = 0.03$ and $\Sigma = 0.8^2 \Sigma_0$.

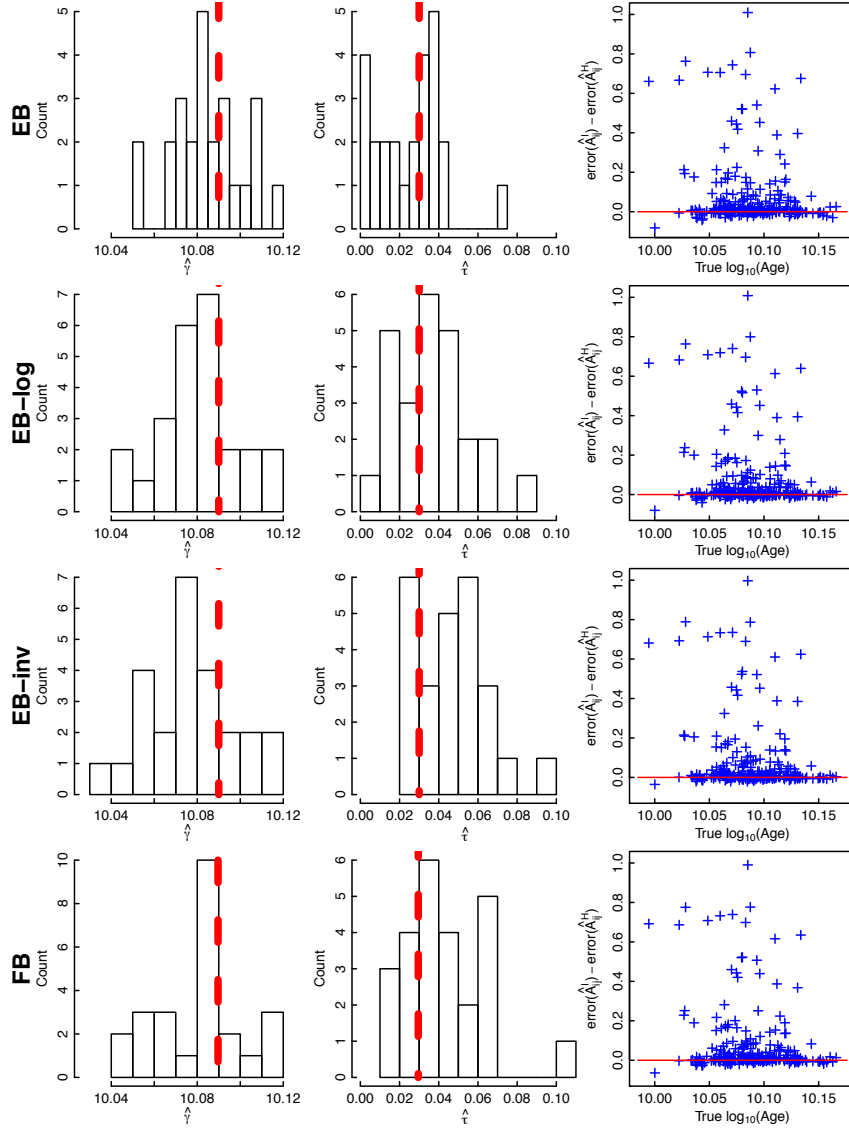


Figure 4.3: Similar to Fig. 4.1, but for the case $\tau = 0.03$ and $\Sigma = 1.2^2 \Sigma_0$.

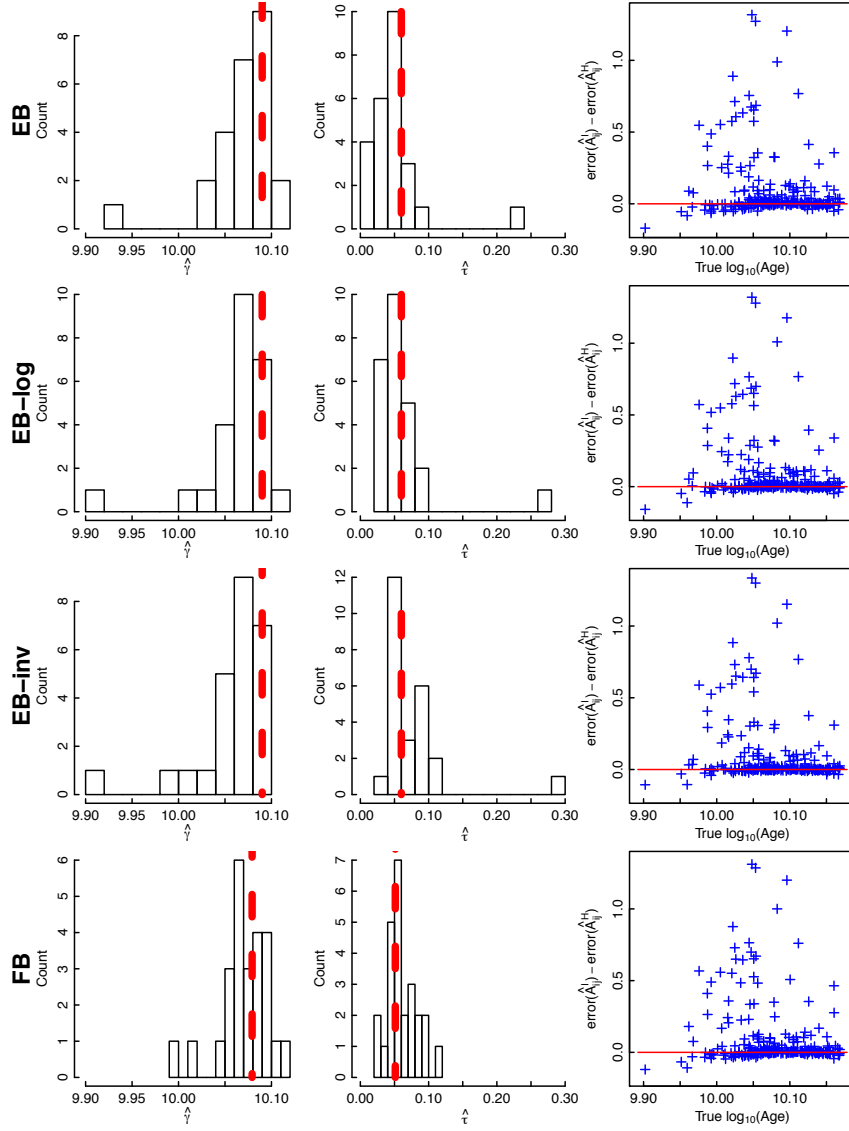


Figure 4.4: Similar to Fig. 4.1, but for the case $\tau = 0.06$ and $\Sigma = 0.8^2 \Sigma_0$.

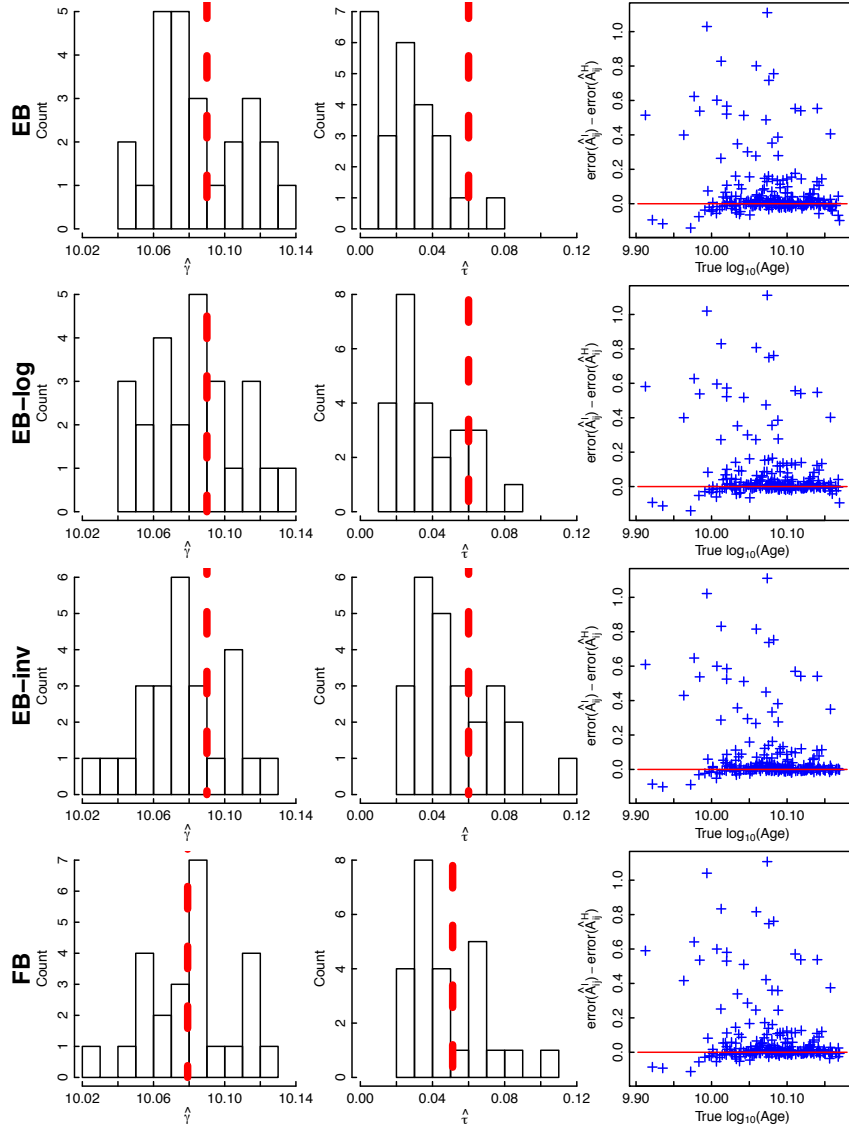


Figure 4.5: Similar to Fig. 4.1, but for the case $\tau = 0.06$ and $\Sigma = 1.2^2 \Sigma_0$.

tings are similar and appear in Figs 4.2–4.5. The histograms in Fig. 4.1 demonstrate that the estimates of γ and τ are generally close to their true values (thick, dashed red lines). Under all 5 settings, however, for some replicate groups of halo WDs, EB produces estimates of τ equal to 0 (see Fig. 3.1 and first row of Fig. 4.1). In these cases, the shrinkage estimate of the age of each WD in these groups converges to one value, which potentially leads to large errors. As we argued in Section 3.2, this highlights a difficulty with EB, and demonstrates the need for the transformed EB-log or EB-inv. Both of these approaches produce similar results to EB, but avoid the possibility of $\hat{\tau} = 0$.

The third column in Fig. 4.1, show the scatter plot of difference of errors between the case-by-case and hierarchical fits, $\text{Diff}(A_{ij})$, against A_{ij} , $i = 1, 2, \dots, N_1, j = 1, 2, \dots, N_2$. Each plus sign above the red line represents a simulated star obtaining a smaller error from the hierarchical fit than the case-by-case fit. Similarly, each plus sign below the red line indicates a star obtaining a larger error from the hierarchical fit than the case-by-case fit. Each plot has 250 simulated stars, and the percentages of stars above the red line are displayed in Table 4.4. Because most of the scatter in the plots are above the solid red zero line, the estimates of $\log_{10}(\text{age})$ from the case-by-case analyses tend to be further from the true values than the shrinkage estimates. Approximately two thirds of the $N_1 \times N_2 = 250$ simulated stars in each setting are better estimated with the shrinkage method than the case-by-case fit. For stars below the red solid lines, nominally the case-by-case fit is better, but the advantage is small. In fact, for almost all simulated stars, $\text{Diff}(A_{ij}) = \text{error}(A_{ij}^I) - \text{error}(A_{ij}^H) > -0.1$, so the shrinkage estimates do not perform much worse than case-by-case estimates for any WD and often perform much better. For some stars, we have $\text{Diff}(A_{ij}) > 0.5$. From the point of view of reliability of the technique, it is comforting that the four hierarchical fits (EB, EB-log, EB-inv, FB) perform similarly, at least when $\hat{\tau} > 0$ for

EB.

Fig. 4.4 and 4.5 display performances of various methods under simulated setting IV ($\tau = 0.06$, $\Sigma = 0.8^2\Sigma_0$) and V ($\tau = 0.06$, $\Sigma = 0.8^2\Sigma_0$), respectively. From Fig. 4.4, the EB tends to underestimate τ when the photometric error is $0.8^2\Sigma_0$. This situation worsens when photometric error increases to $1.2^2\Sigma_0$, as shown in Fig. 4.5. This problem is caused by the shrinkage effects of the hierarchical model. With the increase of photometric error Σ , the ratio of Σ to τ rises, so the shrinkage effect will grow, which means the hierarchical model will drag the fitted object-level parameters closer to each other, i.e., less variation between the fitted ages. Thus the point estimates of τ become smaller in setting V. In fact, all methods exhibit this effect. Their point estimates of τ tend to be smaller when Σ increases from $0.8^2\Sigma_0$ to $1.2^2\Sigma_0$.

Table 4.4 shows the percentage of WDs having closer estimates to their true ages in each simulated setting. In other words, Table 4.4 presents the percentage of plus signs above the red line in Figs 4.1–4.5. In each simulated setting, we have $N_2 = 25$ groups of WDs, each comprising $N_1 = 10$ WDs. So the mathematical formula of percentage figures in Table 4.4 is

$$\text{Percentage} = \frac{1}{N_1 \times N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} I(|\hat{A}_{ij}^I - A_{ij}| > |\hat{A}_{ij}^H - A_{ij}|),$$

where

$$I(|\hat{A}_{ij}^I - A_{ij}| > |\hat{A}_{ij}^H - A_{ij}|) = \begin{cases} 1, & \text{If } |\hat{A}_{ij}^I - A_{ij}| > |\hat{A}_{ij}^H - A_{ij}|, \\ 0, & \text{Otherwise.} \end{cases}$$

Table 4.4 reports the percentage of simulated WDs that are better estimated by shrinkage methods than the case-by-case fits based on 250 simulated WDs under each setting. We conclude that 60%–75% of simulated stars have a more reliable estimate of their ages from the hierarchical analyses than from the case-by-case analyses.

Table 4.4: The percentage of WDs for which EB and FB improved age estimates over case-by-case fits among 250 simulated WDs under the five simulated settings. In other words, these actual figures are the percentages of plus signs above the red lines in Figs 4.1–4.5.

Simulation Settings		EB	EB-log	EB-inv	FB
$\tau = 0.05$	$\Sigma = \Sigma_0$	65.6%	69.6%	68.8%	66.8%
$\tau = 0.03$	$\Sigma = 0.8^2 \Sigma_0$	74.4%	72.8%	76.4%	73.2%
	$\Sigma = 1.2^2 \Sigma_0$	67.6%	69.6%	70.4%	72.8%
$\tau = 0.06$	$\Sigma = 0.8^2 \Sigma_0$	65.6%	67.6%	68.0%	64.4%
	$\Sigma = 1.2^2 \Sigma_0$	60.4%	64.0%	66.8%	63.2%

Table 4.5–4.6 present a numerical comparison of the shrinkage and case-by-case estimates of the object-level parameter $\log_{10}(\text{age})$ and population-level parameters γ and τ . Specifically it presents the average Mean Absolute Error (MAE) and the average root of the mean squared error (RMSE) for each method, i.e.,

$$\begin{aligned}
 \text{MAE}(A) &= \frac{1}{N_1 \cdot N_2} \sum_{j=1}^{N_2} \sum_{i=1}^{N_1} |\hat{A}_{ij} - A_{ij}|, \\
 \text{RMSE}(A) &= \sqrt{\frac{1}{N_1 \cdot N_2} \sum_{j=1}^{N_2} \sum_{i=1}^{N_1} (\hat{A}_{ij} - A_{ij})^2}, \\
 \text{RMSE}(\gamma) &= \sqrt{\frac{1}{N_2} \sum_{j=1}^{N_2} (\hat{\gamma}_j - \gamma_{\text{true}})^2}, \\
 \text{RMSE}(\tau) &= \sqrt{\frac{1}{N_2} \sum_{j=1}^{N_2} (\hat{\tau}_j - \tau_{\text{true}})^2}.
 \end{aligned}$$

Table 4.5 summarises the median absolute errors (MAEs) and root mean squared errors (RMSEs) of different estimates of ages ($\log_{10}$ scale) of WDs under the five simulated settings. In terms of estimates of ages of simulated WDs, from Table 4.5 all four shrinkage methods (EB, EB-log, EB-inv, FB) have significantly smaller MAEs and RMSEs than their case-by-case counterparts, though there are slight difference

Table 4.5: The MAE and RMSE of the various shrinkage and case-by-case estimates of ages ($\log_{10}$ scale) of WDs under five simulated settings.

Simulation Settings	EB		EB-log		EB-inv		FB		Case-by-case	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
I: $\tau = 0.05, \Sigma = \Sigma_0$	0.028	0.038	0.029	0.042	0.033	0.052	0.032	0.047	0.084	0.18
II: $\tau = 0.03, \Sigma = 0.8^2 \Sigma_0$	0.024	0.035	0.025	0.038	0.027	0.044	0.025	0.041	0.086	0.19
III: $\tau = 0.03, \Sigma = 1.2^2 \Sigma_0$	0.023	0.031	0.024	0.034	0.026	0.037	0.025	0.035	0.099	0.20
IV: $\tau = 0.06, \Sigma = 0.8^2 \Sigma_0$	0.038	0.063	0.038	0.070	0.040	0.074	0.034	0.049	0.12	0.26
V: $\tau = 0.06, \Sigma = 1.2^2 \Sigma_0$	0.033	0.046	0.032	0.044	0.032	0.045	0.032	0.044	0.10	0.22

among the four shrinkage estimates. Generally speaking, estimates of ages from case-by-case analyses have errors about three times those errors that shrinkage estimates have.

Simulation Setting III ($\tau = 0.03, \Sigma = 1.2^2 \Sigma_0$) benefits most from the shrinkage estimates in terms of reduced RMSE. The RMSE from the case-by-case fits under Simulation Setting III is approximately 0.20, while the RMSE from the EB is around 0.031, less than one sixth of the former. Simulation Setting IV ($\tau = 0.06, \Sigma = 0.8^2 \Sigma_0$) gains least from the shrinkage estimates (0.063); its RMSE is approximately a quarter of the case-by-case RMSE (0.26). Generally, when Σ is large and τ is small, the advantage of shrinkage estimates is the greatest. With small Σ and large τ , the advantage of shrinkage estimates over case-by-case estimates decreases. This is consistent to statistical theory (see Gelman et al., 2014, Chapter 5). When Σ is small, the photometric data of WDs have less uncertainty, so their fitted ages will be more precise. Therefore, there is less latitude for sharing information across WDs to improve the estimates. On the other hand, when τ is large, then WDs have larger variation between their ages. Because the number of WDs in each group is fixed at 10, when the variation of their ages grows, they tend to be more distant from each other. In other words, in this case the WDs will have less in common. Therefore, there is less advantage is using the hierarchical model.

Table 4.6 demonstrates the RMSEs of multiple estimates of population-level parameters γ and τ from the hierarchical model. Because the case-by-case analysis does not pool WDs into groups, so it fails to produce estimates of population-level parameters and it is excluded in Table 4.6. Because EB can produce $\tau = 0$, so we introduce its variants EB-log and EB-inv to combat this kind of situation. The performances of these four methods are quite similar and EB-inv has the largest RMSE in γ under these five simulated settings. Case IV: $\tau = 0.06, \Sigma = 0.8^2 \Sigma_0$ sees the

Table 4.6: The RMSE of the various hierarchical estimates of population-level parameters under five simulated settings.

Simulation Settings	EB		EB-log		EB-inv		FB	
	γ	τ	γ	τ	γ	τ	γ	τ
I: $\tau = 0.05, \Sigma = \Sigma_0$	0.023	0.025	0.028	0.023	0.037	0.029	0.033	0.028
II: $\tau = 0.03, \Sigma = 0.8^2 \Sigma_0$	0.02	0.02	0.023	0.022	0.029	0.033	0.028	0.034
III: $\tau = 0.03, \Sigma = 1.2^2 \Sigma_0$	0.017	0.017	0.021	0.019	0.024	0.024	0.022	0.024
IV: $\tau = 0.06, \Sigma = 0.8^2 \Sigma_0$	0.041	0.044	0.045	0.044	0.051	0.05	0.032	0.021
V: $\tau = 0.06, \Sigma = 1.2^2 \Sigma_0$	0.025	0.039	0.026	0.029	0.027	0.022	0.026	0.023

worst performance of EB-type approaches, having largest RMSEs among these five settings. This is reasonable, because observed data have smaller variance $0.8^2 \Sigma_0$ but WD-by-WD ages have large variation 0.06, so the shrinkage estimates have less benefit. FB has consistent performance through Case I to V, so it is more stable than EB-type methods in terms of estimation of population-level parameters.

Table 4.7 presents the actual coverage probabilities of 68.3% credible intervals estimates for ages of simulated WDs derived from five methods: EB, EB-log, EB-inv, FB and case-by-case. The 68.3% confidence intervals of the coverage probabilities are computed with the Clopper-Pearson exact method (Clopper and Pearson, 1934). From Table 4.7, 68.3% credible intervals obtained from the EB have coverage probabilities between 42% to 58%, far below the nominal level. One explanation for the poor performance of EB is that it shrinks WD ages too much and it underestimates τ , as shown through 4.1–4.5. When the posterior distribution of τ peaks at 0, EB produces the same age for all WDs in that group. The EB-log algorithm performs slightly better than EB, between 54.8% and 69.2%, but still under four simulated settings its coverage probabilities fall below the nominal value 68.3%. This is because EB-log avoids the case that the posterior of τ peaks at 0, which increases the estimate of τ . The EB-inv algorithm yields a larger estimate of τ than EB-log, so it tends to shrink the ages of WDs less. Table 4.7 shows that EB-inv produces credible intervals that have actual coverage probabilities between 61.6% to 73.2%, which are quite close to the nominal level 68.3%. The case-by-case infers WD ages individually, and the coverage probabilities of its credible intervals are between 60% and 64.8%, slightly under the nominal value. FB performs the best, its actual coverage probabilities varying from 63.6% to 74%, very close to the nominal level.

Table 4.8 presents the coverage probabilities of 68.3% credible intervals for population-level parameters γ and τ obtained via the FB method. EB-type methods do not

Table 4.7: The coverage probabilities (CP) of 68.3% credible intervals of the estimates of ages ($\log_{10}$ scale) of WDs and underneath are their 68.3% confidence intervals (CI).

Simulation Settings	EB		EB-log		EB-inv		FB		Case-by-case	
	CP	(68.3% CI)	CP	(68.3% CI)	CP	(68.3% CI)	CP	(68.3% CI)	CP	(68.3% CI)
I: $\tau = 0.05, \Sigma = \Sigma_0$	0.54	(0.506, 0.573)	0.564	(0.53, 0.597)	0.616	(0.583, 0.648)	0.636	(0.603, 0.668)	0.604	(0.571, 0.636)
II: $\tau = 0.03, \Sigma = 0.8^2 \Sigma_0$	0.456	(0.423, 0.49)	0.564	(0.53, 0.597)	0.684	(0.652, 0.715)	0.672	(0.64, 0.703)	0.636	(0.603, 0.668)
III: $\tau = 0.03, \Sigma = 1.2^2 \Sigma_0$	0.52	(0.486, 0.553)	0.648	(0.615, 0.68)	0.708	(0.676, 0.738)	0.74	(0.709, 0.769)	0.648	(0.615, 0.68)
IV: $\tau = 0.06, \Sigma = 0.8^2 \Sigma_0$	0.58	(0.546, 0.613)	0.692	(0.66, 0.722)	0.732	(0.701, 0.761)	0.724	(0.693, 0.753)	0.6	(0.567, 0.633)
V: $\tau = 0.06, \Sigma = 1.2^2 \Sigma_0$	0.424	(0.391, 0.458)	0.548	(0.514, 0.581)	0.66	(0.627, 0.691)	0.668	(0.636, 0.699)	0.628	(0.595, 0.66)

Table 4.8: FB coverage probabilities (CP) of 68.3% credible intervals for estimates of population-level parameters and their 68.3% confidence intervals (CI).

Simulation Settings	γ	τ
	CP (68.3% CI)	CP (68.3% CI)
I: $\tau = 0.05, \Sigma = \Sigma_0$	0.52 (0.402, 0.636)	0.64 (0.52, 0.747)
II: $\tau = 0.03, \Sigma = 0.8^2 \Sigma_0$	0.64 (0.52, 0.747)	0.8 (0.687, 0.884)
III: $\tau = 0.03, \Sigma = 1.2^2 \Sigma_0$	0.68 (0.561, 0.782)	0.84 (0.731, 0.915)
IV: $\tau = 0.06, \Sigma = 0.8^2 \Sigma_0$	0.64 (0.52, 0.747)	0.84 (0.731, 0.915)
V: $\tau = 0.06, \Sigma = 1.2^2 \Sigma_0$	0.56 (0.441, 0.674)	0.6 (0.48, 0.711)

directly produce credible intervals on γ and τ , so they are not included in this table. The 68.3% confidence intervals of the coverage probabilities are obtained with the Clopper-Pearson exact method (Clopper and Pearson, 1934). Under these five simulated cases, the FB credible interval estimates of γ are under-covered with actual coverage between 52% and 68%, smaller than its nominal value 68.3%. On the other hand, the FB credible intervals on τ have actual coverage between 60% and 84%, and tend to be over its nominal coverage probabilities. The difficulty of the hierarchical method in coverage property is due to the small number, $I = 10$, of WDs in the model. Other research have also reported and discussed this problem, refer to Browne and Goldstein (2002); Browne et al. (2006) for more details.

We present the CPU time consumption report of different algorithms in Table 4.9. These CPU times are based on simulation studies with BASE-9, where WDs have photometric magnitudes in eight filters u, g, r, i, z, J, H, K . The CPU time of all approaches vary subject to a few factors, such as, the number of photometric magnitudes for each WD, the sample size, etc. All methods take more CPU time if

Table 4.9: CPU Times of different methods

Algorithms	CPU Times
Case-by-case Analysis	About 5 hours for 10000 draws
Hierarchical Methods	
Two-stage FB	About 5 hours and 10 minutes for 10000 draws
EB/EB-log/EB-inv	About 20 hours for 10000 draws
Gibbs	About 40 days for 10000 draws

the WDs have more photometric magnitudes. When implementing these algorithms, I always do parallel computing on multiple WDs using the “doParallel” package in R³, so that it takes one unit of time to fit multiple WDs with BASE-9. The case-by-case analysis takes about 5 hours for 10000 samples. FB with the Two-stage sampler takes advantage of samples from case-by-case fits, and it only costs an additional 10 minutes to produce estimates from the hierarchical model. All three EB-type methods (EB, EB-log and EB-inv) require similar amounts of CPU time, approximately 20 hours. EB-type methods take longer because they do not use the samples from the case-by-case analyses. The FB with Gibbs sampler takes the longest time, approximately 40 days because it calls the black-box code (in this case, the BASE-9 package) during each iteration, which requires significant time to explore the conditional posterior distribution for each iteration. The Gibbs is only included to check the accuracy of the two-stage FB fit. I have verified that the MCMC samples from Gibbs sampler is the same as those from the two-stage algorithm that only costs 10 minutes. Therefore, in the WD and IFMR project, I always use the two-stage algorithm to fit the hierarchical models.

From Tables 4.4–4.9, we conclude that shrinkage estimates from both EB-type and FB approaches outperform the case-by-case analyses in terms of smaller RMSE and MAE. Under the five simulated settings, all four computational methods—EB,

³R programming language is the most popular software in the statistical community.

EB-log, EB-inv and FB- behave similarly. Their MAEs and RMSEs are comparable. In terms of coverage properties of 68.3% credible intervals from these methods, the FB approach is the best. EB and EB-log produce credible intervals under their nominal coverage probability. EB-inv yields credible intervals over the nominal value. In terms of computing time, our FB algorithm is less time-consuming than all of our EB algorithms. As discussed in Section 4.2.3, however our FB algorithm can fail when the case-by-case sample cannot represent the case-by-case posterior distribution whereas the EB algorithms tend to be more robust.

Overall, we suggest readers use FB and EB-log approaches to fit hierarchical models that have complicated or computer-based likelihood functions. Our two-stage algorithms for FB and EB-log are very efficient in fitting the hierarchical models. In the next section, we implement FB and EB-log to analyse a group of Galactic halo WDs.

4.4 Data Analysis

This analysis is just a testbed of our statistical techniques, and the astronomical results from our analysis are not definitive. The target of this section is to present how our Bayesian hierarchical methods can be beneficial to the study of ages of WDs in astronomy.

Now we turn to the hierarchical and case-by-case analysis of the 10 field WDs from the Galactic halo listed in Table 4.10. It lays the ground work for large-scale studies using data from the Gaia mission. We acquire prior densities on $M_i, D_i, Z_i, i = 1, \dots, 10$ from the literature (Kilic et al., 2010; Kilic et al., 2012; Gianninas et al., 2015; Dame et al., 2016). The references and priors on distance modulus for these 10 stars are listed in Table 4.10.

Table 4.10: Prior distributions for distance moduli and atmosphere composition for halo WD sample

White Dwarf	Dist. Mod. ($V - M_V$)	Comp.	Reference
J0301−0044	$N(4.35, 0.2)^a$	He	Paper III ^d
J0346+246	$N(2.24, 0.33)^a$	H	Paper II ^c
J0822+3903	$N(5.19, 2.4)$	H	Paper IV ^e
J1024+4920	$N(5.54, 2.4)$	He	Paper IV ^e
J1102+4113	$N(2.64, 0.13)^a$	H	Paper II ^c
J1107+4855	$N(3.41, 0.19)^a$	H	Paper III ^d
J1205+5502	$N(5.04, 2.4)$	He	Paper IV ^e
J2137+1050	$N(4.46, 2.4)$	H	Paper I ^b
J2145+1106N	$N(4.19, 2.4)$	H	Paper I ^b
J2145+1106S	$N(4.23, 2.4)$	H	Paper I ^b

^a Precise distance moduli are from trigonometric parallax measurements.

^b Paper I is Kilic et al. (2010).

^c Paper II is Kilic et al. (2012).

^d Paper III is Gianninas et al. (2015).

^e Paper IV is Dame et al. (2016).

4.4.1 Case-by-case Analysis

For the case-by-case fitting of each WD, we employ a minimally informative flat prior on A_i , that is, $A_i \sim \text{Unif}(8.4, 10.14)$, a ZAMS mass prior IMF from Miller and Scalo (1979) on M_i and a diffuse prior on metallicity $Z_i \sim \text{N}(-1.50, 0.5^2)$. The lower bound 8.40 for the prior of log base 10 age is 251 Myr and it is the lower bound of age of WDs in the BASE-9 package. The upper bound for A_i is 10.14, the widely accepted age of the Universe, 13.82 Gyr. For details of prior distributions, refer to 4.2.1

First we show the case-by-case fits of these 10 WDs and before the end of this section we present the sensitivity analysis of the case-by-case fits to the lower bound of the uniform prior of A_i .

We derive the joint posterior density for the parameters using Bayes' theorem:

$$p(A_i, D_i, Z_i, M_i | \mathbf{X}_i) \propto p(\mathbf{X}_i | A_i, D_i, Z_i, M_i) p(A_i) p(D_i) p(Z_i) p(M_i). \quad (4.10)$$

Before specifying a hierarchical modelling for the 10 WDs, we can obtain their case-by-case fits via BASE-9 (as in O'Malley et al., 2013). By using the priors in Table 4.10 and described above, we fit the 10 halo WDs individually with BASE-9. We present 5 typical stars in Fig. 4.6. Each column in Fig. 4.6 is corresponding to one WD. The rows each give two dimensional marginal density functions of the posterior distributions. The asymmetric errors in the fitted parameters, including age, are evident. The plot shows that the uncertainty of age for these five stars is large, ranging from 8 Gyrs to approximately 14 Gyrs. For more detailed case-by-case results of 10 WDs, classical "triangle" plots for these WDs are shown in Figs A.1–A.10 in Appendix. Assuming each of these ten WDs are bona fide Galactic halo members,

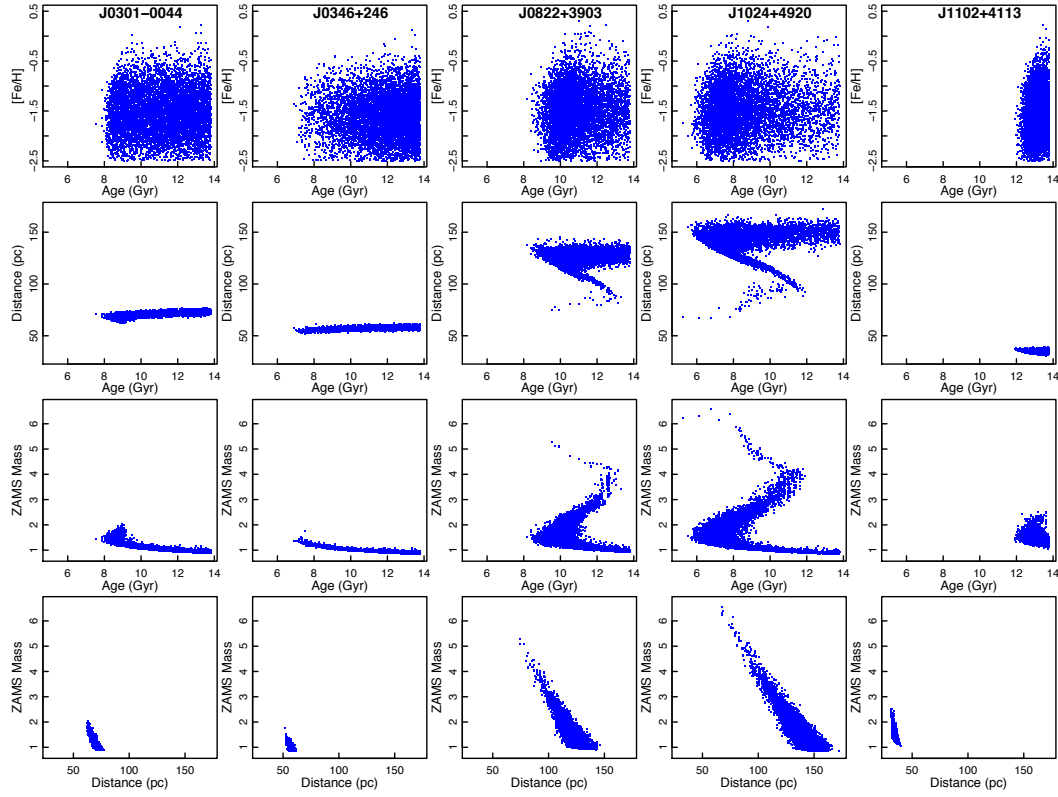


Figure 4.6: Projections of the joint posterior distributions onto the two dimensional planes (from top to bottom) of age–metallicity, age–distance, age–ZAMS mass, and distance–ZAMS mass planes for five of the Galactic halo WDs. The posterior distribution are computed under the case-by-case fits.

we expect their ages to be similar. However, their masses, distance moduli and metallicities may vary substantially. In this case, we can deploy hierarchical modelling only on the ages of these 10 WDs, which provides substantial additional information.

We investigate whether the case-by-case fits is sensitive to the choice of lower bound in the uniform prior distribution of A_i . We set the lower bound prior of A_i to be 8.6, 8.8, 9.0 and 9.2, rerun the black-box code (BASE-9) on these 10 WDs, compare the MCMC draws from their case-by-case posterior distribution in Eq. 4.10. We compare these new fits with the original one when the lower bound is 8.4 by quantile-quantile plots.

Figs 4.7–4.10 presents the quantile-quantile plots of MCMC samples of $\log_{10}(\text{age})$ of 10 WDs based on different lower bounds in the prior distribution of $\log_{10}(\text{age})$. Most circles in these QQ plots lie quite close to the red line, although there is some deviation from the red line in several plots. These deviations are caused by Monte Carlo errors, accounting for only very tiny part (less than 1%) of the distributions. Overall, it is reasonable to conclude that the distributions on the x and y axis are the same. Therefore, the posterior distributions in the case-by-case fits are insensitive to the lower bound in the prior distribution of $\log_{10}(\text{age})$.

4.4.2 Hierarchical Analysis

Here we deploy both EB-log and FB to obtain fits of the hierarchical model in Eq. (4.1) and Eq. (4.4) based on those ten candidate Galactic halo WDs. We present hierarchical analysis in this section. In Fig. 4.11, we compare the posterior density functions of the age of each WD fit with the case-by-case method and with both EB-log and FB.

Fig. 4.11 demonstrates that the posterior distributions of the age under the hierarchical model – both EB-log and FB – peak near a sensible halo age, whereas the

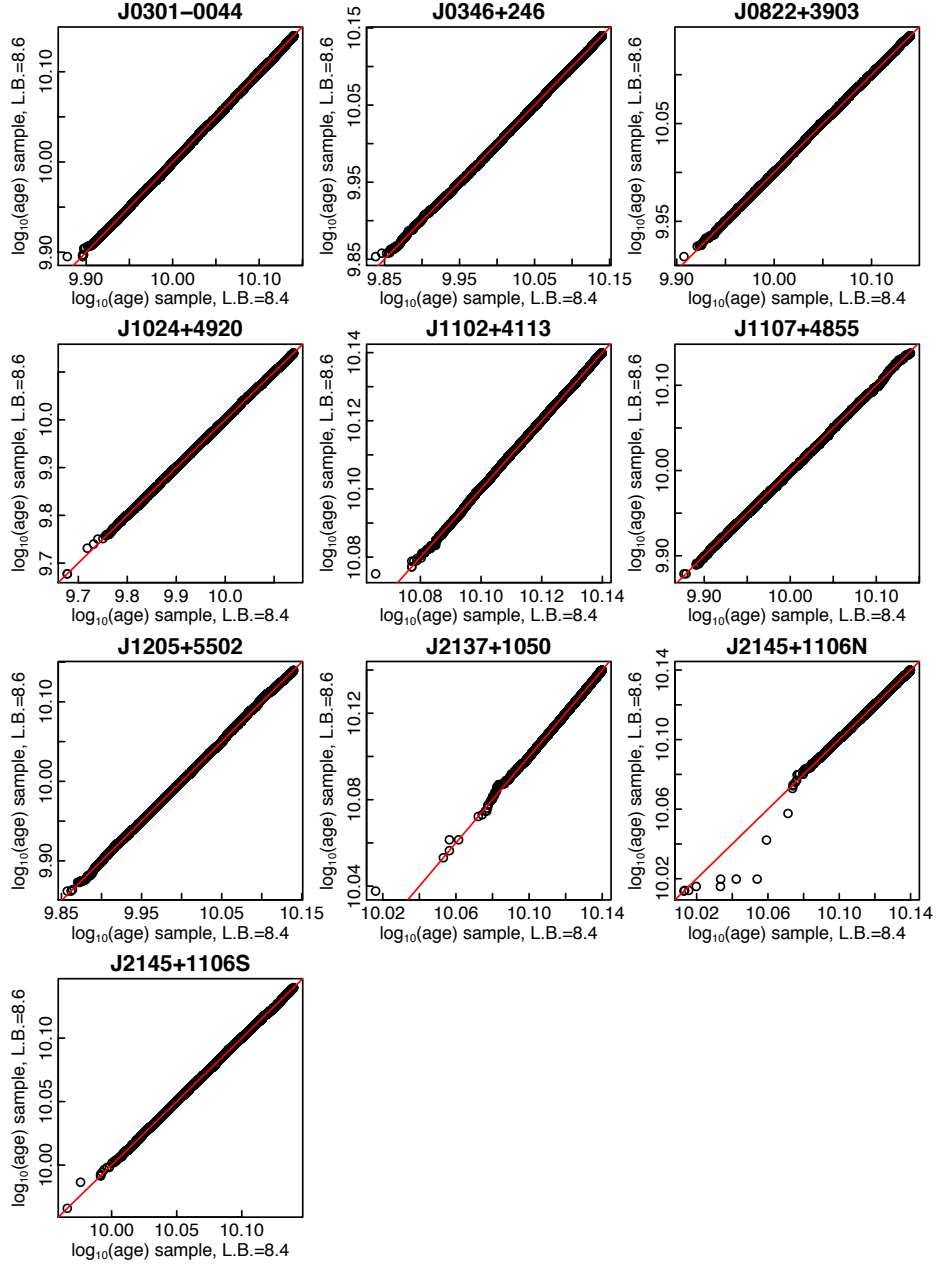


Figure 4.7: Quantile-quantile plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 8.6. The red line represents that quantiles from samples in x and y axes are equal.

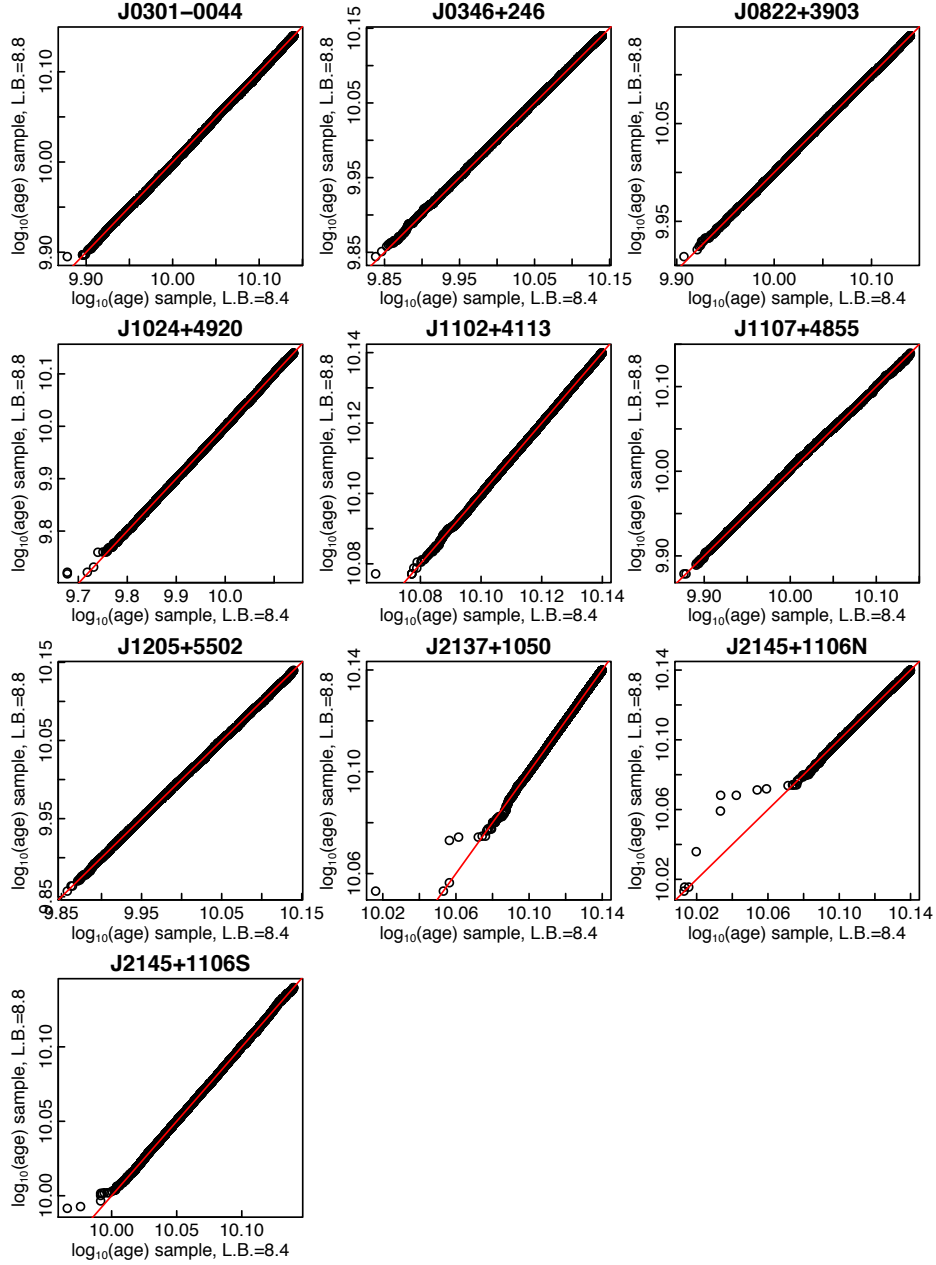


Figure 4.8: Quantile-quantile plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 8.8. The red line represents that quantiles from samples in x and y axes are equal.

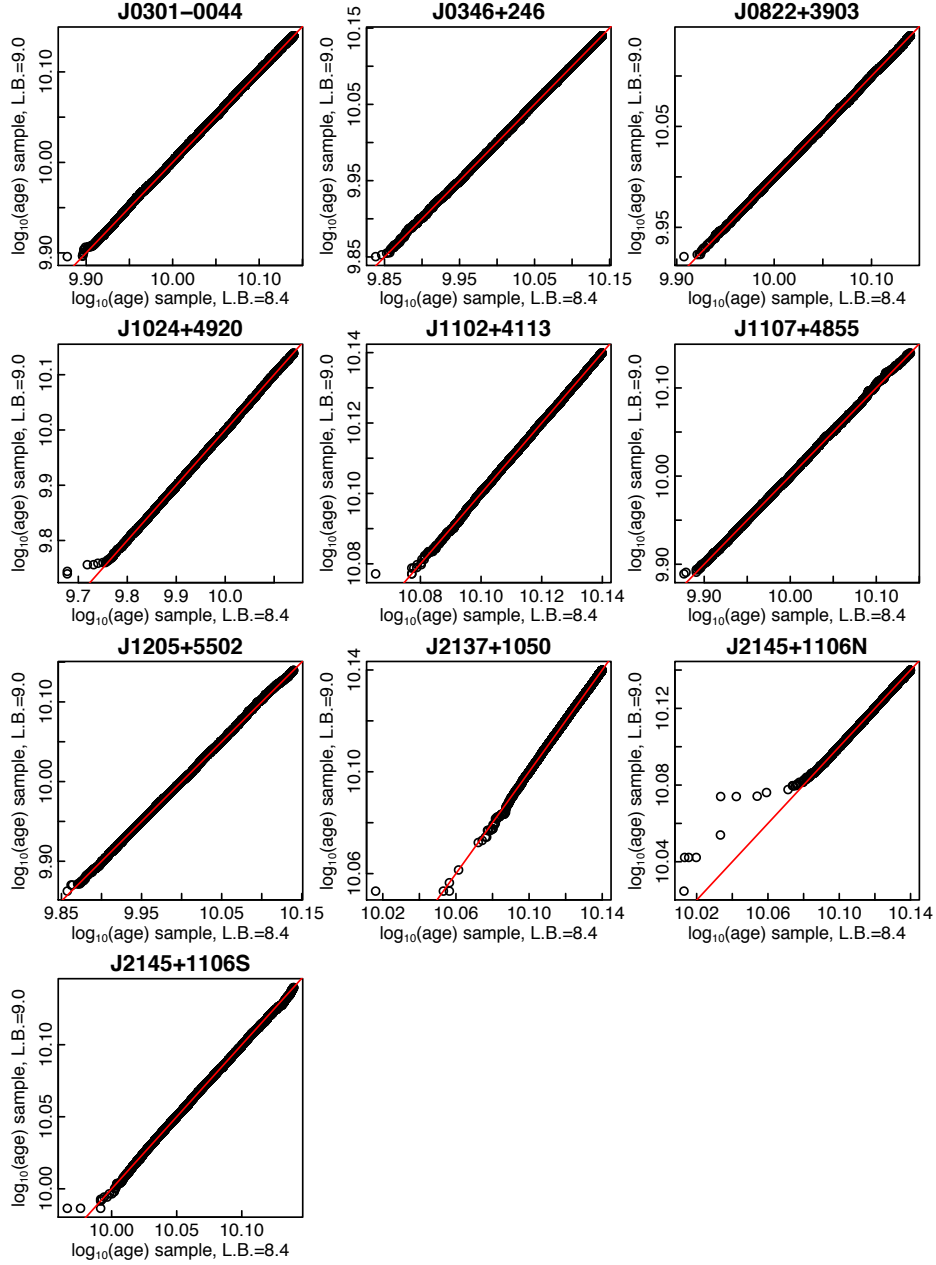


Figure 4.9: Quantile-quantile plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 9.0. The red line represents that quantiles from samples in x and y axes are equal.

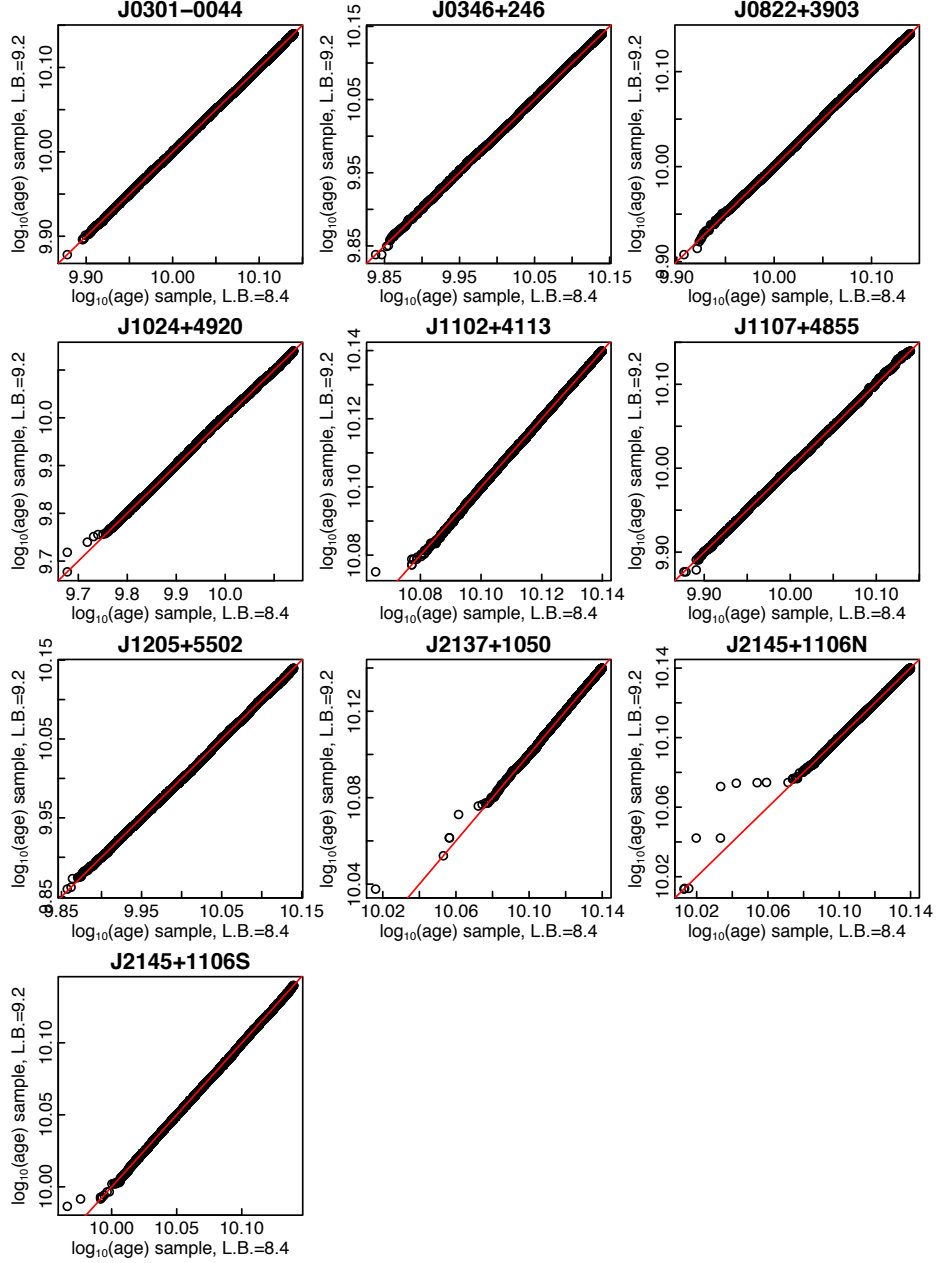


Figure 4.10: Quantile-quantile (QQ) plots of MCMC samples of $\log_{10}(\text{Age})$ of 10 WDs. The x-axis is quantiles from the original fits when lower bound is 8.4 and the y-axis is quantiles based on setting lower bound to be 9.2. The red line represents that quantiles from samples in x and y axes are equal.

Table 4.11: Posterior means and 68.3% credible intervals of the ages (G_{yr}) of ten candidate halo WDs.

WDs	EB-log		FB		Case-by-case	
	Mean	Credible Interval	Mean	Credible Interval	Mean	Credible Interval
J0301-0044	12.18	(11.32, 13.08)	12.01	(10.85, 13.1)	11.04	(8.98, 12.96)
J0346+246	12.33	(11.5, 13.17)	12.25	(11.23, 13.2)	11.77	(10.16, 13.29)
J0822+3903	11.79	(10.81, 12.74)	11.67	(10.32, 12.92)	10.83	(9.78, 12.16)
J1024+4920	11.93	(11, 12.89)	11.53	(9.96, 12.97)	8.58	(7.02, 10.48)
J1102+4113	13	(12.65, 13.41)	13.01	(12.67, 13.41)	13.11	(12.71, 13.56)
J1107+4855	11.54	(10.94, 12.07)	11.3	(9.74, 12.6)	9.71	(8.67, 11.14)
J1205+5502	11.67	(10.81, 12.59)	11.36	(9.67, 12.9)	9.71	(8.65, 10.96)
J2137+1050	13.14	(12.77, 13.51)	13.14	(12.78, 13.52)	13.24	(12.89, 13.61)
J2145+1106N	13	(12.66, 13.35)	13.01	(12.69, 13.35)	13.09	(12.74, 13.47)
J2145+1106S	11.75	(11.04, 12.58)	11.89	(10.99, 12.89)	11.54	(10.83, 12.46)

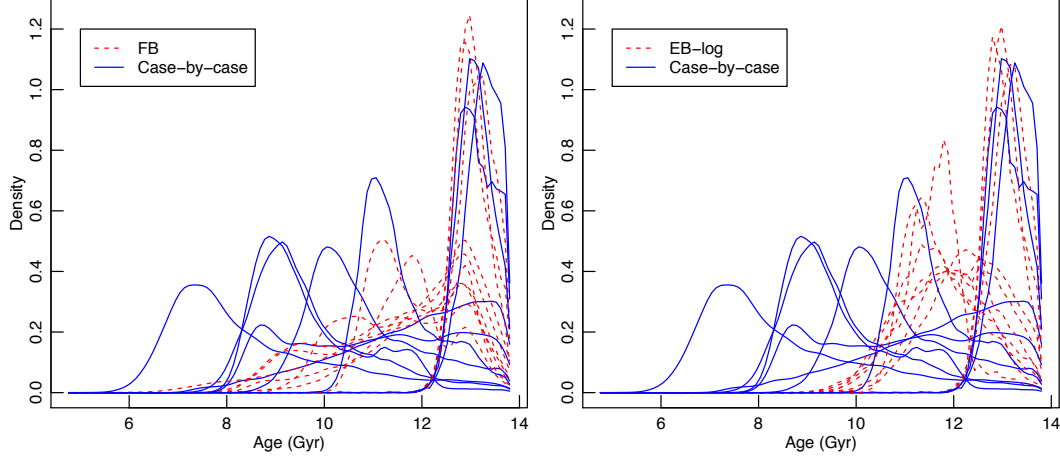


Figure 4.11: The comparison of the posterior distributions of ages of WDs from hierarchical analyses and case-by-case study. The left panel uses FB approach to fit hierarchical model, while the right panel uses EB-log.

Table 4.12: 68.3% intervals for the age distribution, mean age and standard deviation of Galactic halo WDs

Items	Units	FB	EB-log
Prediction interval for the distribution of halo WDs	$\log_{10}(\text{Age})$	10.07 ± 0.05	10.09 ± 0.03
	Age (Gyr)	$11.75^{+1.43}_{-1.28}$	$12.30^{+0.88}_{-0.82}$
Credible interval for the mean age γ of halo WDs	$\log_{10}(\text{Age})$	10.08 ± 0.03	n.a.
	Age (Gyr)	$12.04^{+0.80}_{-0.75}$	n.a.
Credible interval for the standard deviation τ of ages of halo WDs	$\log_{10}(\text{Age})$	0.045 ± 0.029	n.a.
	Age (Gyr)	0.95 ± 0.48	n.a.

case-by-case estimates (solid lines) disperse over a much wider range.

Table 4.12 presents the 68.3% credible interval estimates of population-level parameters. The EB-log produces MAP estimates of population-level parameters $\hat{\gamma} \approx 10.09$ and $\hat{\tau} \approx 0.03$. The FB provides estimates of γ and τ with credible intervals, $\gamma = 10.08 \pm 0.03$ and $\tau = 0.045 \pm 0.029$. The photometric errors of these ten WDs are close to Σ_0 in the simulation study. So the data is similar to Simulation Setting I ($\tau = 0.05, \Sigma = \Sigma_0$). Hence, the advantage of the shrinkage estimates over the case-by-case estimates shown in Simulation Setting I should be predictive and we expect

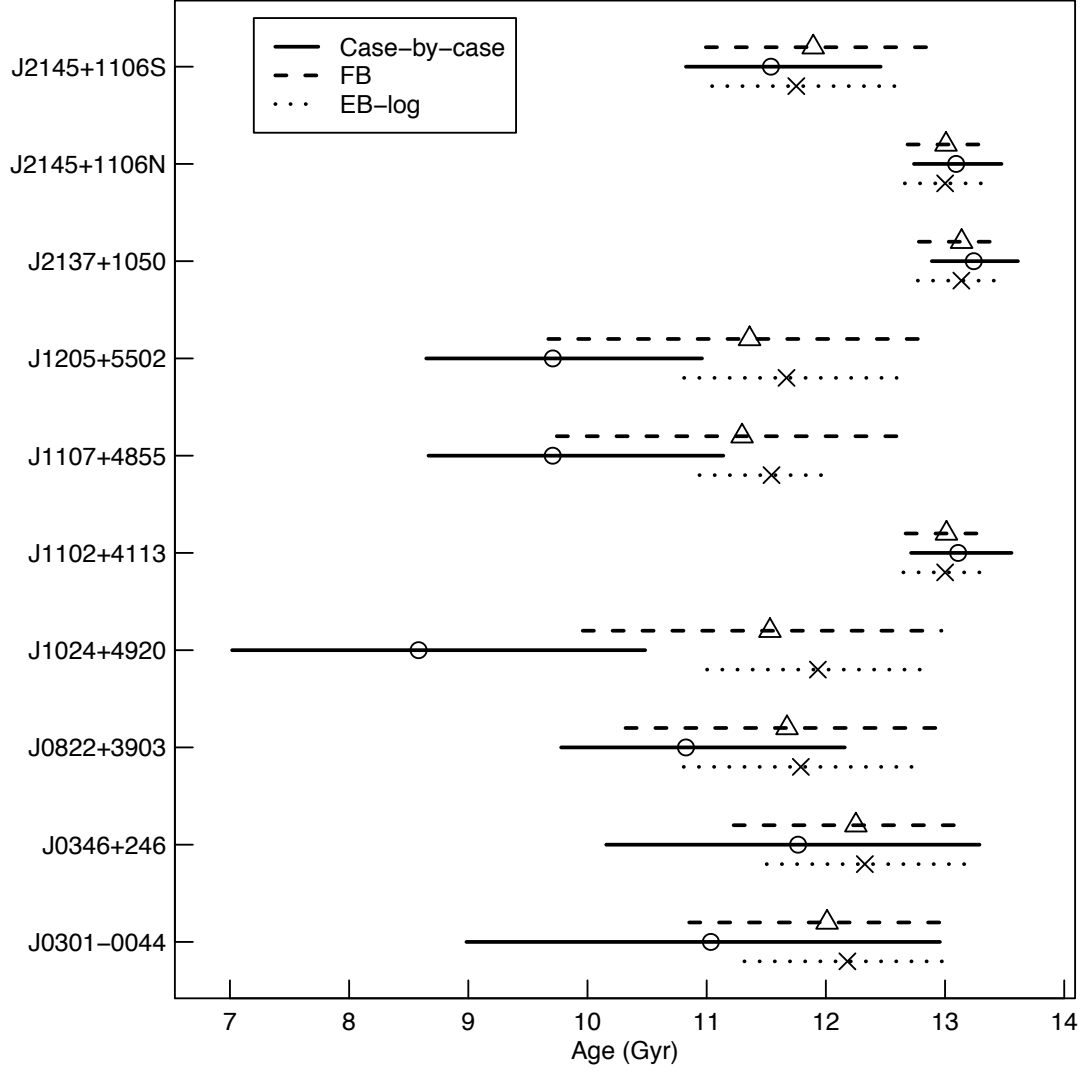


Figure 4.12: The EB-log, FB, and case-by-case estimates of the age of the ten halo WDs. The interval estimates are 68.3% credible intervals of the posterior distributions. The dashed, solid, and dotted intervals are computed from FB, case-by-case, and EB-log, respectively. The triangle, circle and cross signs are the posterior means of the ages of WDs from FB, case-by-case, and EB-log, respectively.

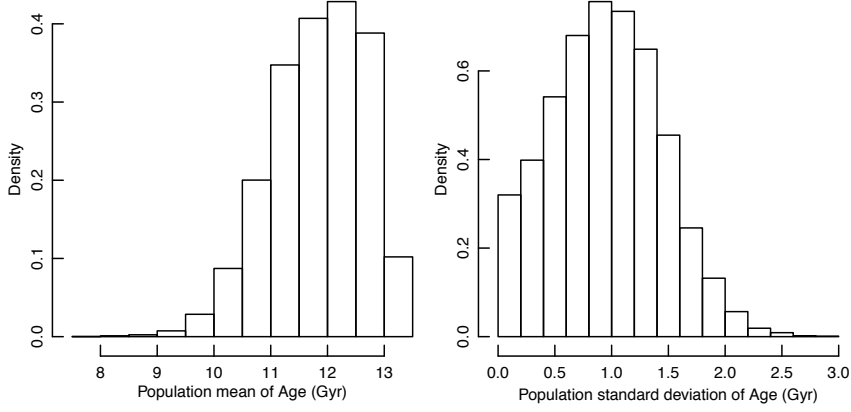


Figure 4.13: Histograms of the Monte Carlo sample for the population mean age and population standard deviation obtained using FB fit to ten candidate galactic halo WDs. Here the population mean of Age (Gyr) = 10^γ and the population standard deviation of Age (Gyr) = $10^{E(\gamma)} \log_e(10)\tau$.

that the dotted lines in Fig. 4.11 are better estimates of the halo WD ages.

Table 4.11 and Fig. 4.12 also summarise the estimated ages. The 68.3% credible intervals of ages of WDs from EB-log and FB are generally shorter than those from the case-by-case analyses, which means that shrinkage estimates (FB and EB-log) produce more precise estimates. The posterior means from the FB and EB-log are quite consistent, but the error bars from EB-log are much shorter than those from FB. This is because the point estimate of τ (0.045) is much larger than that from EB-log (0.03).

In Table 4.12, we report the 68.3% intervals from FB and EB-log for (i) the distribution of halo WDs ages, (ii) the mean age of halo WDs and (iii) the standard deviation of ages of halo WDs. With FB, we obtain a Monte Carlo sample of γ and τ from their posterior distributions. We transform the samples of γ and τ to that of mean and standard deviation of Galactic halo WD ages, with histograms appearing in Fig. 4.13.

Under FB, the 68.3% predictive interval for the distribution of halo WD ages is

10.07 ± 0.05 on $\log_{10}(\text{age})$ scale, i.e., $11.75^{+1.43}_{-1.28}$ Gyr. In other words, we predict that 68.3% of WDs in the Galactic halo have ages between 10.87 Gyr and 13.78 Gyr. The estimates based on EB-log is much narrower, $12.30^{+0.88}_{-0.82}$ Gyr.

Using FB, the 68.3% interval for the mean of $\log_{10}(\text{age})$ of halo WDs (γ) is 10.08 ± 0.03 , i.e., $12.04^{+0.80}_{-0.75}$ Gyr. Based on the simulation study, the actual coverage probability of this interval is 52% with 68.3% confidence interval (40.2%, 63.6%). EB-log does not directly provide an error bar for $\hat{\gamma}$, but the bootstrap techniques (Efron, 1979) could be used. We do not pursue this here, however, because it is computationally expensive and error bars are provided by FB.

Figs 4.14–4.15 display the time series, autocorrelation function and histogram plots of the samples of $\log_{10}(\text{age})$ for 10 WDs from the EB-log approach using the Two-stage approach. Fig. 4.16 illustrate the MCEM updates of γ and τ from the EB-log approach via Two-stage algorithm.

Figs 4.17–4.19 demonstrate the time series and autocorrelation function plot of the samples of $\log_{10}(\text{age})$ for 10 WDs from the FB approach using the Two-stage approach.

Based on Figs 4.14–4.19, we observe that the Two-stage algorithms for both the FB and EB-log approaches converge well and work reasonably well. The time series plots for all parameters are stationary.

4.5 Sensitivity Analysis for the Assumption of the Hierarchical Model

Section 4.3 performs a thorough simulation study to verify that the hierarchical model (in Eq. (4.5)) outperforms the star-by-star analysis (in Eq. (4.3)). Here we investigate the robustness of the hierarchical model to its assumption that halo WD ages follow a

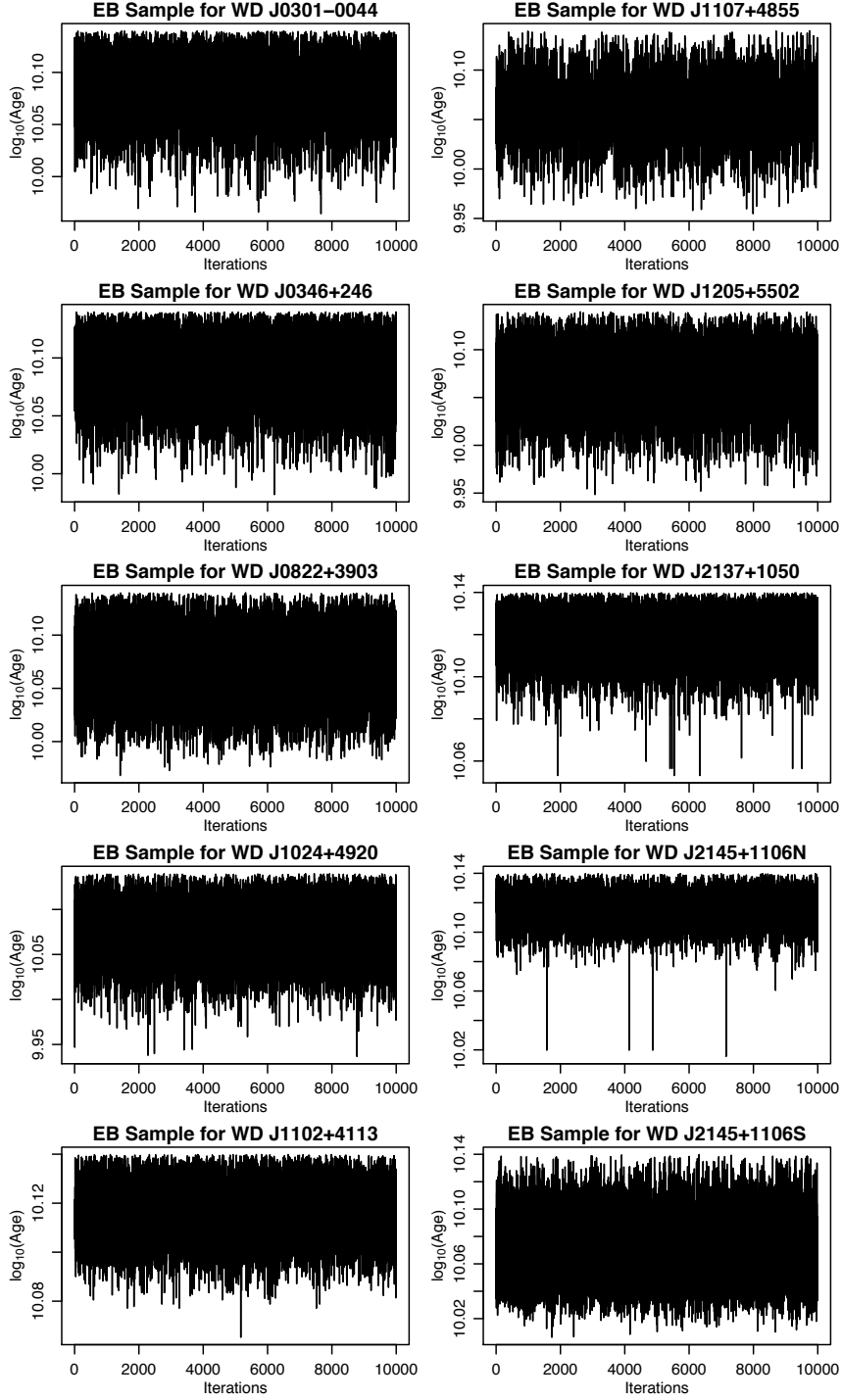


Figure 4.14: Time series plots of samples of $\log_{10}(\text{Age})$ for 10 WDs from the EB-log approach.

Estimating the Ages of a Group of Halo White Dwarfs

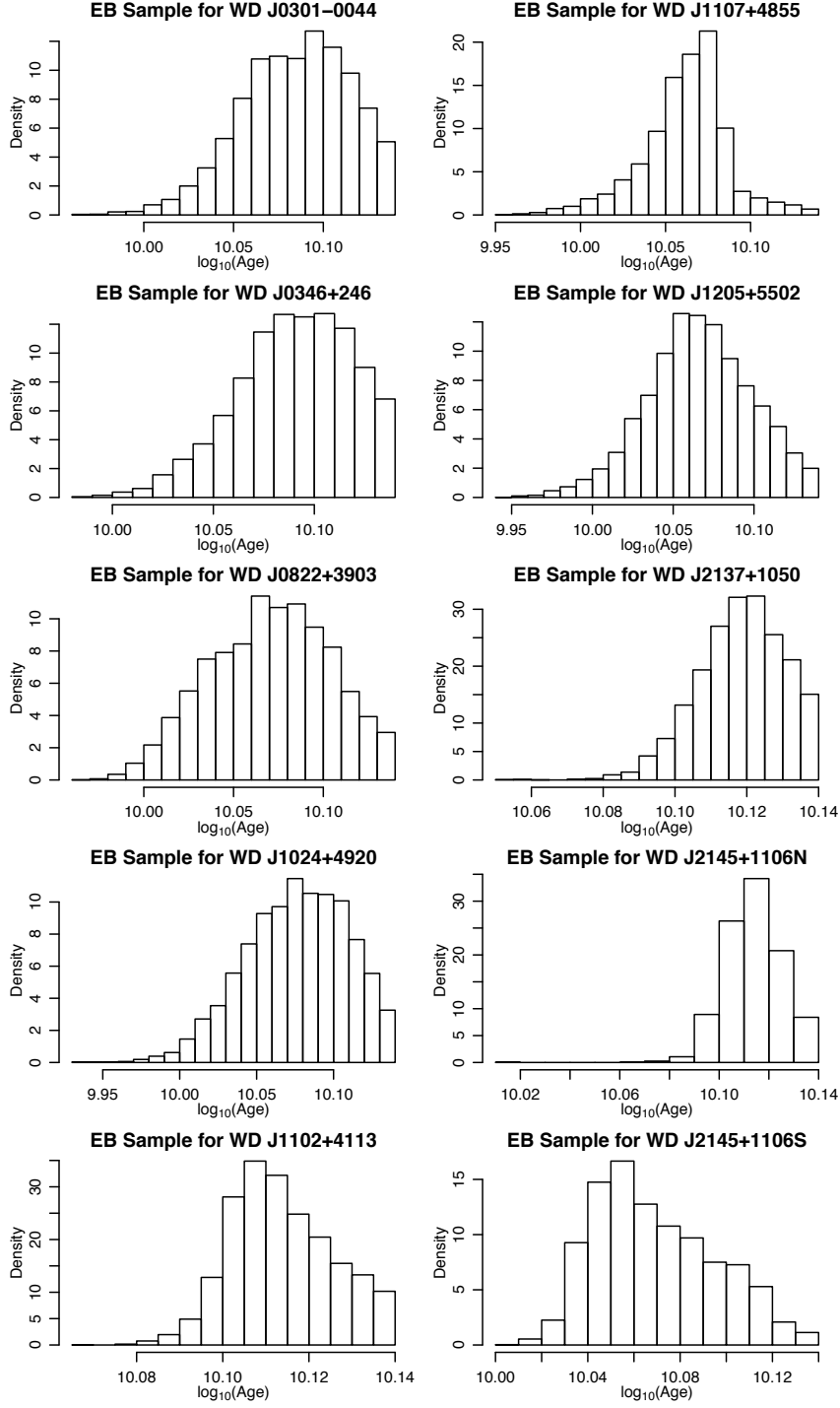


Figure 4.15: Histograms of samples of $\log_{10}(\text{Age})$ for 10 WDs from the EB-log approach.

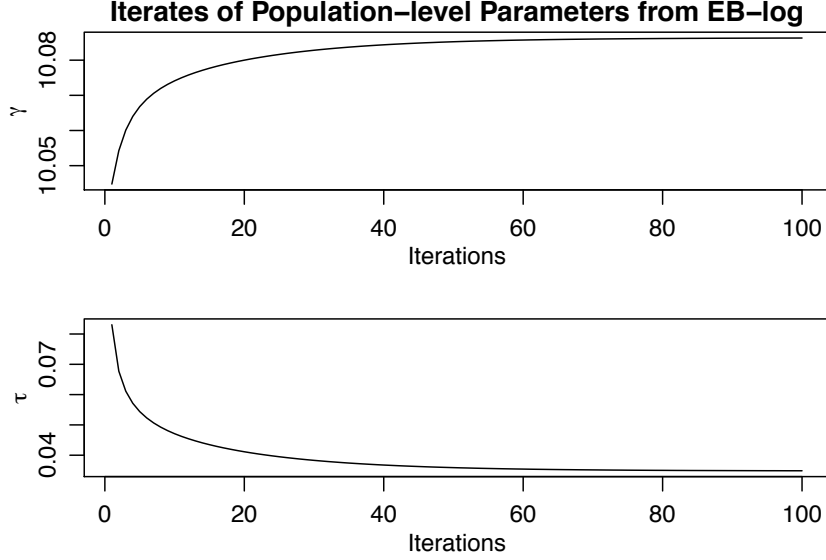


Figure 4.16: Iterates of population-level parameters γ and τ from the EB-log approach.

Gaussian distribution, see Eq. (4.4). Instead of simulating $\log_{10}(\text{Age})$ from Gaussian distributions as in the simulation study in Si et al. (2017b), here we simulate WD ages from the scaled gamma distribution in Fig. 4.20.

The distribution in Fig. 4.20 is excellent to try because it is quite different from the bell-shaped normal distribution. It has a long left tail from 9.9 to 10.06 and peaks at 10.06, then it plunges to 0 at 10.09. This population distribution of $\log_{10}(\text{Age})$ means that more WDs have their $\log_{10}(\text{Age})$ smaller than 10.06 and 10.09 is the upper limit for $\log_{10}(\text{Age})$. It has a mode at 10.06, but it also has a quick cut-off at large ages because 10.09 is the upper limit. It could capture different astrophysics, e.g. infalling satellite galaxies that start their star formation somewhat later. The statistical algorithms developed in this chapter can be readily adapted to using this distribution in the second level of our hierarchical model.

This simulation study consists of the four settings detailed in Table 4.13. In each setting, we generate $N_1 = 25$ groups of $N_2 = 20$ WDs. Let $A_{jk}, D_{jk}, Z_{jk}, M_{jk}$ the

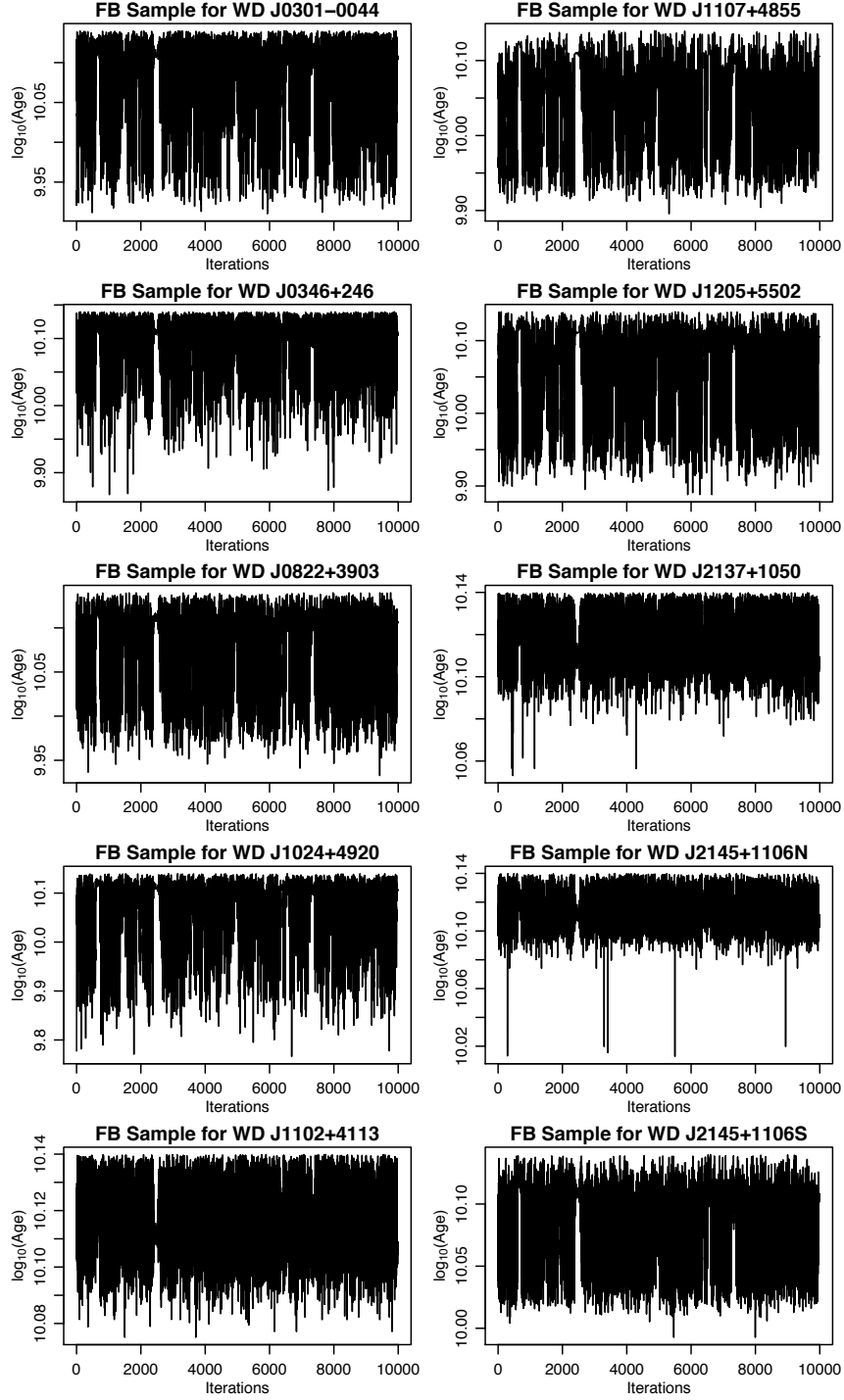


Figure 4.17: Time series plots of samples of $\log_{10}(\text{Age})$ for 10 WDs from the FB approach using the Two-stage algorithm.

Estimating the Ages of a Group of Halo White Dwarfs

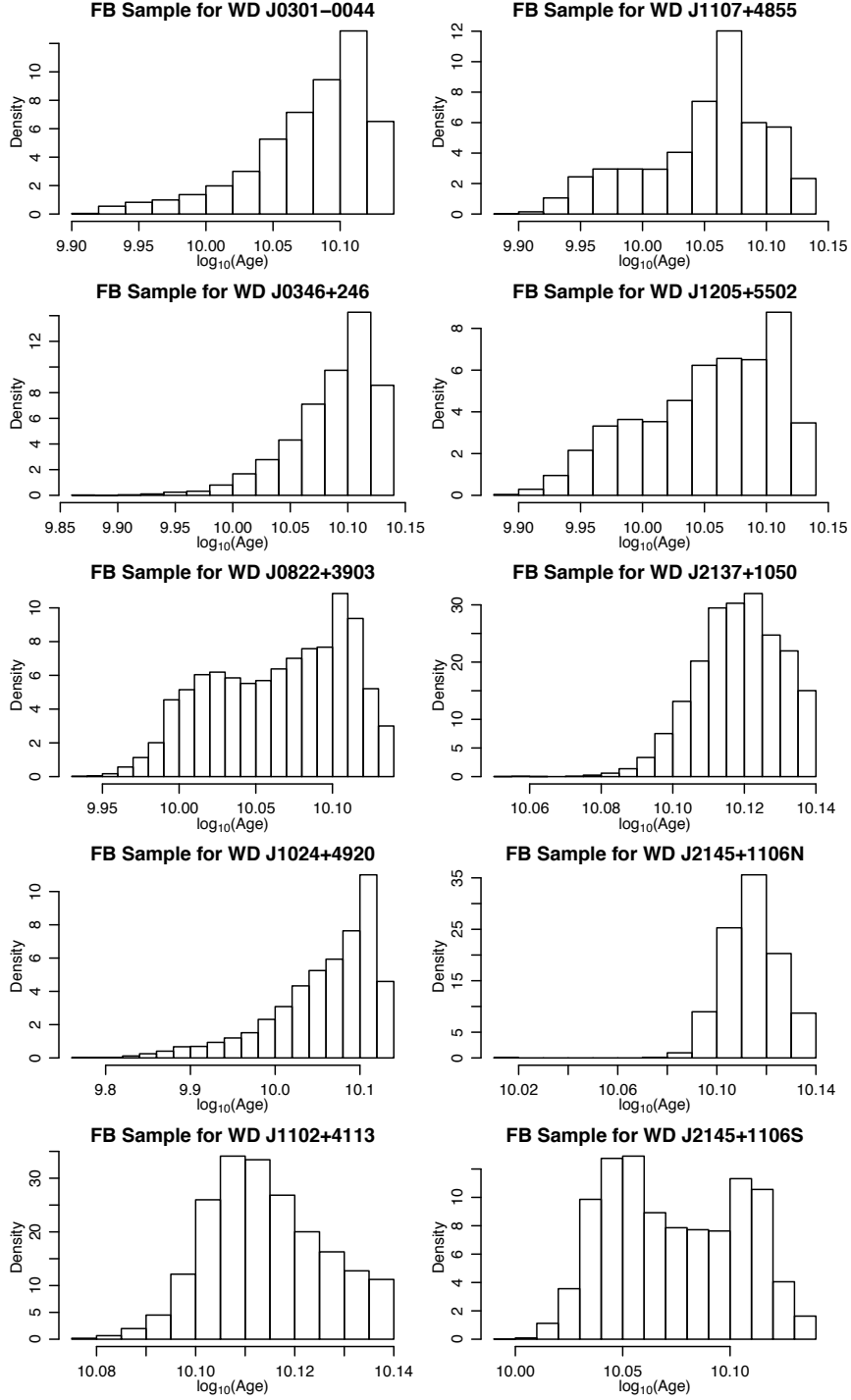


Figure 4.18: Histograms of samples of $\log_{10}(\text{Age})$ for 10 WDs from the FB approach using the Two-stage algorithm.

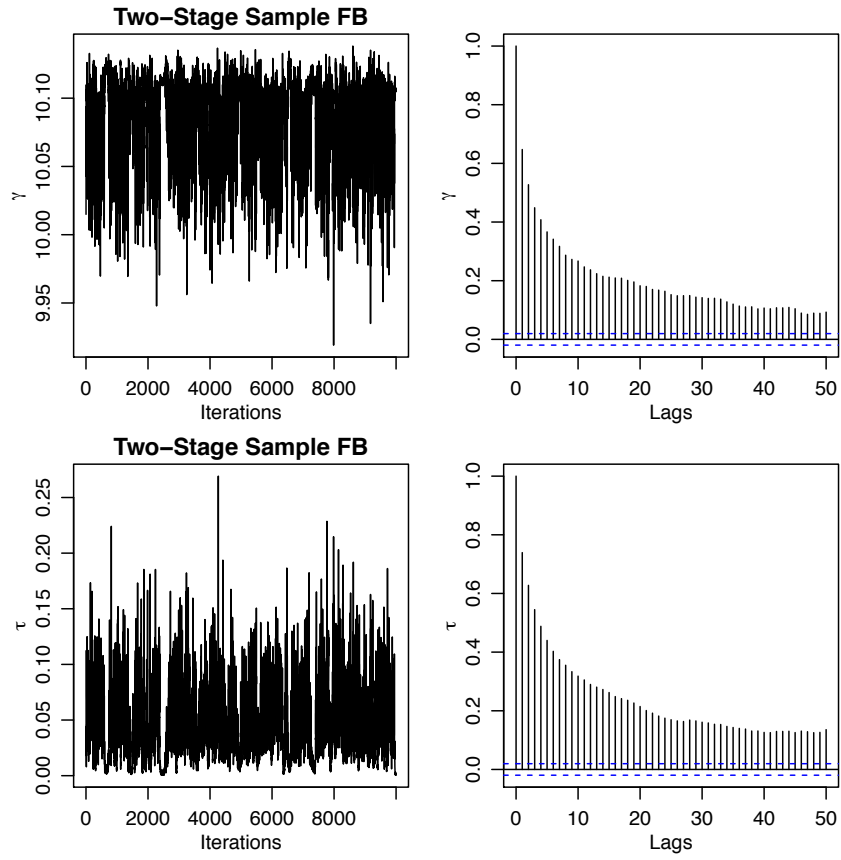


Figure 4.19: Time series and autocorrelation plots of population-level parameters γ and τ from the FB approach using the Two-stage algorithm.

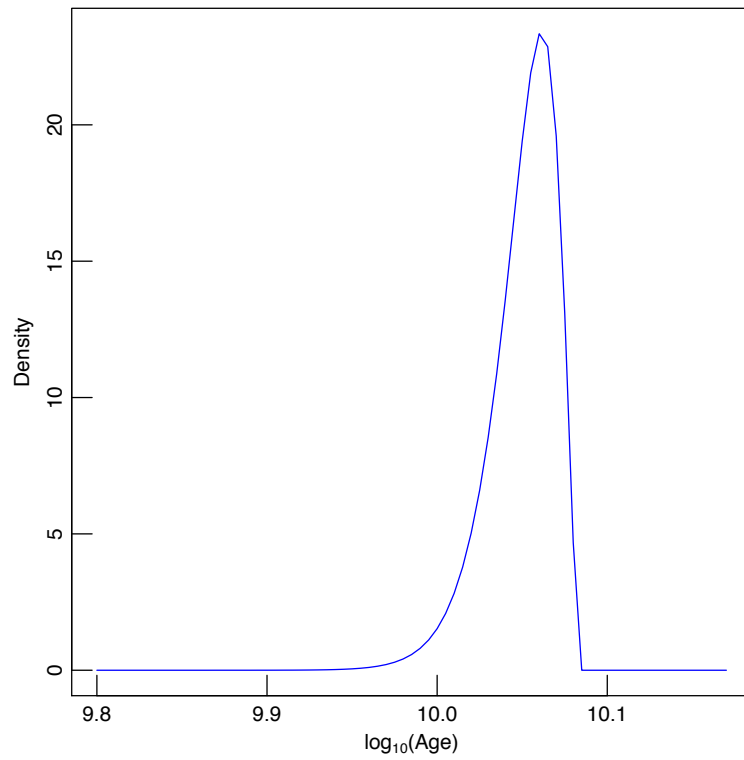


Figure 4.20: The scaled gamma distribution used to generate $\log_{10}(\text{Age})$ in our sensitivity analysis. It has a long left tail from 9.9 to 10.06 and peaks at 10.06, then it plunges to 0 at 10.09. In astronomy, this population distribution of $\log_{10}(\text{Age})$ means that more WDs have their $\log_{10}(\text{Age})$ smaller than 10.06 and 10.09 is the upper limit for $\log_{10}(\text{Age})$.

$\log_{10}(\text{Age})$, distance modulus, metallicity and ZAMS mass of the WD k in group j for $j = 1, \dots, N_1, k = 1, \dots, N_2$. For simplicity, we fix $D_{jk} = 4.0, M_{jk} = 2.0, Z_{jk} = -1.5$ and in each setting we simulate A_{jk} from a scaled gamma distribution with mean 10.05 and standard deviation τ . With $A_{jk}, D_{jk}, Z_{jk}, M_{jk}$ in place, we generate the photometry in filters U, B, V, R, I, J, H, K with measurement error σ by using BASE-9. The values of the standard deviation of the stellar ages, τ , and the photometric errors, σ , are specified in Table 4.13 for each simulated setting.

We employ Montgomery WD cooling model (Montgomery et al., 1999), Dotter MS/RG model (Dotter et al., 2008), and Williams IFMR model (Williams et al., 2009). We use both the star-by-star and fully Bayesian hierarchical approaches to analyze the simulated datasets in each setting. We express the estimates of A_{jk} from hierarchical and star-by-star as $\hat{A}_{jk}^H$ (for “Hierarchical”) and $\hat{A}_{jk}^I$ (for “Individual”), respectively.

Fig 4.21 makes comparison between the hierarchical and star-by-star fits. Under all the four simulation settings, the majority of WD ages are better estimated with the FB hierarchical models. Table 4.13 also demonstrates this point in the “Percentage” (The column name “Perce.” in the table) column under both approaches. These columns gives the percentage of WDs with age better estimated by the hierarchical and star-by-star fits, respectively. In the Setting I, 73.4% of WD ages are more precisely estimated with the hierarchical model, whereas only 26.6% of WD ages are better estimated with the star-by-star analysis. In the Setting II, the benefits of the hierarchical analysis is even more clearer. Even in the Setting IV, the best scenario for the star-by-star analysis, 63.4% of WD ages are better estimated with the hierarchical model.

The advantage of the hierarchical fits is also clear in terms of the mean of absolute

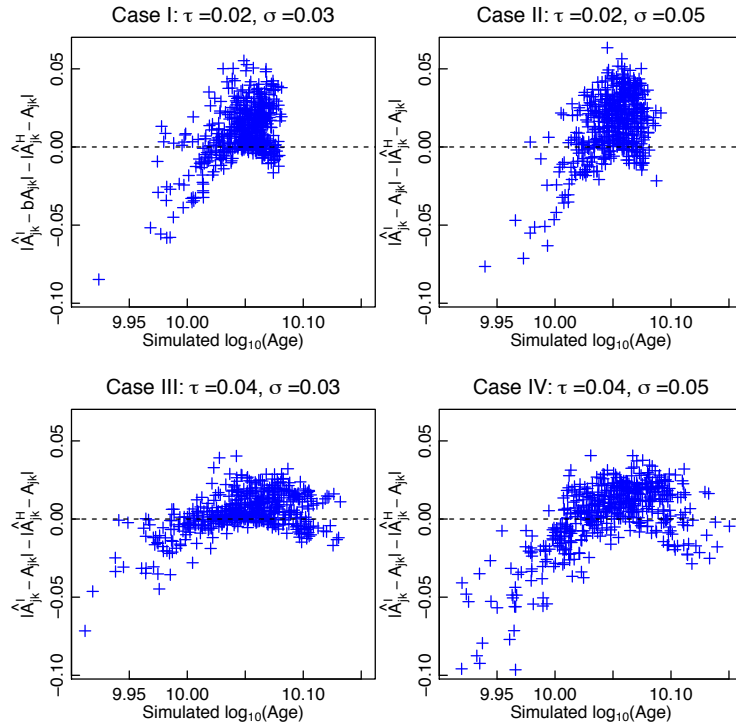


Figure 4.21: The absolute value of errors from the star-by-star fits minus that of the hierarchical fits under four simulation settings. Each plus sign represents one WD and dashed lines indicate 0. Plus signs above the dashed lines correspond to WDs with ages that are better estimated with the hierarchical fits. For plus signs below the dashed lines, star-by-star fits produce better estimates.

Table 4.13: Summary of the relative performance of hierarchical and star-by-star fits in each of four simulated settings.

Settings	Hierarchical			Star-by-star		
	RMSE	MAE	Perce.	RMSE	MAE	Perce.
I: $\tau = 0.02, \sigma = 0.03$	0.019	0.012	73.4%	0.027	0.022	26.6%
II: $\tau = 0.02, \sigma = 0.05$	0.021	0.015	79.2%	0.034	0.029	20.8%
III: $\tau = 0.04, \sigma = 0.03$	0.022	0.016	66.6%	0.027	0.021	33.4%
IV: $\tau = 0.04, \sigma = 0.05$	0.036	0.026	63.4%	0.034	0.028	36.6%

Table 4.14: Coverage probabilities (CP) and their 68.3% confidence intervals (CI) of interval estimates for ages of WDs in each of four simulated settings. The nominal level is 68.3%.

Settings	Hierarchical		Star-by-star	
	CP	68.3% CI	CP	68.3% CI
I: $\tau = 0.02, \sigma = 0.03$	0.602	(0.579, 0.625)	0.66	(0.637, 0.682)
II: $\tau = 0.02, \sigma = 0.05$	0.518	(0.495, 0.541)	0.63	(0.607, 0.652)
III: $\tau = 0.04, \sigma = 0.03$	0.662	(0.64, 0.684)	0.636	(0.613, 0.658)
IV: $\tau = 0.04, \sigma = 0.05$	0.584	(0.561, 0.607)	0.68	(0.658, 0.701)

errors (MAE) in Table 4.13, which is defined as

$$MAE = \frac{1}{N_1 \cdot N_2} \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} |\hat{A}_{jk} - A_{jk}|,$$

where A_{jk} is the true $\log_{10}(\text{Age})$ of the star k in the group j and $\hat{A}_{jk}$ is its estimate. The MAE from the hierarchical fits is smaller than that of the star-by-star under all settings.

Table 4.14 presents the actual coverage probabilities of 68.3% credible intervals estimates for ages of simulated WDs derived from two methods: hierarchical (FB) and case-by-case. The 68.3% confidence intervals of the coverage probabilities are computed with the Clopper-Pearson exact method. From this table, the case-by-case performs better than FB, and the coverage probabilities of its credible intervals are between 63% and 68%, slightly under the nominal value. The actual coverage

probabilities of credible intervals from FB vary from 51.8% to 66.2%, significantly below the nominal level 68.3%. This situation is because the case-by-case method fits WDs individually, so it is not affected by the population distribution of $\log_{10}(\text{age})$. The poor performance of FB is caused by model misspecification.

We examined the advantages of hierarchical fits over star-by-star fits when the distribution of $\log_{10}(\text{age})$ is non-gaussian. We show that even when the population distribution of $\log_{10}(\text{age})$ is asymmetric, the hierarchical model still performs better in terms of the precision of estimates. However, the magnitude of improvement is not as good as in Section 4.3. This is caused by the fact that the population distribution is skewed and non-normal, but we used a Gaussian population in the hierarchical analysis.

From the simulation results, we conclude that the hierarchical estimates have smaller MSEs than the case-by-case fits, which means that pooling WDs in hierarchical model can be beneficial even when the population distribution of WDs' $\log_{10}(\text{age})$ deviates from the normal assumption. This research focuses on statistical side, mainly statistical algorithms to compute hierarchical models for astronomy problems. So we have not investigated how the equation of state of white dwarfs affect the estimates of ages (Winget and Kepler, 2008; Potekhin and Chabrier, 2013; Salaris et al., 2013). This can be a good project for my future research. If WDs are from accretion events, their population distribution of $\log_{10}(\text{Age})$ is less likely to be normal. In the future, we will attempt to put more flexible distributions for the $\log\text{Age}$, maybe mixture normal distributions. Also our statistical techniques can be easily adapted to fitting the hierarchical models with more complicated population distributions and more hyper-parameters. But for now, because of the limited datasets, we are unable to fit complicated population distributions well.

4.6 Concluding Remarks

WDs are the last evolutionary stage of most stars with initial masses high enough to have evolved off of the main sequence and less than $8.0 M_{\odot}$. Precise determination of the ages of WDs is vital in studying how they evolve von Hippel et al. (2006); DeGennaro et al. (2009); van Dyk et al. (2009); O'Malley et al. (2013); Si et al. (2017b). Bayesian hierarchical models are statistical tools that combine multiple objects and produce partial pooling (shrinkage) estimates that have better statistical properties such as less MSE and more robustness Gelman (2006a); Morris and Lysy (2012); Gelman et al. (2014). This chapter applies the Bayesian hierarchical modelling technique to a group of WDs in the hope that we can obtain more accurate estimates of their ages. Due to the complicated nature of stellar evolution, the likelihood function for each WD is intricate and computer-based, so Bayesian hierarchical modelling of a group of WDs is computationally challenging. This fact motivates this chapter.

In this chapter, we have developed a method for fitting hierarchical models when we have output (particularly, MCMC samples) from black-box codes (in our case, the BASE-9 package) that do not require opening up the box. We have developed our methods in the context of a prototype analysis for Gaia, which will measure exquisite distances to tens of thousands of WDs. Our new hierarchical approach is expected to produce better estimates of WD ages in terms of MSE than the case-by-case analyses, and to provide better estimates for the age distributions of the underlying Galactic populations. This is merely a prototype analysis for what we will do with Gaia data.

We note that there are a few methods or black-box codes (now or in the future) (e.g., Hernandez et al., 1999; Pont and Eyer, 2004; da Silva et al., 2006; Hernandez and Valls-Gabaud, 2008; Bailer-Jones, 2011) that can fit WDs or stellar clusters one-at-a-time. Our statistical methods in this chapter were developed for use with BASE-9,

but they are not reliant on BASE-9. Other researchers could choose their preferred black-box codes and use our techniques to obtain the hierarchical fits.

Secondly, as to the Gaussian assumption on the distribution of $\log_{10}(\text{age})$ for a group of halo WDs in Eq. (4.4), this is just a preliminary start. Readers are encouraged to use their own distributions and our statistical techniques will still work. The computing time will remain similar, even when the population distribution becomes complicated. In the future, we might experiment with flexible distributions on $\log_{10}(\text{age})$, such as a Gaussian mixture distribution to model the ages of WDs after the merger of galaxies. For now with 10 candidate WDs for the Galactic halo, a complicated population distribution with many parameters may cause insufficient constraints of hyper-parameters. When the Gaia data are available and the number of WDs are enough, we can use these methods to fit population distributions with more parameters.

Thirdly, because we focus on the development of statistical techniques, we have not investigated the uncertainties of the underlying stellar evolution models yet, for instance, the effects of different Main-sequence/red giant models on the estimates of WDs' ages. We will investigate these kind of problems in future research.

5 Initial-Final Mass Relations

The IFMR provides a mapping between the ZAMS mass of a star to its WD mass and is vital for the determination of parameters in stellar evolution. Many researchers have investigated the IFMR using data from different star clusters, leading to numerous versions of the IFMR. For instance, Williams et al. (2004); Williams et al. (2009) probe the IFMR using WDs from many clusters such as, the Hyades, NGC 2168 (M35) and so on. Salaris et al. (2009) provides an empirical estimate of the IFMR using published results about WDs in ten clusters. Stein et al. (2013) treats the parametrised IFMR as cluster-specific parameters and develops simultaneous principled Bayesian estimates of all cluster-specific parameters, including those describing the IFMR. Stein et al. (2013) implements their methods via a black-box code, BASE-9. Given the photometric data of a star cluster, BASE-9 can produce MCMC samples for cluster parameters such as, age, distance, and IFMR parameters.

Bayesian hierarchical models combine multiple objects (i.e., units of data) and produce partially pooled (shrinkage) estimates that have better statistical properties such as smaller mean squared errors (MSE) and better predictive accuracy (Gelman, 2006a; Morris and Lysy, 2012; Gelman et al., 2014; Katahira, 2016). To derive precise estimates of IFMRs, we propose a Bayesian hierarchical model that pools photometric data from multiple star clusters. However, this black-box contains complicated stellar evolution models that are unavailable in closed forms, which causes computa-

tional challenges when combining multiple star clusters in the hierarchical model. We developed efficient algorithms for fitting hierarchical models in this situation, taking advantage of the black box that we do not open and do not need to alter.

We perform a simulation study to show the benefits of the Bayesian hierarchical model. We apply this model to five star clusters: the Hyades, M67, NGC 188, NGC 2168, and NGC 2477, leading to reasonable and consistent estimates of IFMRs for these clusters. We illustrate how cluster-specific (case-by-case) analysis of NGC 188 using its own photometric data can produce an unreasonable IFMR since its WDs have a narrow range of Zero-age Main Sequence masses (ZAMS masses). However, the Bayesian hierarchical model corrects the cluster-specific analysis by borrowing strength from other clusters, thus generating more reliable estimates of IFMR parameters.

This analysis is just a testbed of our statistical techniques, and the astronomical results from this chapter are not definitive. The target of this chapter is to present how our Bayesian hierarchical methods can be beneficial to the study of IFMRs in astronomy.

5.1 Case-by-case Analysis

Stein et al. (2013) developed a Bayesian approach to incorporate parametrised IFMRs with other astrophysical models such as Main-sequence/Red Giant (MS/RG) models, WD cooling models, etc. They inferred all parameters by using a state-of-the-art Markov chain Monte Carlo (MCMC) algorithm and implemented their methods with the software package BASE-9 (von Hippel et al., 2006; DeGennaro et al., 2009; van Dyk et al., 2009). BASE-9, short for Bayesian Analysis of Stellar Evolution with 9 parameters, deploys MCMC techniques to perform reliable Bayesian analysis for stellar parameters including age, distance modules, metallicity and mass, based on

the photometry of stars in a star cluster. Stein et al. (2013) treats IFMR parameters in a cluster-specific way, so various star clusters may have different IFMR parameters. In summary, Stein et al. (2013) provides a Bayesian analysis for one single stellar cluster, i.e., cluster-specific analysis. In this chapter we employ the mathematical notations of Stein et al. (2013), while the subscript is extending to accommodate multiple star clusters.

Suppose we have photometry for K star clusters, along with measurement errors. The number of stars in each cluster can vary, as can the number of photometric magnitudes among the clusters or even among the stars within the clusters. We use k to index clusters and i to index stars within clusters. To simplify notations, we assume the number of stars within each cluster is N and the observed photometry vector for star i within cluster k is $\mathbf{X}_{ki}$, with known measurement variance-covariance matrix Σ_{ki} . We assume that age (θ_{age}), metallicity ($\theta_{[\text{Fe}/\text{H}]}$), distance modulus ($\theta_{\text{m-M}_V}$), and absorption (θ_{A_V}) are common to all stars in each cluster, and we denote $\Theta_k = (\theta_{\text{age},k}, \theta_{[\text{Fe}/\text{H}],k}, \theta_{\text{m-M}_V,k}, \theta_{\text{A}_V,k})$. We denote the parameters describing the IFMR of cluster k by α_k ; below we use a linear IFMR model so each α_k consists of a intercept and a slope. Since α_k is the same for all WDs in cluster k , we treat α_k as a cluster parameter. See Table 5.1 for a concise summary of the model parameters. We denote the ZAMS mass of star i within cluster k by M_{ki} . Also, a star may be a field star, i.e., not a member of a specific cluster. We define $\mathbf{Z}_k = (Z_{k1}, \dots, Z_{kN})$, where $Z_{ki} = 1$ if star i observed within cluster k is indeed a member, otherwise $Z_{ki} = 0$. Though we have distinct evolution models for MS/RG and WD stars, we denote them indistinguishably by $\mathbf{G}(\cdot)$, which comprises MS/RG evolution models, WD cooling models, WD atmosphere models, and IFMR models.

Table 5.1: Cluster and stellar parameters.

Cluster parameters	
$\theta_{\text{age},k}$	$\log_{10}$ age of the cluster k
$\theta_{[\text{Fe}/\text{H}],k}$	metallicity of the cluster k
$\theta_{\text{m-M}_V,k}$	distance modulus of the cluster k
$\theta_{\text{Av},k}$	$\log_{10}$ absorption of the cluster k
$\boldsymbol{\alpha}_k$	IFMR parameters of the cluster k
Stellar parameters	
M_{ki}	the initial mass of the observed star i of the cluster k
Z_{ki}	indicator for the membership of the observed star i in the cluster k

We parametrise the IFMR of cluster k as a linear form

$$M_{\text{WD},ki} = \alpha_{k0} + \alpha_{k1}(M_{ki} - 3.0) \text{ for WD } i \text{ in cluster } k, \quad (5.1)$$

where $\boldsymbol{\alpha}_k = (\alpha_{k0}, \alpha_{k1})$ are the intercept¹ and slope parameters, $M_{\text{WD},ki}$ is the mass of WD i within cluster k . Specifically α_{k0} is the WD mass of a star in cluster k with progenitor ZAMS mass equal to $3.0 M_{\odot}$. For every additional increment of $1.0 M_{\odot}$ in ZAMS mass, we expect the WD mass to increase by α_{k1} . Because the expected photometric magnitudes of WDs depend on the WD masses, $\boldsymbol{G}(\cdot)$ must incorporate $\boldsymbol{\alpha}_k$. Thus, for the remainder of this Chapter, the stellar evolution model $\boldsymbol{G}(\cdot)$, is viewed as a function of $\boldsymbol{\alpha}_k$ in addition to $\boldsymbol{\Theta}_k$ and M_{ki} .

The cluster-specific model for the cluster k is

$$\boldsymbol{X}_{ki} \mid (\boldsymbol{\Theta}_k, M_{ki}, \boldsymbol{\alpha}_k) \sim N(\boldsymbol{G}(\boldsymbol{\Theta}_k, M_{ki}, \boldsymbol{\alpha}_k), \boldsymbol{\Sigma}_{ki}), \text{ if } Z_{ki} = 1, \quad (5.2)$$

where $\boldsymbol{G}(\boldsymbol{\Theta}_k, M_{ki}, \boldsymbol{\alpha}_k)$ is the predicted vector of photometric magnitudes for star

¹ α_{k0} is the IFMR intercept of the line when $M_{ki} = 3.0$, i.e., the predicted final mass of a progenitor star with a ZAMS of 3.0.

i within cluster k . Eq. (5.2) summarise the probabilistic relationship between the photometry of stars ($Z_{ki} = 1$) that are members of cluster k and the model parameters. If a star is not a member of cluster k ($Z_{ki} = 0$), i.e., it is a field star, we assume that its magnitudes are uniformly distributed on a hyper-rectangle which includes the range of observed magnitudes of stars for that cluster. We use $p_{\text{field}}(\cdot)$ to denote the distribution of photometric magnitudes for field stars, which is simply the reciprocal of the volume of the hyper-rectangle. To be specific, for example, star cluster k has photometric magnitudes in the U, B, V filters. Then we find the range of U, B, V by using the maximum value of each filter minus its minimum and denote them ℓ_U, ℓ_B and ℓ_V . If a star with magnitudes $\mathbf{X}_{ki}$ is a field star, its likelihood is $p_{\text{field}}(\mathbf{X}_{ki}) = \frac{1}{\ell_U \ell_B \ell_V}$. The uniform model for field stars is unrealistic, but one study (Stenning et al., 2016) has investigated whether inclusion of a more realistic and complicated field-star model (Robin et al., 2003) is necessary and found that the simple model is sufficient. Stenning et al. (2016) performed simulation study using the uniform distribution for field stars, and found that the black-box (BASE-9) misidentifies field stars as cluster stars $< 2\%$ of the time, and never misidentifies cluster stars as field stars. For details, refer to Page 9 of Stenning et al. (2016). Therefore, for simplicity we use uniform model for field stars in each cluster.

Taken together, this means that the likelihood function for cluster k is

$$L(\mathbf{M}_k, \boldsymbol{\Theta}_k, \mathbf{Z}_k, \boldsymbol{\alpha}_k \mid \mathbf{X}_k, \boldsymbol{\Sigma}_k) = \prod_{i=1}^N \left[\frac{Z_{ki}}{\sqrt{|2\pi\boldsymbol{\Sigma}_{ki}|}} \times \exp \left(-\frac{1}{2} \left(\mathbf{X}_{ki} - \mathbf{G}(M_{ki}, \boldsymbol{\Theta}_k, \boldsymbol{\alpha}_k) \right)^\top \boldsymbol{\Sigma}_{ki}^{-1} \left(\mathbf{X}_{ki} - \mathbf{G}(M_{ki}, \boldsymbol{\Theta}_k, \boldsymbol{\alpha}_k) \right) \right) + (1 - Z_{ki}) p_{\text{field}}(\mathbf{X}_{ki}) \right], \quad (5.3)$$

where $\mathbf{M}_k = (M_{k1}, \dots, M_{kN})$, $\mathbf{X}_k = (\mathbf{X}_{k1}, \dots, \mathbf{X}_{kN})$, and $\boldsymbol{\Sigma}_k = (\boldsymbol{\Sigma}_{k1}, \dots, \boldsymbol{\Sigma}_{kN})$. The

prior distribution for the parameters is assumed to be

$$\begin{aligned}
 p(\boldsymbol{\Theta}_k, \boldsymbol{\alpha}_k, \mathbf{M}_k, \mathbf{Z}_k) &= p(\boldsymbol{\Theta}_k) p(\boldsymbol{\alpha}_k) p(\mathbf{M}_k, \mathbf{Z}_k) \\
 &= p(\theta_{\text{age},k}) p(\theta_{[\text{Fe}/\text{H}],k}) p(\theta_{m-M_V,k}) p(\theta_{A_V,k}) p(\boldsymbol{\alpha}_k) \times \\
 &\quad \prod_{i=1}^N p(M_{ki} \mid Z_{ki}) p(Z_{ki}).
 \end{aligned} \tag{5.4}$$

Specifically, for $\theta_{\text{age},k}$, $\theta_{[\text{Fe}/\text{H}],k}$, and $\theta_{m-M_V,k}$ we use independent Gaussian prior distributions, with means set in accordance to the most recently published result and variances chosen to be reasonably non-informative. We use a uniform prior distribution for the IFMR intercept α_{k0} , on the real line, while the IFMR slope α_{k1} is uniform on the positive part of the real line, to exclude the possibility of a monotonically decreasing IFMR. The prior probability of star i being a member of cluster k , $p(Z_{ki} = 1)$ is set based on best external information, typically using proper motions and/or radial velocities. Finally, $p(M_{ki} \mid Z_{ki} = 1)$ is the prior distribution of the ZAMS for star i , for which we use the initial mass function (IMF) of Miller and Scalo (1979) in Table 4.2.

In their cluster-specific analysis of cluster k , Stein et al. (2013) based statistical inference, including parameters' estimates and error bars, on the joint posterior distribution,

$$\begin{aligned}
 p(\boldsymbol{\Theta}_k, \mathbf{M}_k, \mathbf{Z}_k, \boldsymbol{\alpha}_k \mid \mathbf{X}_k, \boldsymbol{\Sigma}_k) \\
 \propto p(\boldsymbol{\Theta}_k, \boldsymbol{\alpha}_k, \mathbf{M}_k, \mathbf{Z}_k) L(\mathbf{M}_k, \boldsymbol{\Theta}_k, \mathbf{Z}_k, \boldsymbol{\alpha}_k \mid \mathbf{X}_k, \boldsymbol{\Sigma}_k).
 \end{aligned} \tag{5.5}$$

The BASE-9 package can draw a reliable sample for all parameters from their joint posterior distribution in Eq. (5.5) using the partially collapsed sampler (van Dyk and Park, 2008). This sampler proceeds in three steps:

- 1.) It integrates out $\mathbf{M}_k = (M_{k1}, \dots, M_{kN})$ and $\mathbf{Z}_k = (Z_{k1}, \dots, Z_{kN})$, i.e.,

$$\begin{aligned} p(\boldsymbol{\Theta}_k, \boldsymbol{\alpha}_k | \mathbf{X}_k, \boldsymbol{\Sigma}_k) &= \int \cdots \int \left(\Sigma_{Z_{k1}} \cdots \Sigma_{Z_{kN}} p(\boldsymbol{\Theta}_k, \mathbf{M}_k, \mathbf{Z}_k, \boldsymbol{\alpha}_k | \mathbf{X}_k, \boldsymbol{\Sigma}_k) \right) d\mathbf{M}_k d\mathbf{Z}_k \\ &= p(\boldsymbol{\Theta}_k) p(\boldsymbol{\alpha}_k) \prod_{i=1}^N \left[c(\boldsymbol{\Theta}_k, \boldsymbol{\alpha}_k) P(Z_{ki} = 1) + p_{\text{field}}(\mathbf{X}_{ki}) P(Z_{ki} = 0) \right], \end{aligned} \quad (5.6)$$

where

$$c(\boldsymbol{\Theta}_k, \boldsymbol{\alpha}_k) = \int p(\mathbf{X}_{ki} | \boldsymbol{\Theta}_k, M_{ki}, \boldsymbol{\alpha}_k, Z_{ki} = 1) p(M_{ki} | Z_{ki} = 1) dM_{ki}.$$

We include the prior probabilities $P(Z_{ki} = 1)$ of each star's membership in Eq. 5.6, which we obtain from recent external research.

- 2.) We use an adaptive Metropolis-Hastings algorithm to draw samples of $\boldsymbol{\Theta}_k$ and $\boldsymbol{\alpha}_k$ from the marginal posterior distribution in Eq. (5.6).
- 3.) Given samples of $\boldsymbol{\Theta}_k$ and $\boldsymbol{\alpha}_k$, we draw samples of $\mathbf{M}_k$ and $\mathbf{Z}_k$. For more details of the sampler, refer to Stein et al. (2013).

This is a cluster-specific study of the IFMRs, since the fits of the IFMR parameters only rely on one cluster. We aim to perform a comprehensive analysis of the IFMR by combining multiple star clusters into a hierarchical model. This allows us to simultaneously obtain superior estimates of each cluster's IFMR and to estimate the intrinsic variance of cluster-specific IFMRs.

5.2 Hierarchical Models for a Group of Star Clusters

In this section, we describe how to pool data from multiple star clusters using a hierarchical model and how we fit this comprehensive model. For K star clusters, our

hierarchical model is

$$\begin{cases} \mathbf{X}_{ki} \mid (\boldsymbol{\Theta}_k, M_{ki}, \boldsymbol{\alpha}_k) \sim \mathcal{N}\left(\mathbf{G}(\boldsymbol{\Theta}_k, M_{ki}, \boldsymbol{\alpha}_k), \boldsymbol{\Sigma}_{ki}\right), & \text{if } Z_{ki} = 1 \\ \text{where } \boldsymbol{\alpha}_k \sim \mathcal{N}(\boldsymbol{\gamma}, \boldsymbol{\Gamma}), k = 1, \dots, K, \end{cases} \quad (5.7)$$

with

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

For field stars ($Z_{ki} = 0$), $\mathbf{X}_{ki}$ is uniformly distributed on the aforementioned hyper-rectangle. We set prior distributions on $\boldsymbol{\Theta}_k, \mathbf{M}_k, \mathbf{Z}_k, k = 1, \dots, K$ as in the aforementioned cluster-specific analysis. The only new parameters in the hierarchical model in Eq. (5.7) are $\boldsymbol{\gamma}$ which is the mean of the IFMR intercept and slope and the $\boldsymbol{\Gamma}$ which is the variance of IFMR parameters of different clusters. We must set prior distribution for $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ and set $p(\boldsymbol{\gamma}, \boldsymbol{\Gamma}) = p(\boldsymbol{\gamma}|\boldsymbol{\Gamma})p(\boldsymbol{\Gamma})$ with $p(\boldsymbol{\gamma} \mid \boldsymbol{\Gamma})$ uniform on its range. For $\boldsymbol{\Gamma}$, we use the marginally non-informative prior distribution proposed by Huang and Wand (2013), and its hierarchical form is shown as follows,

$$\boldsymbol{\Gamma} \mid \lambda_1, \lambda_2 \sim \text{Inverse Wishart}\left(2\nu \begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{pmatrix}, \nu + 1\right),$$

where Inverse Wishart ² is prior distribution for the variance parameter $\mathbf{\Gamma}$, $\lambda_1, \lambda_2 \sim \text{Inverse Gamma}(1/2, 1/\Lambda)$ ³. Huang and Wand (2013) suggests that $\nu = 2$ leads to marginal uniform distribution for correlation ρ and arbitrarily large positive Λ leads to arbitrarily weakly informative prior distributions for σ_1 and σ_2 . To have a marginally non-informative prior distribution on ρ , we set $\nu = 2$. As for Λ , we choose a large value of 1000 and we perform a sensitivity analysis to show in 5.6.1 that the hierarchical result is insensitive to the value of Λ when $\Lambda \geq 1000$.

For clarity, we categorise parameters in the hierarchical model (Eq. (5.7)) into 3 types, as shown in Table 5.2. Later we discuss the Bayesian inference of the hierarchical model.

The joint posterior density for all parameters in Eq. (5.7) is

$$\begin{aligned}
 & p(\boldsymbol{\gamma}, \mathbf{\Gamma}, \lambda_1, \lambda_2, \boldsymbol{\Theta}_1, \mathbf{M}_1, \mathbf{Z}_1, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\Theta}_K, \mathbf{M}_K, \mathbf{Z}_K, \boldsymbol{\alpha}_K \\
 & \quad | \mathbf{X}_1, \boldsymbol{\Sigma}_1, \dots, \mathbf{X}_K, \boldsymbol{\Sigma}_K) \\
 & = p(\lambda_1)p(\lambda_2)p(\boldsymbol{\gamma}, \mathbf{\Gamma} | \lambda_1, \lambda_2) \times \prod_{k=1}^K \\
 & \quad \left(p(\boldsymbol{\Theta}_k)p(\mathbf{M}_k, \mathbf{Z}_k)p(\boldsymbol{\alpha}_k | \boldsymbol{\gamma}, \mathbf{\Gamma})L(\mathbf{M}_k, \boldsymbol{\Theta}_k, \mathbf{Z}_k, \boldsymbol{\alpha}_k | \mathbf{X}_k, \boldsymbol{\Sigma}_k) \right)
 \end{aligned} \tag{5.8}$$

² Inverse Wishart distribution is the conjugate prior for the covariance matrix of a multivariate normal distribution, having two parameters: scale matrix $\boldsymbol{\Psi}$ and degrees of freedom ν , and its probability density function is

$$p(\mathbf{X}|\boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{p\nu}{2}}\Gamma_p(\frac{\nu}{2})}|\mathbf{X}|^{-\frac{\nu+p+1}{2}}e^{-\frac{1}{2}tr(\boldsymbol{\Psi}\mathbf{X}^{-1})},$$

in which Γ_p is the multivariate gamma function, and $tr(\cdot)$ is the trace function.

³Inverse Gamma is the reciprocal of of Gamma distribution, with two parameters: shape α and rate β , and its density function is

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right),$$

where Γ is the Gamma function.

Table 5.2: Categories of parameters in the hierarchical model

Cluster(Object)-level parameters	
$\Theta_k = (\theta_{\text{age},k}, \theta_{[\text{Fe}/\text{H}],k}, \theta_{\text{m-M}_V,k}, \theta_{\text{A}_V,k})$	Cluster parameters of the cluster k
$\mathbf{M}_k$	the initial masses of all stars in cluster k
$\mathbf{Z}_k$	membership parameters for stars in cluster k
α_k	IFMR parameters of the cluster k , this is the only part that has hierarchical modelling
Population-level parameters	
γ	population mean of IFMR parameters
Γ	population variance of IFMR parameters
Hyper-parameters in priors	
ν	the degree of freedom in the inverse Wishart prior distribution
λ_1, λ_2	parameters in the scale matrix in inverse Wishart prior distribution

5.3 Two-stage Algorithm for Hierarchical Modelling

The joint posterior distribution in Eq. (5.8) is high-dimensional, with $2N \times K + 6K + 7$ parameters. In this section, we show how to take advantage of the cluster-specific fittings in BASE-9 via a two-stage (TS) algorithm to fit the hierarchical model.

To simplify the description of TS algorithm, we introduce $\Omega_k = (\mathbf{M}_k, \mathbf{R}_k, \mathbf{Z}_k, \Theta_k)$ for $k = 1, 2, \dots, K$. We use $\Psi = \text{diag}(1/\lambda_1, 1/\lambda_2)$.

Step 0a: For each star cluster run BASE-9 to obtain a Monte Carlo sample of $p(\mathbf{M}_k, \mathbf{R}_k, \alpha_k, \mathbf{Z}_k, \Theta_k | \mathbf{X}_k, \Sigma_k)$ via the cluster-specific analysis. Thin each chain to obtain an essentially independent Monte Carlo sample and label the samples as $\{\Omega_1^{(t)}, \alpha_1^{(t)}, \dots, \Omega_K^{(t)}, \alpha_K^{(t)}, t = 1, 2, \dots, t_{MC}\}$.

Step 0b: We denote the TS samples by the tilde notation. Simulate K random integers between 1 and t_{MC} , denote them $r_1^*, \dots, r_K^*$, and initialise the parameters $\tilde{\alpha}_k^{(1)} = \alpha_k^{(r_k^*)}$ and $\tilde{\Omega}_k^{(1)} = \Omega_k^{(r_k^*)}$ for $k = 1, \dots, K$.

Step 0c: Given $\tilde{\Omega}_1^{(1)}, \tilde{\alpha}_1^{(1)}, \dots, \tilde{\Omega}_K^{(1)}, \tilde{\alpha}_K^{(1)}$, simulate $\tilde{\gamma}^{(1)}$ and $\tilde{\Gamma}^{(1)}$ via the partially collapse Gibbs (PCG) sampler (van Dyk and Park, 2008),

$$\begin{aligned}\tilde{\Gamma}^{(1)} &\sim \text{Inverse Wishart}\left(K + \nu, \sum_{k=1}^K (\tilde{\alpha}_k^{(1)} - \tilde{\alpha}^{(1)})(\tilde{\alpha}_k^{(1)} - \tilde{\alpha}^{(1)})^\top\right), \\ \tilde{\gamma}^{(1)} &\sim N_2\left(\tilde{\alpha}^{(1)}, \tilde{\Gamma}^{(1)}/K\right),\end{aligned}$$

$$\text{with } \tilde{\alpha}^{(1)} = \frac{1}{K} \sum_{k=1}^K \tilde{\alpha}_k^{(1)}.$$

Step 0d: Given $\tilde{\Gamma}^{(1)}$, we simulate λ_1, λ_2 from their conditional posterior distributions,

$$\tilde{\lambda}_\ell^{(1)} \sim \text{Inverse Gamma}\left(1 + \nu/2, \frac{1}{\Lambda_\ell} + \nu(\tilde{\Gamma}^{(1)})_{\ell\ell}^{-1}\right)$$

for $\ell = 1, 2$.

For $s = 1, \dots$, run Step 1 and Step 2 iteratively.

Step 1: Randomly generate K integers between 1 and t_{MC} , and denote them $r_1, \dots, r_K$.

For each $k = 1, \dots, K$, set $\alpha_k^* = \alpha_k^{(r_k)}$ and $\Omega_k^* = \Omega_k^{(r_k)}$ as the new proposal and set $\tilde{\alpha}_k^{(s+1)} = \alpha_k^*, \tilde{\Omega}^{(s+1)} = \Omega_k^*$ with probability $\min\left\{1, \frac{p(\alpha_i^* | \tilde{\gamma}^{(s)}, \tilde{\Gamma}^{(s)})}{p(\alpha_i^{(s)} | \tilde{\gamma}^{(s)}, \tilde{\Gamma}^{(s)})}\right\}$. Otherwise, set $\tilde{\alpha}_k^{(s+1)} = \tilde{\alpha}_k^{(s)}, \tilde{\Omega}^{(s+1)} = \tilde{\Omega}^{(s)}$.

Step 2: Given $\tilde{\lambda}_\ell^{(s)}, \ell = 1, 2$ and $\tilde{\alpha}_k^{(s+1)}$, and $\tilde{\Omega}_k^{(s+1)}, k = 1, \dots, K$, update $\tilde{\gamma}, \tilde{\Gamma}$ via

$$\begin{aligned}\tilde{\Gamma}^{(s+1)} &\sim \text{Inverse Wishart}\left(K + \nu, \sum_{k=1}^K \left(\tilde{\alpha}_k^{(s+1)} - \tilde{\alpha}^{(s+1)}\right)\left(\tilde{\alpha}_k^{(s+1)} - \tilde{\alpha}^{(s+1)}\right)^\top + 2\nu \text{diag}\left(1/\tilde{\lambda}_1^{(s)}, 1/\tilde{\lambda}_2^{(s)}\right)\right), \\ \tilde{\gamma}^{(s+1)} &\sim N_2\left(\tilde{\alpha}^{(s+1)}, \tilde{\Gamma}^{(s+1)}/K\right),\end{aligned}$$

$$\text{with } \tilde{\alpha}^{(s+1)} = \frac{1}{K} \sum_{k=1}^K \tilde{\alpha}_k^{(s+1)}.$$

Table 5.3: Parameter distributions for Simulated Clusters

$\theta_{\text{age},k} \sim N(9.0, 0.1^2),$
$\theta_{[\text{Fe}/\text{H}],k} \sim N(-0.08, 0.05^2),$
$\theta_{\text{m-M}_V,k} = 10.0,$
$\theta_{\text{A}_V,k} \sim N(0.51, 0.43^2)$ truncated to positive values,
$\boldsymbol{\alpha}_k \sim N_2(\boldsymbol{\gamma}_0, \boldsymbol{\Gamma}_0),$ with
$\boldsymbol{\gamma}_0 = \begin{pmatrix} 0.722 \\ 0.107 \end{pmatrix}, \boldsymbol{\Gamma}_0 = \begin{pmatrix} 0.05^2 & -0.2 \cdot 0.05 \cdot 0.02 \\ -0.2 \cdot 0.05 \cdot 0.02 & 0.02^2 \end{pmatrix}$

Given $\tilde{\boldsymbol{\Gamma}}^{(s+1)}$, simulate $\tilde{\lambda}_\ell, \ell = 1, 2$ via

$$\tilde{\lambda}_\ell^{(s+1)} \sim \text{Inverse Gamma} \left(1 + \nu/2, \frac{1}{\Lambda_\ell} + \nu(\tilde{\boldsymbol{\Gamma}}^{(s+1)})_{\ell\ell}^{-1} \right).$$

5.4 Simulation Study

To illustrate the statistical advantages of our hierarchical model, we simulate $K = 10$ star clusters with BASE-9 and we recover their IFMRs via both cluster-specific and hierarchical analyses. We simulate the cluster parameters using the distributions in Table 5.3. These parameter values in Table 5.3 are set to be similar to those of the observed star clusters that we analyse in Section 5.5. To mimic the errors of the observed photometry, we compute the average standard deviations for filters B, V , and I of the WDs in the datasets analysed in Section 5.5 and use them as the corresponding errors in the simulated datasets. Specifically, the observed errors for B, V, I are $\sigma_B = 0.026, \sigma_V = 0.035$, and $\sigma_I = 0.185$.

After simulating the parameters $\theta_{\text{age},k}, \theta_{[\text{Fe}/\text{H}],k}, \theta_{\text{m-M}_V,k}$, and $\theta_{\text{A}_V,k}$ of cluster k , we simulate photometric data for each star in the cluster using BASE-9 (see the BASE-9 User’s Manual, von Hippel et al., 2014). For each cluster, we simulate 200 MS/RG stars brighter than $V = 15$ and 10 WDs. Subsequently, we recover parameters by fitting the simulated datasets with BASE-9 in the cluster-specific analysis in Eq.

Table 5.4: RMSE and coverage probabilities (CP) of 68.3% credible intervals (CI) for IFMR parameters from FB and case-by-case analyses.

Items	FB			Case-by-case		
	RMSE	CP	68.3% CI	RMSE	CP	68.3% CI
IFMR Const.	0.067	0.764	(0.734, 0.792)	0.202	0.8	(0.771, 0.826)
IFMR Slope	0.019	0.628	(0.595, 0.66)	0.331	0.844	(0.818, 0.867)

(5.2) and obtain samples for parameters from the posterior distribution in Eq. (5.5). For the entire simulation study, we employ the Dotter et al. (2008) MS/RG models, Montgomery et al. (1999) WD interior models, Bergeron et al. (1995, , with updates posted online) WD atmospheres, and the IFMR model in Eq. (5.1).

We repeat the data generation and parameter estimation 25 times, record results from both the FB and case-by-case methods. We compare estimates in terms of two criteria: i.) root mean squared error (RMSE) of point estimates, and 2.) actual versus nominal coverage probabilities of interval estimates. When we require point estimates, we utilise posterior means from MCMC samples, and for interval estimates we use the 68.3% credible intervals from posterior distributions by finding the 15.8% and 84.1% quantiles from the MCMC samples.

Table 5.4 presents the RMSE of point estimates and the actual coverage probabilities of 68.3% credible interval estimates for IFMR constants and slopes using two methods: FB and case-by-case. The 68.3% confidence intervals of the coverage probabilities are computed with the Clopper-Pearson exact method (Clopper and Pearson, 1934). From this table, the RMSE of FB estimates of IFMR constants is 0.067, about a third of that from the case-by-case method (0.202). The performance of FB is even better on IFMR slope with its RMSE, 0.019, about 1/17 of that from the case-by-case analysis. In terms of interval estimates of IFMR parameters, the case-by-case (cluster-specific) method has actual coverage probabilities, 80% and 84.4% for IFMR constant and slope respectively, higher than the nominal value, 68.3%. The actual

Table 5.5: RMSE and coverage probabilities (CP) of 68.3% credible intervals (CI) for population-level parameters from FB.

Items	FB		
	RMSE	CP	68.3% CI
γ_1	0.021	0.92	(0.824, 0.972)
γ_2	0.009	0.68	(0.561, 0.782)
$\tilde{\gamma}_1 = \gamma_1 - 3\gamma_2$	0.034	0.76	(0.644, 0.851)
σ_1	0.057	0.6	(0.48, 0.711)
σ_2	0.007	0.88	(0.777, 0.945)
ρ	0.281	0.96	(0.874, 0.993)

coverage probabilities of interval estimates from FB, 76.4% and 62.8% for IFMR constant and slope respectively, are closer to the nominal value. In summary, the estimates of IFMR parameters from FB fits outperform that from case-by-case fits in terms of RMSE and coverage property.

Here are population-level parameters: the population mean of IFMR parameters, $\gamma = (\gamma_1, \gamma_2)$, in which γ_1 and γ_2 are the mean of IFMR constants and slopes, respectively; the population variance-covariance matrix of IFMR parameters,

$$\mathbf{\Gamma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

in which σ_1, σ_2 and ρ are the standard deviations of IFMR constants, slopes and the correlation between them, respectively. In our analysis, IFMR is parameterised as $M_{\text{final}} = \alpha_0 + \alpha_1(M_{\text{initial}} - 3.0)$, but many researchers use another form $M_{\text{final}} = \tilde{\alpha}_0 + \alpha_1 M_{\text{initial}}$, $\tilde{\alpha}_0 = \alpha_0 - 3\alpha_1$, which is the intercept of the IFMR line with the $M_{\text{initial}} = 0.0$. To make our results comparable with others, we transform our results in the $\tilde{\alpha}_0$ (intercept) and α_1 form. We denote the population mean of intercept at 0 to be $\tilde{\gamma}_1$.

Table 5.5 presents the RMSE of point estimates and the actual coverage proba-

bilities of 68.3% credible interval estimates for population-level parameters using the FB approach. The 68.3% confidence intervals of the coverage probabilities are computed with the Clopper-Pearson exact method (Clopper and Pearson, 1934). The case-by-case method does not pool star clusters into a population, so it fails to produce estimates of the population. From this table, the RMSE of FB estimates of population-level parameters are small except σ_1 and ρ . From the actual coverage property of interval estimates, the FB method tends to produce over-covered interval estimates. For parameters like γ_2 , $\tilde{\gamma}_1$ and σ_1 , their interval estimates perform well, and their actual coverage probabilities are close to the nominal level 68.3%, and their 68.3% confidence interval of coverage contains the nominal value. The difficulty of FB method in estimating population-level parameters is mainly due to two reasons, the small number of objects, $I = 10$. Other research have also reported and discussed this problem, refer to Browne and Goldstein (2002); Browne et al. (2006) for more details.

To further investigate the advantage of the hierarchical analysis in Eq. (5.7), we take one group of 10 simulated clusters as an example and compare the estimates of IFMR parameters from the case-by-case and FB fits. After we obtain the MCMC sample for the IFMR parameters, we use its sample means as point estimates of $\boldsymbol{\alpha}_k, k = 1, \dots, K$. We denote the sample means of IFMR parameters from the cluster-specific fits as $\hat{\boldsymbol{\alpha}}_k^{\text{CS}}$, and denote those from the hierarchical analysis as $\hat{\boldsymbol{\alpha}}_k^{\text{Hier}}$.

Table 5.6 reports the true values of 10 simulated IFMR parameters and their estimates from both the hierarchical and cluster-specific fits. For most of the clusters, the hierarchical estimates are closer to their true values than are their cluster-specific counterparts. Specifically, the hierarchical estimates are better for clusters 2, 3, 4, 6, 7, 9, and 10. Overall, the RMSE of the hierarchical estimates is 0.012, i.e., the average deviation of hierarchical estimates from the true values is 0.012, while that

Table 5.6: Point Estimates of IFMR parameters for 10 simulated clusters from both the cluster-specific and hierarchical fits

Cluster (k)	Hierarchical Estimates		True Values		Cluster-specific Estimates	
	IFMR Const. ($\hat{\alpha}_{k0}^{\text{Hier}}$)	IFMR Slope ($\hat{\alpha}_{k1}^{\text{Hier}}$)	IFMR Const. (α_{k0})	IFMR Slope (α_{k1})	IFMR Const. ($\hat{\alpha}_{k0}^{\text{CS}}$)	IFMR Slope ($\hat{\alpha}_{k1}^{\text{CS}}$)
1	0.789	0.09	0.794	0.081	0.801	0.078
2	0.777	0.144	0.761	0.145	0.78	0.176
3	0.688	0.133	0.7	0.115	0.638	0.205
4	0.696	0.11	0.698	0.089	0.69	0.119
5	0.681	0.11	0.656	0.121	0.663	0.125
6	0.735	0.114	0.728	0.112	0.736	0.118
7	0.788	0.104	0.798	0.097	0.829	0.068
8	0.677	0.102	0.669	0.12	0.673	0.108
9	0.68	0.083	0.685	0.08	0.69	0.074
10	0.623	0.072	0.61	0.089	0.626	0.067

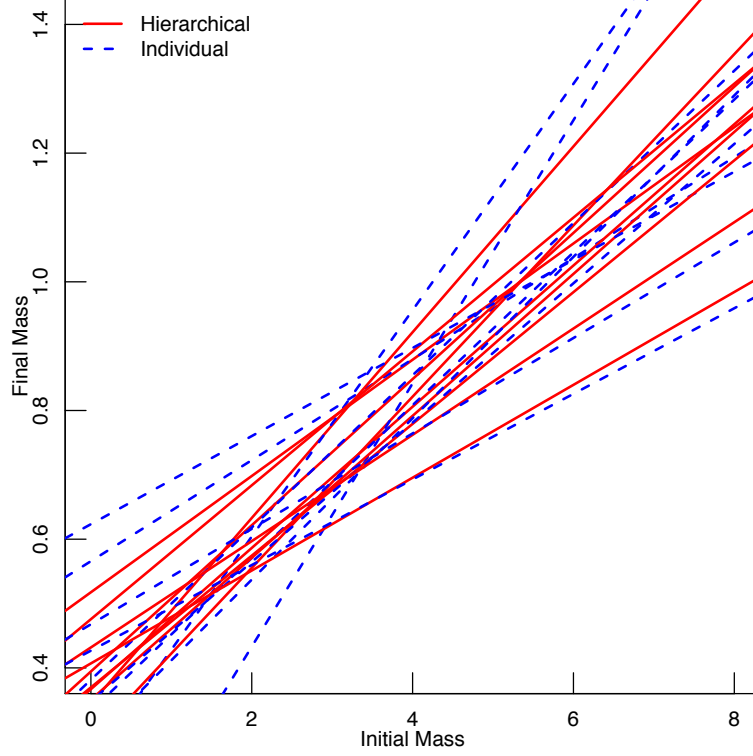


Figure 5.1: Estimated IFMR parameters for 10 simulated clusters using both the hierarchical (red solid lines) and cluster-specific (blue dashed lines) analyses.

of the cluster-specific estimates is 0.029, more than twice the hierarchical value.

Fig. 5.1 illustrates the recovered IFMRs for the 10 simulated clusters. The red solid lines are the hierarchical fits and the blue dashed lines are the cluster-specific fits. A key feature of the hierarchical estimates is that they tend to cluster toward the centre, displaying the shrinkage effect (Morris and Lysy, 2012; Gelman et al., 2014) of hierarchical models. Statistically, this property stems from the assumption that the IFMR parameters of different clusters are generated from the same bivariate normal distribution in Eq. (5.7). Astrophysically, this corresponds to the expectation that the IFMRs of different clusters, although not identical, are similar. Our Bayesian hierarchical model is similar in spirit to the method in Si et al. (2017b), in which we pool ten Galactic halo white dwarfs in a Bayesian hierarchical analysis in which

Table 5.7: CPU Times of two methods in the IFMR project

Algorithms	Time
Case-by-case Analysis	About 30 hours for 10000 draws
Hierarchical Model	
FB Two-stage	About 30 hours and 10 minutes for 10000 draws

we assume that their ages follow a common normal distribution. Si et al. (2017a) verifies that even when this normality assumption is deviated, estimates based on the Bayesian hierarchical model still outperform their case-by-case counterparts.

Table 5.7 presents the computing time of the case-by-case and FB fits. Because each star cluster may have different number of stars, which affects the computing time, here is the time for a cluster consisting of 200 stars, each of which has three photometric magnitudes U , V , and I . The case-by-case analysis takes about 30 hours for 10000 draws and the FB with two-stage sampler uses the case-by-case fits and costs additional 10 minutes to produce estimates from the hierarchical model. The Gibbs sampler is too slow for this complicated hierarchical model, so I have not used it in this project.

In summary, the FB approach with two-stage sampler produces shrinkage estimates of the IFMR parameters that have smaller RMSEs than the case-by-case analysis and better coverage properties. Also, if the case-by-case samples are available, it only takes the two-stage algorithm ten minutes to obtain the estimates under the hierarchical model. So we recommend readers to use our two-stage algorithm to fit hierarchical models.

5.5 Data Analysis

In this section we deploy both cluster-specific and hierarchical analyses using photometry for five star clusters: the Hyades, M67, NGC 188, NGC 2168, and NGC

2477. In the data analysis, we use Montgomery et al. (1999) WD interior models, and Bergeron et al. (1995) WD atmospheres models. For the MS/RG models, we use Dotter et al. (2008) models for all clusters except NGC2168, which is too young for the Dotter et al. (2008) models, so we choose Girardi et al. (2000) models for NGC 2168 in this section.

The Bayesian hierarchical analysis of IFMRs for five star clusters is just a testbed for our statistical techniques. The astronomy results from this analysis are preliminary but not meant to be definitive.

5.5.1 Case-by-case Analysis

We perform the cluster-specific analysis developed by Stein et al. (2013), which uses BASE-9 to deliver MCMC samples of all model parameters for each cluster. The prior distributions that we use for distance moduli, metallicities, and absorptions of five clusters are shown in Table 5.8. Because the MS/RG models tend to poorly predict the photometry of faint main sequence stars, we removed main sequence stars with V magnitudes greater than the cluster-specific thresholds given in Table 5.8. Table 5.8 also gives the references where we obtain the cluster-specific prior distributions and cut-off for the V magnitude.

Here we explain the prior distribution of distance modulus for the Hyades in Table 5.8. In the case-by-case (cluster-specific) analysis via BASE-9, we assume that all stars in a specific cluster have the same distance modulus. This assumption is approximately true for clusters fairly far from the Earth. For the Hyades, due to the fact that its closeness (~ 50 pc) to the Solar System is comparable to its depth ($\sim 10\text{--}20$ pc), so its member stars have significantly different distances, which violates the equal distance assumption in the BASE-9. To address this problem, DeGennaro (2009) adjusted the magnitudes of each star for its distance using the precise distance

Table 5.8: Prior distributions, maximum V magnitudes and their references for five clusters

Cluster	Dist. Mod.	Metallicity	Absorption ^a	$V_{\max}$	Reference
Hyades	$N(0.0, 0.03^2)$	$N(0.07, 0.05^2)$	$N(0.009, 0.006^2)$	4.5	DeGennaro (2009); Stein et al. (2013)
M67	$N(9.62, 0.091^2)$	$N(-0.009, 0.009^2)$	$N(0.127, 0.013^2)$	15.0	VandenBerg and Stetson (2004); Taylor (2006)
NGC 188	$N(11.24, 0.1^2)$	$N(-0.03, 0.1^2)$	$N(0.27, 0.1^2)$	15.5	Meibom et al. (2009)
NGC 2168	$N(10.3, 0.1^2)$	$N(-0.2, 0.3^2)$	$N(0.682, 0.1^2)$	30.0	Stein et al. (2013)
NGC 2477	$N(11.46, 0.0^2)$	$N(-0.1, 0.0^2)$	$N(0.75, 0.0^2)$	15.5	Jeffery et al. (2011); Stein et al. (2013)

^a All prior distribution are truncated to positive values

estimates obtained with the Hipparcos Satellite (Perryman et al., 1997). Each Hyades star was offset to a nominal distance modulus, $m - M_V = 0.0$, i.e., 10 pc. That is why we set the prior distance modulus to be a Gaussian distribution with mean 0. Additionally, because the Hyades is well-studied and the uncertainty of its distance modulus is small, we take 0.03 as the prior standard deviation. After we obtain the MCMC samples for the Hyades from the case-by-case analysis, we add the average distance modulus from multiple studies, 3.40, (Perryman et al., 1998; DeGennaro, 2009) to the MCMC sample of distance modulus and this is how we procure the posterior sample of distance modulus for Hyades with BASE-9. For details, refer to (DeGennaro, 2009; DeGennaro et al., 2009; Stein et al., 2013).

Because NGC 2477 includes a large number (533) of stars and we are primarily interested in the IFMR parameters, we set the prior standard deviations for the distance modulus, absorption, and metallicity to zero. This allows BASE-9 to fix these parameters at their prior means and only obtains MCMC sample for the other parameters, which saves significant computational time.

The MCMC samples from the cluster-specific analyses appear in Fig. 5.2. Each row corresponds to one cluster, and the columns provide scatter plots of various parameter combinations.

Because the prior standard deviations of metallicity, distance modulus, and absorption are set to zero for NGC 2477, the scatter plots of age–metallicity, age–distance, age–absorption degenerate into lines. The scatter plot of IFMR parameters for NGC 2477 has two separate modes. The upper one, accounting for 90.44% of the whole distribution, tends to have larger IFMR slope than the lower one, constituting 9.56% of the posterior distribution. The most likely explanation for the bimodal nature is uncertainty in cluster membership of one or more stars.

The rightmost column of Fig. 5.2 displays the scatterplots of IFMR parameters

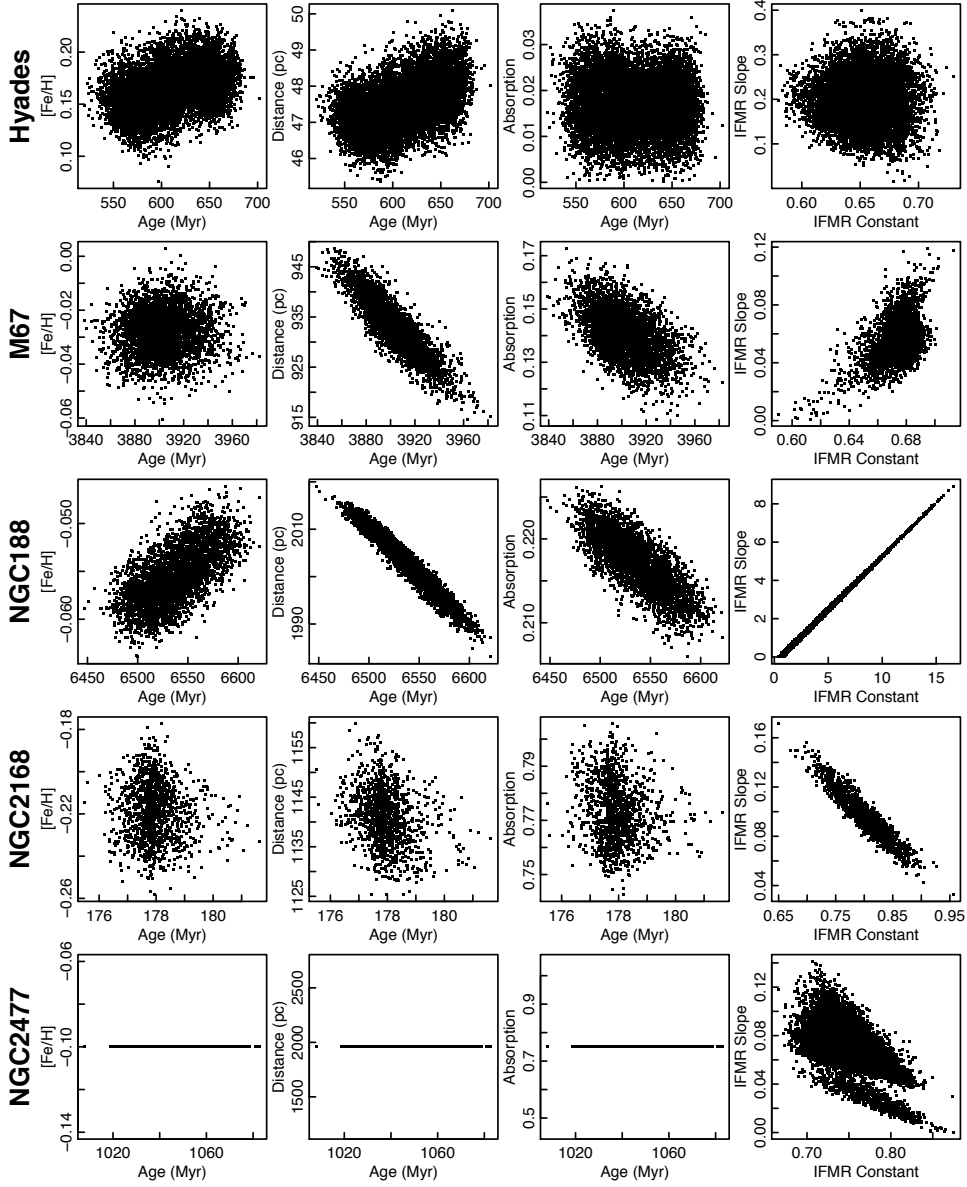


Figure 5.2: Cluster-specific results: projections of the joint posterior distributions onto the two dimensional planes of (from left to right) age–metallicity, age–distance, age–absorption, and IFMR intercept–IFMR slope for five clusters.

for the five clusters under the cluster-specific analyses. For all of the clusters except NGC 188, the range of the IFMR intercept is 0.55 to 0.95 and that of IFMR slope is 0.0 to 0.4, which are both quite consistent to the results in Salaris et al. (2009) and Williams et al. (2009). However, the IFMR parameters for NGC 188 are both far from their commonly accepted range. This appears to be because the WDs in NGC 188 have similar ZAMS masses (1.17 to 1.24 $M_{\odot}$), but their WD masses vary significantly (0.52 to 0.80 $M_{\odot}$). In the next section, we simultaneously analyse the five clusters with a hierarchical model. This allows us to borrow strength among the clusters and provides more reliable estimates of the IFMR parameters of NGC 188.

We have made pairwise scatterplots of each of these five clusters. To save space, we place them in the Appendix. For details, refer to Fig. B.1– B.5 in the Appendix. In these figures, the pairwise scatterplots of different parameters and histograms of all parameters are presented.

5.5.2 Hierarchical Analysis

In this section, we present the result obtained under the hierarchical model of Eq. (5.7). We deploy the two-stage algorithm in 5.3 to obtain the MCMC samples for all of the model parameters. For simplicity, we compute posterior sample means and standard deviations to summarise the posterior distributions of each parameter. Table 5.9 compares the statistical results for all five clusters using both the hierarchical and cluster-specific fits. In all five cases, the estimates of $\log_{10}(\text{Age})$, $m - M_V$, $[\text{Fe}/\text{H}]$ and absorption are about the same for the hierarchical and cluster-specific fits. The estimates of the IFMR parameters for NGC 188, however, vary substantially. From the cluster-specific analysis, the posterior average of the IFMR for NGC 188 is $M_{\text{WD}} = 4.522 + 2.195(M_{\text{ZAMS}} - 3.0)$. That means that a star with ZAMS mass of 3 $M_{\odot}$ has a WD mass of 4.522 $M_{\odot}$ and that the WD mass increases by 2.195 $M_{\odot}$ for each

Table 5.9: Estimates of parameters for five clusters from both the Hierarchical and Cluster-specific fits

	Cluster	$\log_{10}(\text{Age})$	$m - M_V$	$[\text{Fe}/\text{H}]$	Absorption	IFMR Intercept	IFMR Slope
Hierarchical	Hyades	8.773 ± 0.026	3.40 ± 0.03	0.157 ± 0.02	0.017 ± 0.006	0.66 ± 0.02	0.14 ± 0.05
	M67	9.591 ± 0.002	9.85 ± 0.01	-0.029 ± 0.007	0.142 ± 0.008	0.68 ± 0.01	0.06 ± 0.01
Estimates	NGC188	9.815 ± 0.002	11.51 ± 0.01	-0.056 ± 0.003	0.218 ± 0.003	0.75 ± 0.12	0.09 ± 0.05
	NGC2168	8.25 ± 0.001	10.29 ± 0.01	-0.219 ± 0.015	0.774 ± 0.012	0.79 ± 0.04	0.10 ± 0.02
	NGC2477	9.019 ± 0.004	11.46 ± 0.0	-0.1 ± 0	0.75 ± 0	0.75 ± 0.03	0.07 ± 0.02
Cluster-specific	Hyades	8.785 ± 0.028	3.41 ± 0.03	0.164 ± 0.021	0.017 ± 0.006	0.65 ± 0.02	0.20 ± 0.06
	M67	9.591 ± 0.002	9.85 ± 0.01	-0.029 ± 0.007	0.142 ± 0.008	0.68 ± 0.01	0.05 ± 0.01
Estimates	NGC188	9.815 ± 0.002	11.51 ± 0.01	-0.056 ± 0.003	0.217 ± 0.003	4.52 ± 3.13	2.20 ± 1.75
	NGC2168	8.25 ± 0.001	10.29 ± 0.01	-0.221 ± 0.015	0.775 ± 0.012	0.81 ± 0.05	0.10 ± 0.02
	NGC2477	9.019 ± 0.004	11.46 ± 0.0	-0.1 ± 0	0.75 ± 0	0.76 ± 0.03	0.07 ± 0.02

additional $M_{\odot}$ in its ZAMS mass. Unfortunately, this result is nonsense: it violates conservation of mass. The reason the cluster-specific analysis results in this bizarre IFMR is that the ZAMS masses of WDs in NGC 188 are in a narrow range, 1.17 to 1.33 $M_{\odot}$, so they fail to constrain the IFMR parameters over the whole ZAMS mass range. By contrast, the hierarchical model yields reasonable estimates for the IFMR parameters of NGC 188, $M_{\text{WD}} = 0.749 + 0.088(M_{\text{ZAMS}} - 3.0)$. For other clusters, the hierarchical and cluster-specific estimates have slight differences due to the shrinkage effects of the hierarchical model, which are further illustrated in Fig. 5.5.

In Table 5.9, another significant observation is that both the hierarchical and cluster-specific estimates of IFMR slopes of the Hyades and M67 are significantly different. One possible explanation is that M67 (age about 3.9 Gyr) is much older than the Hyades (about 590 Myr). The number of WDs in the M67, 35, is much more than that of Hyades (6), which might also lead to a shallow IFMR for M67.

From Table 5.9, ages and metallicities vary significantly across star clusters. Numerous studies have shown that the metallicity may also have an effect on the IFMR parameters of a cluster (see, e.g., Kalirai et al., 2005; Catalan et al., 2008; Meng et al., 2008; Zhao et al., 2012). We can investigate the effect of metallicity on the IFMR parameters via a Bayesian hierarchical model. To achieve this aim, we adjust the second level model in Eq. (5.7) to be

$$\boldsymbol{\alpha}_k = \begin{pmatrix} \alpha_{k0} \\ \alpha_{k1} \end{pmatrix} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\eta}_k, \boldsymbol{\Gamma}),$$

with

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}, \boldsymbol{\eta}_k = \begin{pmatrix} 1 \\ \theta_{[\text{Fe}/\text{H}],k} \end{pmatrix},$$

where the $\mathbf{B}$ matrix is the effect of metallicity $\theta_{[\text{Fe}/\text{H}],k}$ of cluster k on its IFMR

parameters α_k . Our two-stage algorithm has the capacity to fit this complicated hierarchical model. This model introduces four more parameters in the effect matrix $\mathbf{B}$, but we only have five clusters in this study. So we stick to the simple model in Eq. (5.7) for now and we will investigate the effect of metallicity on IFMR parameters once we have a sufficient number of star clusters.

Figs 5.3–5.4 plot the colour magnitude diagrams (CMD) for the five clusters. Fig. 5.3 presents the U-V and B-V CMD plots for the Hyades and the rows of Fig. 5.4 plot CMDs of clusters (from top to bottom) M67, NGC 188, NGC 2168, and NGC 2477. The solid (red) lines are from the hierarchical fits and the dashed (blue) lines are from the cluster-specific fits. In each row of Figs 5.3–5.4, the left panel displays a close up of the WD region of the CMD and the right panel shows the MS/RG stars. The CMD plots for MS/RG from both the hierarchical and cluster-specific fits are similar for all five clusters. For WDs, the same holds for all clusters except NGC 188. For NGC 188, the cluster-specific CMD (dashed blue) is quite far off from the dimmest WD, while the hierarchical CMD (solid red) is consistent with all of the cluster’s WDs. This illustrates the advantage of the hierarchical model.

Fig. 5.5 compares the estimated IFMR (plotted in lines) for the five clusters and the 68.3% contours (plotted as ovals) of the joint distributions of initial (ZAMS) and final (WD) masses for each WD in each cluster. Results for both the hierarchical (solid) and cluster-specific (thickdashed) analyses are plotted. Colors (red, blue, green, purple, black) correspond to five clusters (Hyades, M67, NGC 188, NGC 2168 and NGC 2477, respectively). It can be seen that the solid IFMR lines from the hierarchical fits tend to be in the centre and are consistent with the Williams, Salaris and Salaris piecewise IFMRs (plotted as grey solid, dashed and dotted lines, respectively). The thick dashed fitted IFMRs from the cluster-specific approach have more uncertainty. The most striking feature of this figure is that the cluster-specific

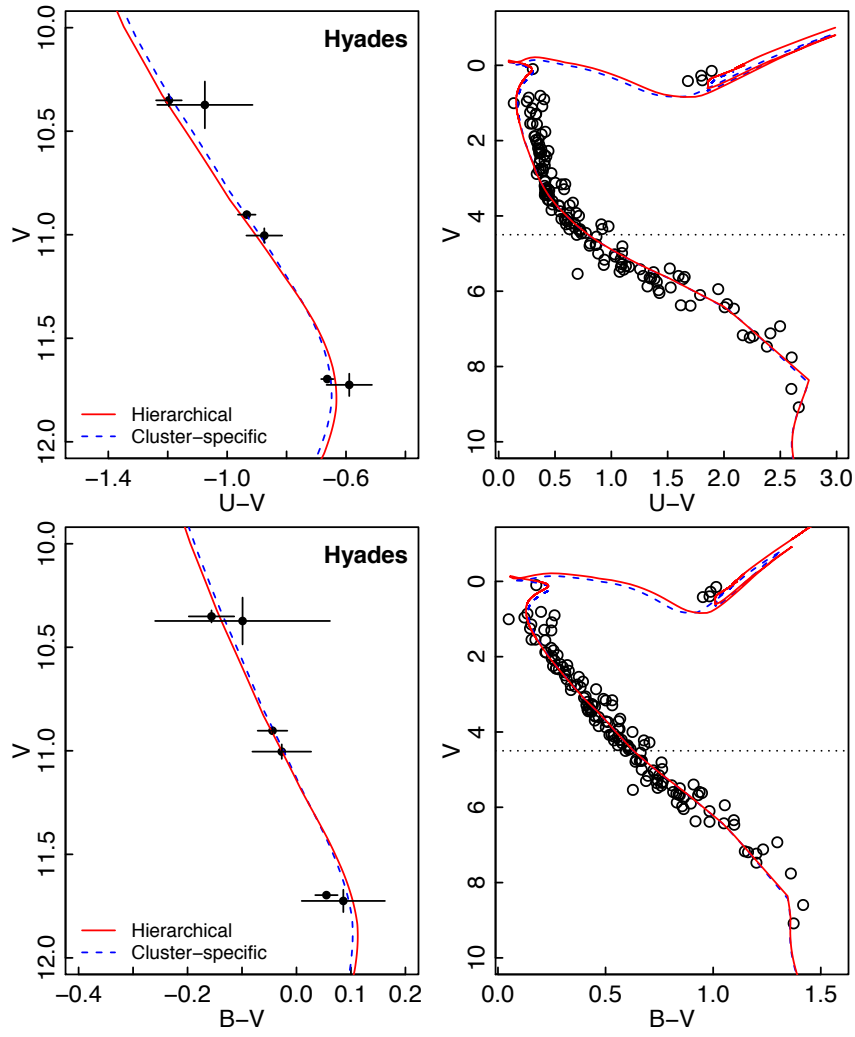


Figure 5.3: Colour magnitude diagrams for the Hyades based on both the hierarchical and cluster-specific estimates. The solid (red) lines are from the hierarchical fits and dashed (blue) lines are from the cluster-specific fits.

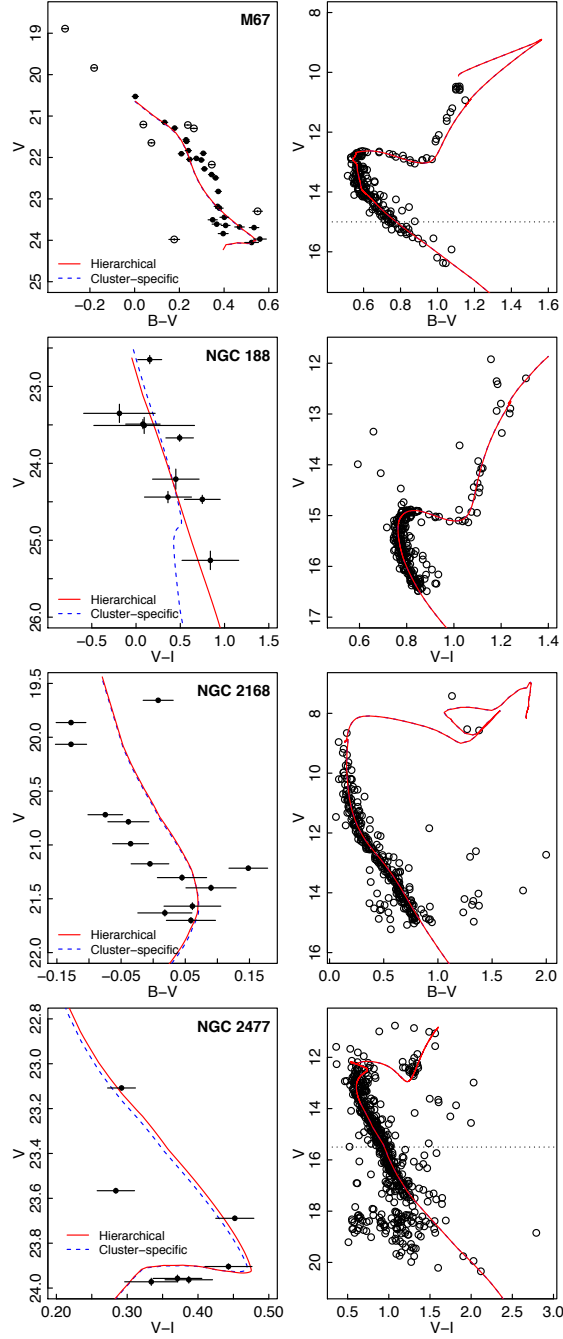


Figure 5.4: Colour magnitude diagrams for (from top to bottom) M67, NGC 188, NGC 2168, NGC 2477, based on both the hierarchical and cluster-specific estimates. The solid (red) lines are from the hierarchical fits and dashed (blue) lines are from the cluster-specific fits. The solid points in the left panel of M67 are inferred as members of the cluster with BASE-9, while the hollow points in the CMD of M67 are inferred as non-members.

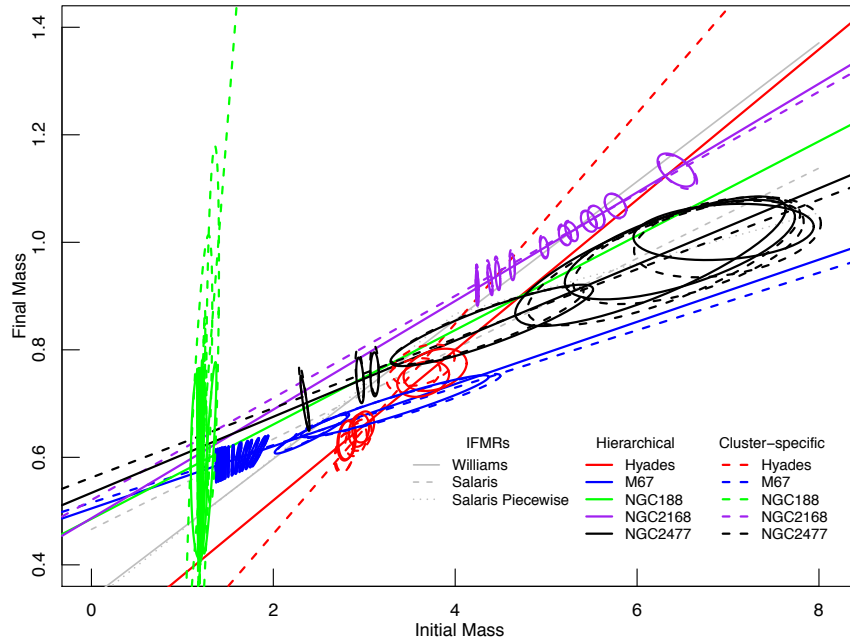


Figure 5.5: The average IFMR for five clusters obtained from both the hierarchical and cluster-specific analyses. The ovals are 68.3% contours of the joint distribution of ZAMS and WD masses for WDs in each cluster. The colour scheme of red, blue, green, purple, and black corresponds to the Hyades, M67, NGC 188, NGC 2168, and NGC 2477, respectively. The colour scheme of Williams, Salaris and Salaris piecewise are in solid, dashed and dotted grey lines.

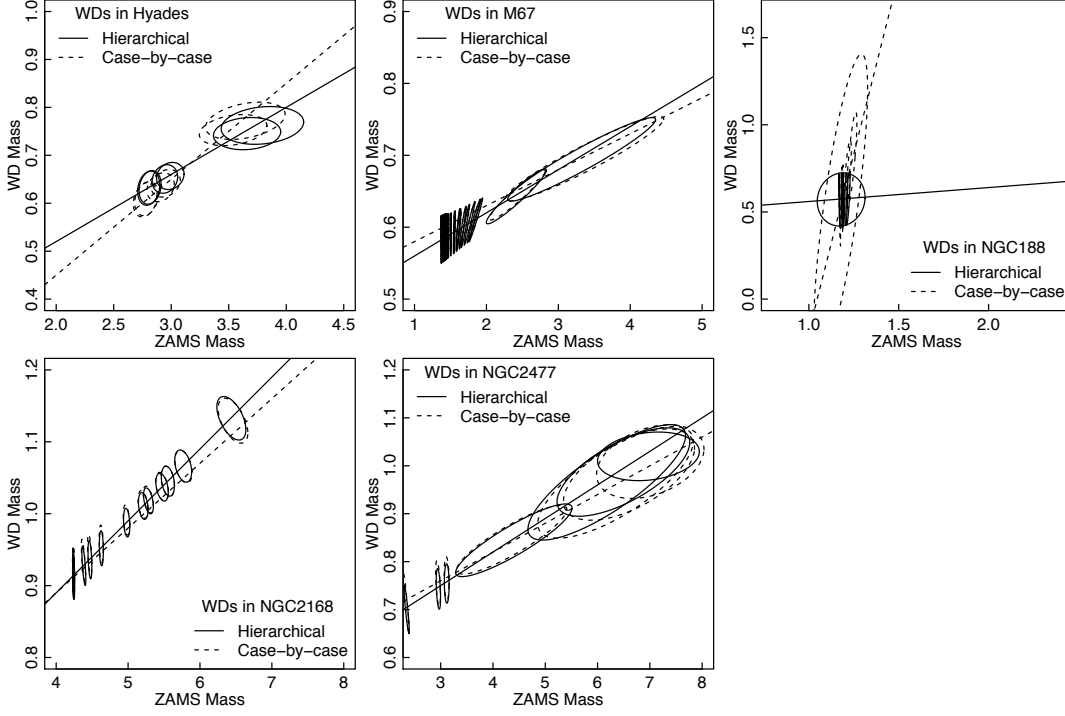


Figure 5.6: The IFMR estimates and joint distributions of the initial and final masses for WDs in each cluster from both the cluster-specific (case-by-case) and hierarchical analysis. The lines are IFMR estimates and the ovals are 68.3% contours of the joint distribution of ZAMS and WD masses for WDs in each cluster. The dashed lines and ovals are based on cluster-specific (case-by-case) fits, while their solid counterparts are based on the hierarchical analysis.

analysis of NGC 188 yields an unreasonable steep IFMR line (plotted as a dashed green line), whereas the hierarchical model produces a much shallower IFMR fit for NGC 188 (plotted as a solid green line).

Fig. 5.6 presents the comparison of the cluster-specific and hierarchical analysis for those five clusters. We can observe that the joint distributions of the initial and final masses from the cluster-specific and hierarchical analysis for clusters M67, NGC 2168 and NGC 2477 are consistent. The hierarchical and cluster-specific estimates of masses for WDs in Hyades differ slightly. The hierarchical and cluster-specific fits of masses for WDs in NGC 188 present a striking difference. With the cluster-specific

analysis, one WD in NGC 188 shows large uncertainty in its final mass, but the hierarchical fit reduces this uncertainty, providing a more precise estimate.

Fig.s 5.7–5.8 displays the time series plots of case-by-case samples of IFMR parameters of five clusters. IFMR parameters of five clusters all converge well. Fig.s 5.9–5.10 and Fig.s 5.11– 5.12 demonstrate the diagnostics of the two-stage sampler for the hierarchical analysis. All MCMC chains converge well, so the two-stage algorithm works well.

Tables 5.10–5.12 illustrate the initial masses, final masses, and membership probabilities for WDs in the five clusters.

Table 5.13 presents the estimates of the average IFMR parameters under the Bayesian hierarchical model and compares them with results from Kalirai et al. (2008) and Williams et al. (2009). In our analysis, we include five clusters: the Hyades, M67, NGC 188, NGC 2168 and NGC 2477. The 68.3% credible intervals for the IFMR intercept and slope are 0.440 ± 0.140 (i.e., $[0.30, 0.58]$) and 0.090 ± 0.040 (i.e., $[0.05, 0.13]$), respectively. The point estimates of IFMR parameters from Kalirai et al. (2008) and Williams et al. (2009) falls into the credible intervals from our hierarchical model, so we consider that the average IFMR from our analysis is consistent with results from these studies. Additionally, from the simulation study, the actual coverage probabilities of our interval estimates of IFMR intercept and slope are 76% and 68%, respectively, and they are quite close to the nominal level 68.3%.

The point estimate of IFMR intercept from our FB analysis is 0.440, greater than intercepts from both Kalirai et al. (2008), 0.394, and Williams et al. (2009), 0.339. On the other hand, our estimate of IFMR slope is the shallowest one, 0.090, and Williams et al. (2009) has the steepest IFMR, at 0.129.

The most obvious characteristic of our result lies in the large error bars, about 6~10 times the error bars in other analyses. There are three reasons for this. 1.) The

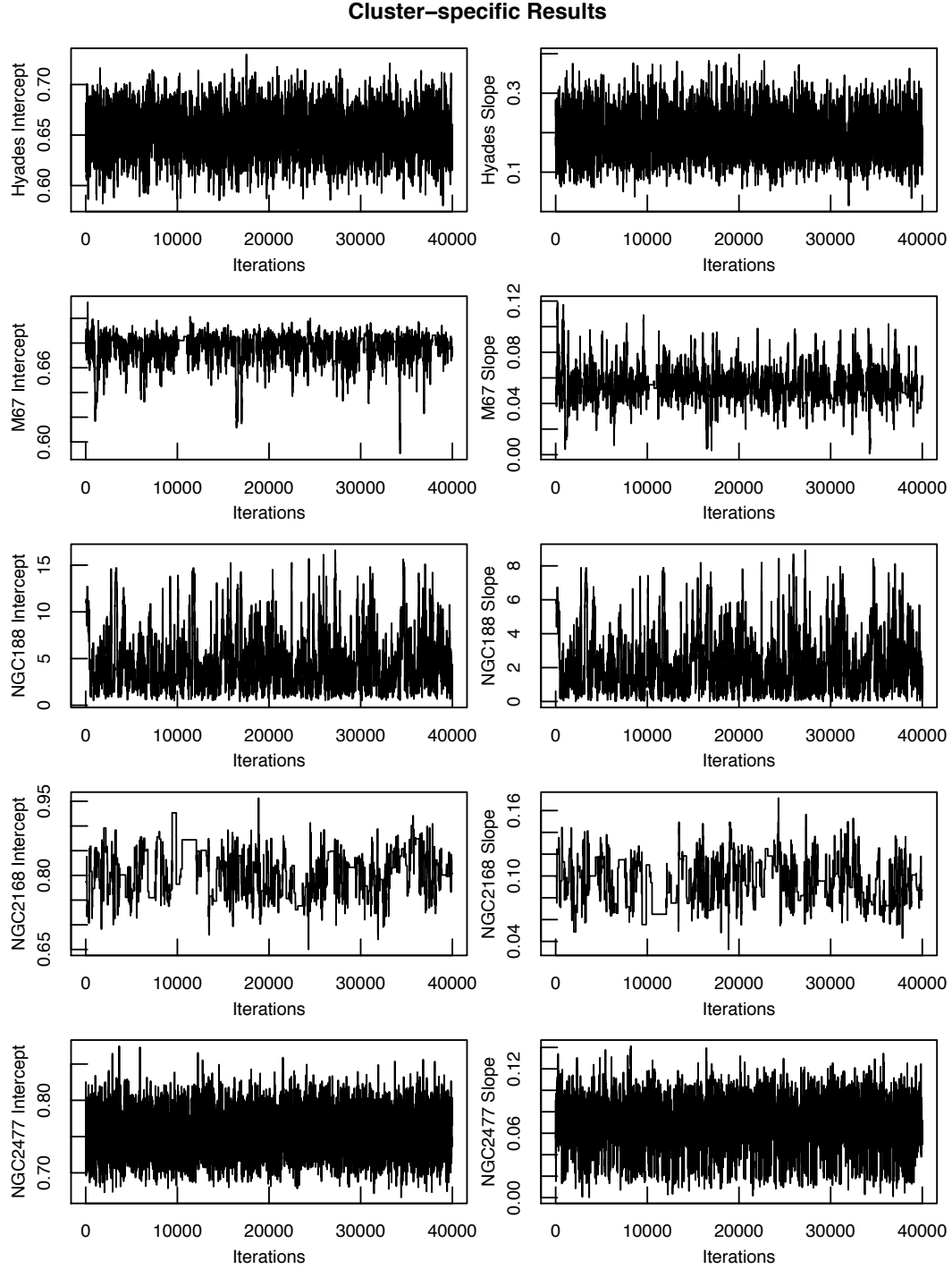


Figure 5.7: The time series plots of IFMR parameters from the case-by-case (cluster-specific) analysis.

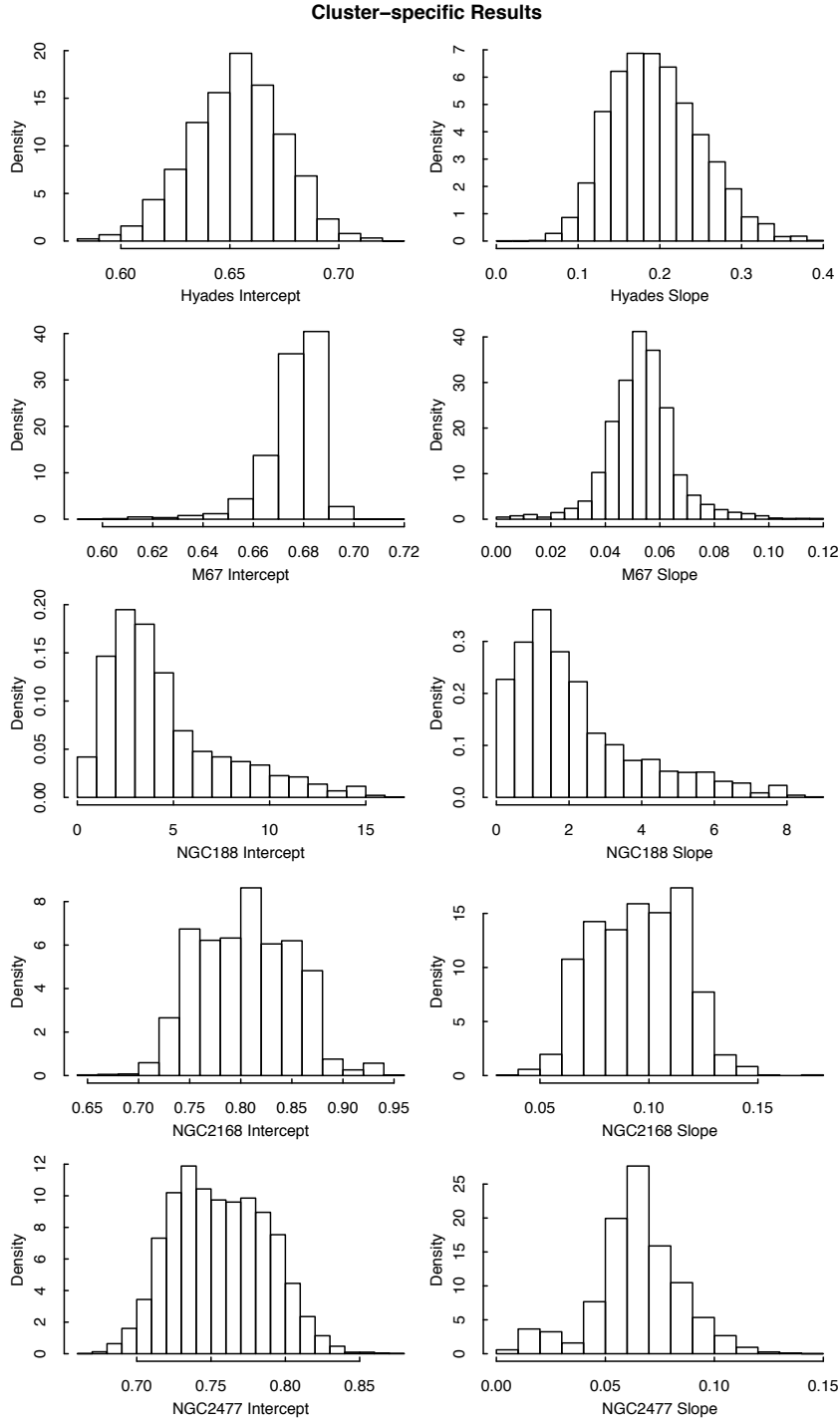


Figure 5.8: Histograms of IFMR parameters from the case-by-case (cluster-specific) analysis.

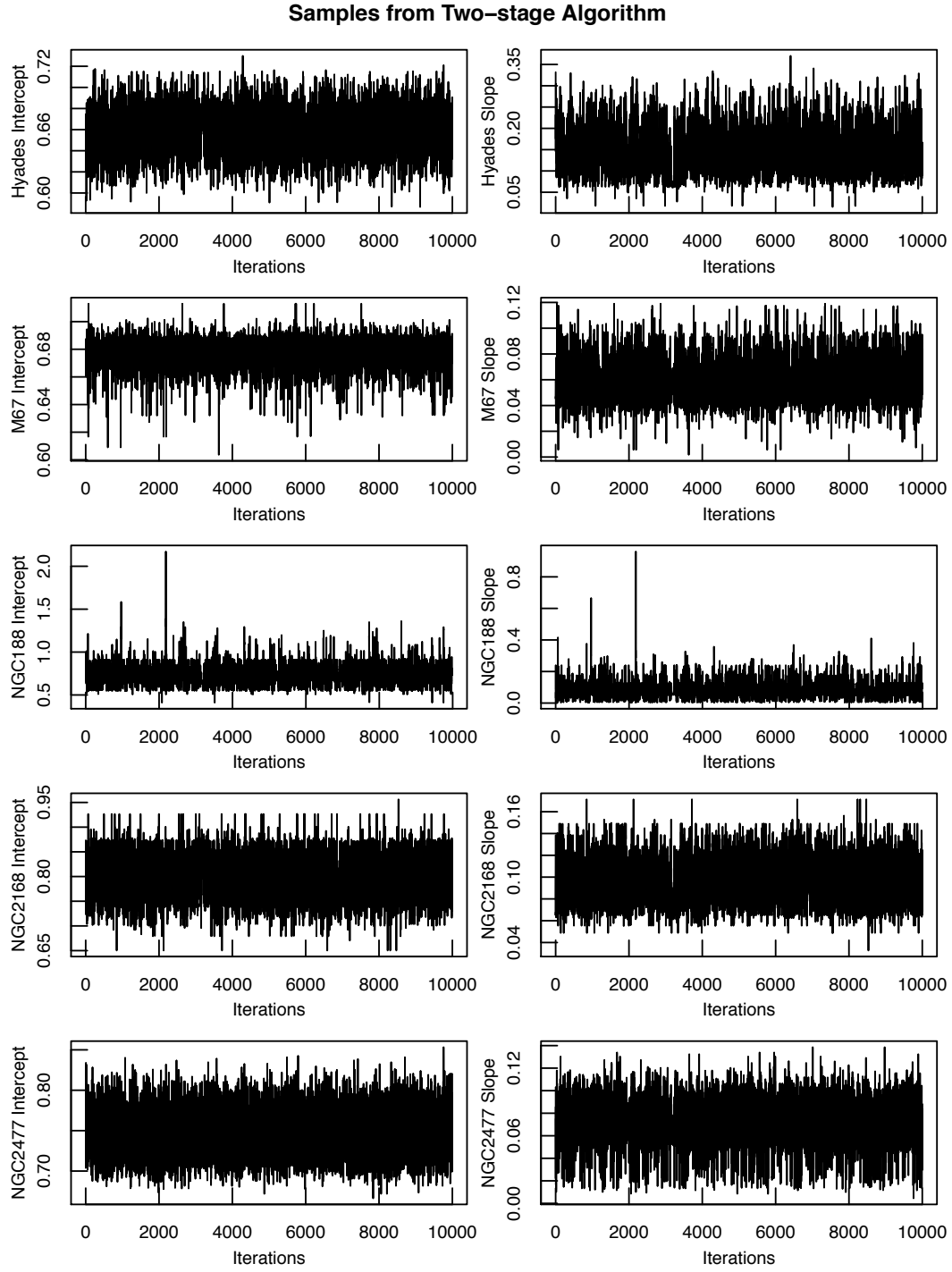


Figure 5.9: The time series plots of IFMR parameters from the hierarchical analysis.

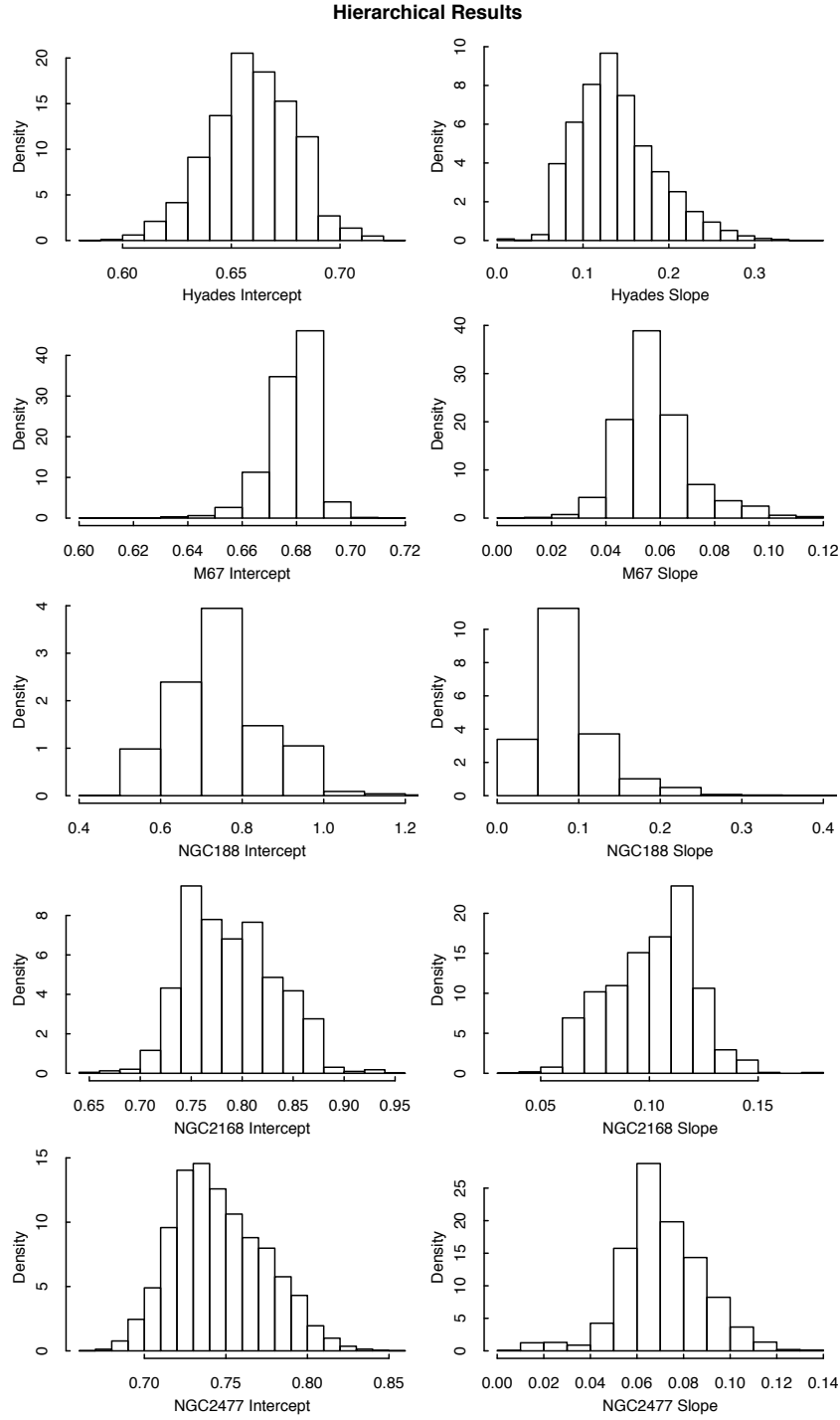


Figure 5.10: Histograms of IFMR parameters from the hierarchical analysis.

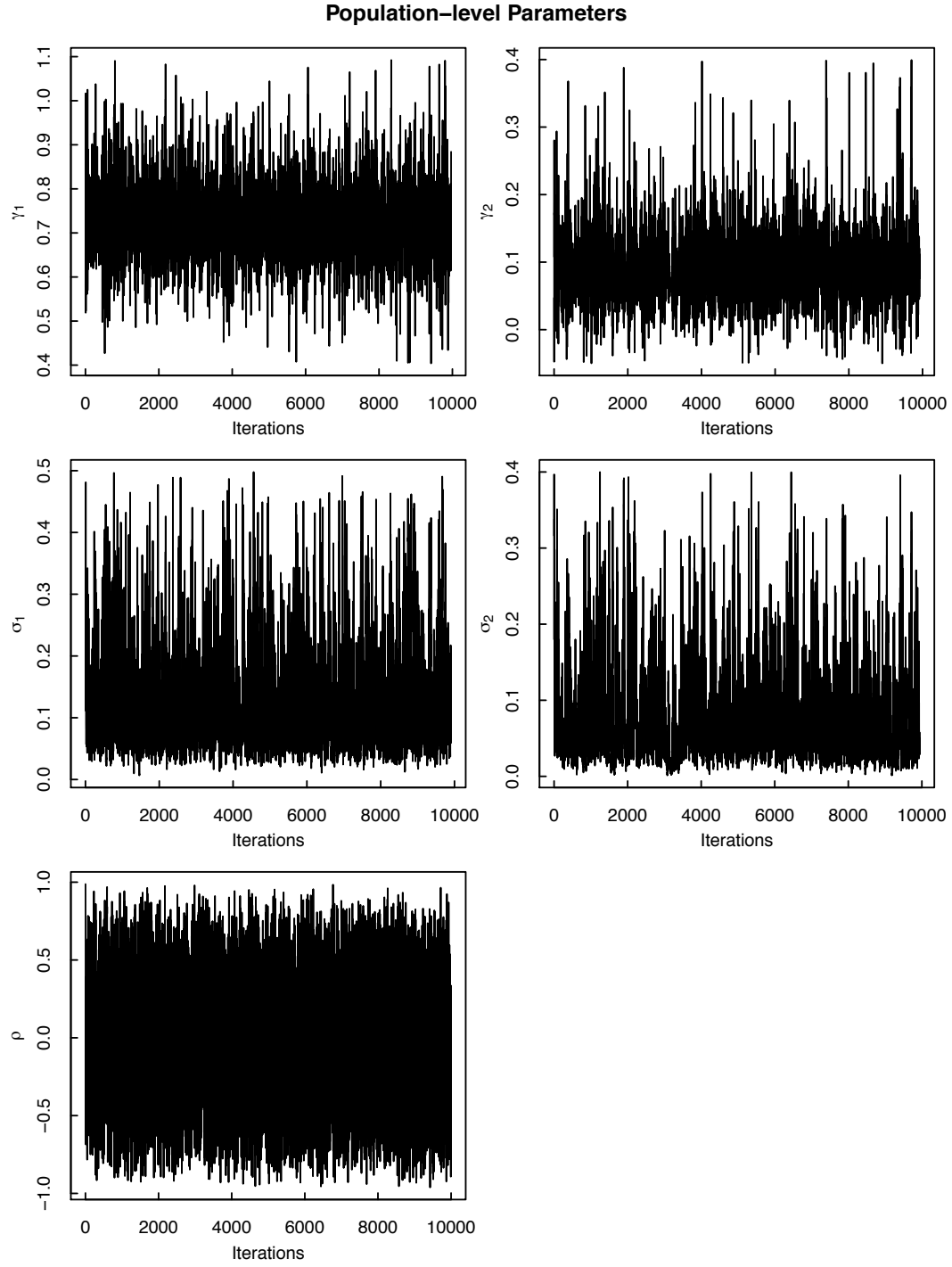


Figure 5.11: Time series plots for population-level parameters in the IFMR hierarchical model.

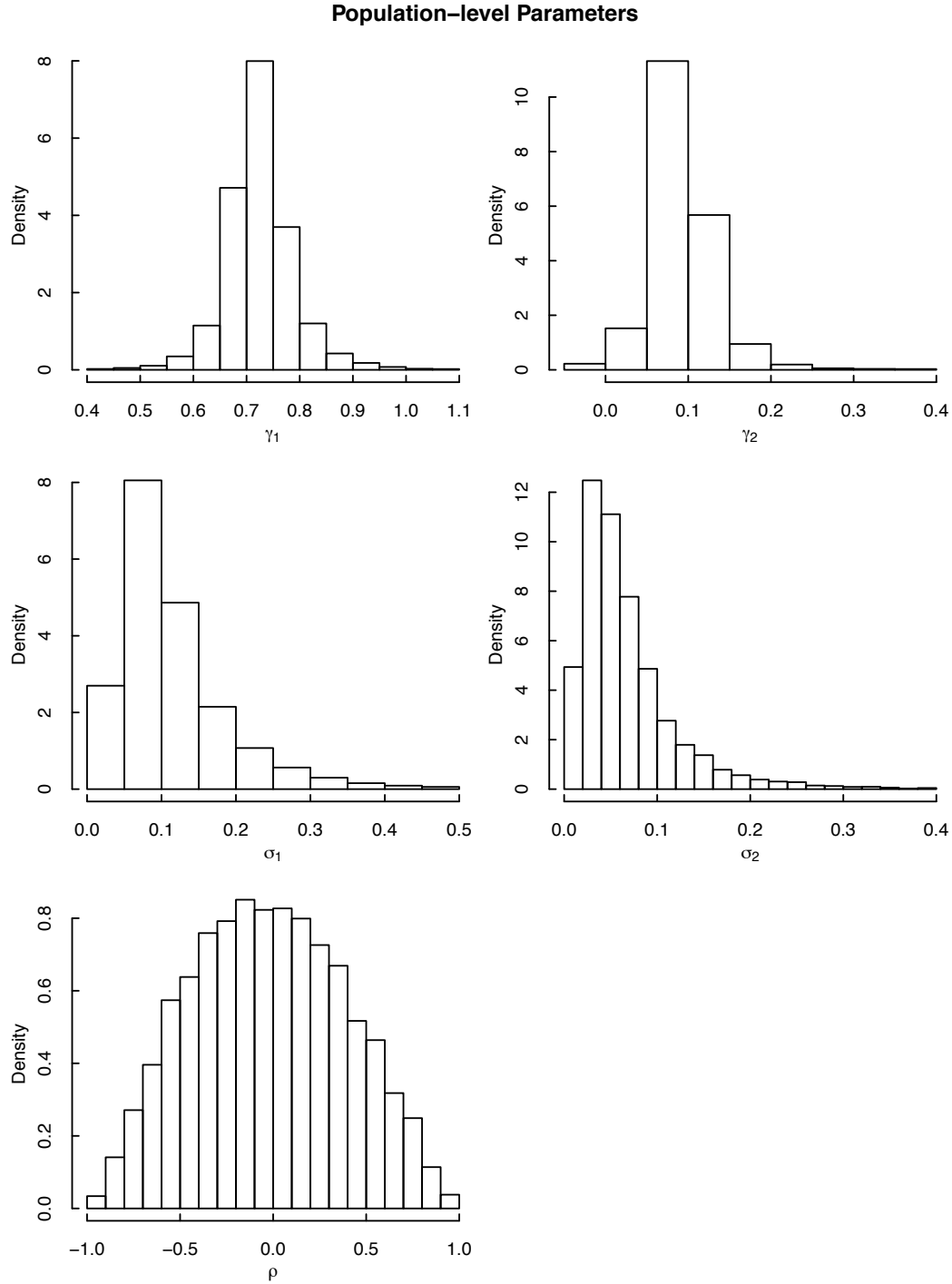


Figure 5.12: Histograms for population-level parameters in the IFMR hierarchical model.

Table 5.10: Results from both the Hierarchical and Cluster-specific fits for WDs in Hyades

Cluster	WD	Hierarchical Estimates			Cluster-specific Estimates		
		ZAMS Mass	WD Mass	Mem. Prob.	ZAMS Mass	WD Mass	Mem. Prob.
Hyades	HZ14	2.801 ± 0.064	0.631 ± 0.024	1	2.776 ± 0.067	0.607 ± 0.024	1
	VR16	2.813 ± 0.064	0.632 ± 0.024	1	2.787 ± 0.067	0.609 ± 0.024	1
	HZ7	2.942 ± 0.074	0.65 ± 0.02	1	2.91 ± 0.078	0.633 ± 0.02	1
	VR7	2.988 ± 0.079	0.657 ± 0.019	1	2.954 ± 0.084	0.641 ± 0.019	1
	HZ4	3.779 ± 0.229	0.763 ± 0.026	1	3.649 ± 0.228	0.772 ± 0.025	1
	LB227	3.648 ± 0.188	0.746 ± 0.022	1	3.543 ± 0.194	0.752 ± 0.021	1

Table 5.11: Results from the Hierarchical and Cluster-specific fits on WDs in M67

Cluster	WD	Hierarchical Estimates			Cluster-specific Estimates		
		ZAMS Mass	WD Mass	Mem. Prob.	ZAMS Mass	WD Mass	Mem. Prob.
M67	WD 1	1.615 ± 0.021	0.598 ± 0.019	1	1.619 ± 0.021	0.602 ± 0.017	1
	WD 2	1.438 ± 0.003	0.588 ± 0.02	1	1.438 ± 0.003	0.593 ± 0.018	1
	WD 3	1.429 ± 0.003	0.588 ± 0.02	1	1.429 ± 0.003	0.592 ± 0.018	1
	WD 4	1.412 ± 0.003	0.587 ± 0.02	0	1.412 ± 0.003	0.591 ± 0.018	0
	WD 5	1.364 ± 0.003	0.584 ± 0.021	0.209	1.364 ± 0.003	0.589 ± 0.019	0.191
	WD 6	1.426 ± 0.003	0.587 ± 0.02	1	1.426 ± 0.003	0.592 ± 0.018	1
	WD 7	1.414 ± 0.003	0.587 ± 0.02	1	1.414 ± 0.003	0.591 ± 0.018	1
	WD 8	1.396 ± 0.003	0.586 ± 0.02	1	1.395 ± 0.003	0.59 ± 0.018	1
	WD 9	1.389 ± 0.003	0.585 ± 0.02	1	1.388 ± 0.003	0.59 ± 0.018	1
	WD 10	1.556 ± 0.006	0.595 ± 0.019	1	1.558 ± 0.006	0.599 ± 0.017	1
	WD 11	1.431 ± 0.003	0.588 ± 0.02	0.983	1.431 ± 0.003	0.592 ± 0.018	0.981
	WD 12	1.394 ± 0.003	0.586 ± 0.02	0.329	1.393 ± 0.003	0.59 ± 0.018	0.253
	WD 13	1.37 ± 0.003	0.584 ± 0.021	1	1.37 ± 0.003	0.589 ± 0.018	1
	WD 14	1.506 ± 0.003	0.592 ± 0.019	0.999	1.507 ± 0.003	0.596 ± 0.017	0.999
	WD 15	1.454 ± 0.003	0.589 ± 0.02	1	1.454 ± 0.003	0.594 ± 0.018	1
	WD 16	1.828 ± 0.056	0.61 ± 0.019	1	1.838 ± 0.051	0.614 ± 0.017	1
	WD 17	1.44 ± 0.003	0.588 ± 0.02	1	1.44 ± 0.003	0.593 ± 0.018	1
	WD 18	1.473 ± 0.003	0.59 ± 0.019	0.595	1.473 ± 0.003	0.595 ± 0.017	0.544
	WD 19	1.549 ± 0.005	0.594 ± 0.019	1	1.55 ± 0.006	0.599 ± 0.017	1
	WD 20	1.389 ± 0.003	0.585 ± 0.02	0.173	1.388 ± 0.003	0.59 ± 0.018	0.221
	WD 21	1.363 ± 0.003	0.584 ± 0.021	0.033	1.363 ± 0.003	0.589 ± 0.019	0.03
	WD 22	2.424 ± 0.27	0.644 ± 0.025	1	2.424 ± 0.252	0.645 ± 0.023	1
	WD 23	3.335 ± 0.682	0.695 ± 0.038	1	3.365 ± 0.747	0.695 ± 0.04	1
	WD 24	1.398 ± 0.003	0.586 ± 0.02	0.092	1.398 ± 0.003	0.591 ± 0.018	0.056
	WD 25	1.671 ± 0.03	0.601 ± 0.019	0.999	1.677 ± 0.029	0.605 ± 0.017	1
	WD 26	1.768 ± 0.047	0.607 ± 0.019	1	1.777 ± 0.044	0.61 ± 0.017	1
	WD 27	1.466 ± 0.003	0.59 ± 0.02	0.976	1.466 ± 0.003	0.594 ± 0.018	0.971
	WD 28	1.708 ± 0.036	0.603 ± 0.019	1	1.715 ± 0.034	0.607 ± 0.017	1
	WD 29	1.607 ± 0.019	0.598 ± 0.019	1	1.611 ± 0.019	0.602 ± 0.017	1
	WD 30	1.812 ± 0.055	0.609 ± 0.019	0	1.822 ± 0.05	0.613 ± 0.017	0
	WD 31	1.437 ± 0.003	0.588 ± 0.02	1	1.437 ± 0.003	0.593 ± 0.018	1
	WD 32	1.856 ± 0.058	0.612 ± 0.019	0.996	1.867 ± 0.052	0.615 ± 0.017	0.998
	WD 33	1.412 ± 0.003	0.587 ± 0.02	1	1.412 ± 0.003	0.591 ± 0.018	1
	WD 34	1.449 ± 0.003	0.589 ± 0.02	0.098	1.449 ± 0.003	0.593 ± 0.018	0.073
	WD 35	1.599 ± 0.016	0.597 ± 0.019	0.093	1.602 ± 0.017	0.601 ± 0.017	0.06

Table 5.12: Results from the Hierarchical and Cluster-specific fits for WDs in NGC188, NGC2168 and NGC2477

Cluster	WD	Hierarchical Estimates			Cluster-specific Estimates		
		ZAMS Mass	WD Mass	Mem. Prob.	ZAMS Mass	WD Mass	Mem. Prob.
NGC188	WD 1	1.169 ± 0.003	0.588 ± 0.117	1	1.171 ± 0.003	0.51 ± 0.128	1
	WD 2	1.184 ± 0.003	0.59 ± 0.116	1	1.185 ± 0.003	0.538 ± 0.122	1
	WD 3	1.187 ± 0.002	0.59 ± 0.116	1	1.188 ± 0.002	0.544 ± 0.121	1
	WD 4	1.184 ± 0.06	0.59 ± 0.117	1	1.177 ± 0.1	0.519 ± 0.326	1
	WD 5	1.191 ± 0.002	0.59 ± 0.116	1	1.191 ± 0.002	0.551 ± 0.12	1
	WD 6	1.208 ± 0.006	0.592 ± 0.116	1	1.21 ± 0.01	0.595 ± 0.13	1
	WD 7	1.219 ± 0.009	0.592 ± 0.116	1	1.226 ± 0.017	0.639 ± 0.157	1
	WD 8	1.219 ± 0.009	0.592 ± 0.116	1	1.224 ± 0.015	0.628 ± 0.146	1
	WD 9	1.302 ± 0.048	0.599 ± 0.118	1	1.333 ± 0.046	0.845 ± 0.22	1
NGC2168	WD 1	4.627 ± 0.019	0.954 ± 0.016	1	4.625 ± 0.019	0.96 ± 0.017	1
	WD 2	5.762 ± 0.078	1.068 ± 0.015	1	5.764 ± 0.079	1.067 ± 0.015	1
	WD 3	5.279 ± 0.045	1.02 ± 0.012	1	5.277 ± 0.047	1.021 ± 0.012	1
	WD 4	4.244 ± 0.008	0.916 ± 0.022	1	4.243 ± 0.008	0.923 ± 0.023	1
	WD 5	4.24 ± 0.008	0.915 ± 0.022	1	4.239 ± 0.008	0.923 ± 0.023	1
	WD 6	5.208 ± 0.045	1.013 ± 0.012	1	5.203 ± 0.046	1.015 ± 0.013	1
	WD 7	6.43 ± 0.135	1.134 ± 0.02	1	6.45 ± 0.138	1.131 ± 0.021	1
	WD 8	4.468 ± 0.018	0.938 ± 0.018	1	4.465 ± 0.02	0.945 ± 0.019	1
	WD 9	4.24 ± 0.008	0.915 ± 0.022	1	4.239 ± 0.008	0.923 ± 0.023	1
	WD 10	5.469 ± 0.057	1.039 ± 0.013	1	5.467 ± 0.059	1.039 ± 0.013	1
	WD 11	5.549 ± 0.061	1.047 ± 0.014	1	5.546 ± 0.062	1.047 ± 0.014	1
	WD 12	4.383 ± 0.019	0.93 ± 0.019	1	4.378 ± 0.02	0.936 ± 0.02	1
	WD 13	4.979 ± 0.029	0.99 ± 0.013	1	4.976 ± 0.03	0.993 ± 0.014	1
NGC2477	WD 1	2.348 ± 0.033	0.701 ± 0.034	1	2.337 ± 0.036	0.714 ± 0.04	1
	WD 2	3.114 ± 0.031	0.754 ± 0.027	1	3.112 ± 0.031	0.763 ± 0.03	1
	WD 3	2.954 ± 0.028	0.743 ± 0.028	0.95	2.949 ± 0.03	0.753 ± 0.031	0.964
	WD 4	6.951 ± 0.654	1.019 ± 0.034	1	6.994 ± 0.678	1.01 ± 0.049	1
	WD 5	6.471 ± 0.835	0.986 ± 0.061	1	6.565 ± 0.836	0.981 ± 0.063	1
	WD 6	6.147 ± 0.981	0.965 ± 0.08	1	6.294 ± 0.99	0.964 ± 0.078	1
	WD 7	4.402 ± 0.735	0.845 ± 0.05	1	4.354 ± 0.698	0.845 ± 0.048	1

Table 5.13: Estimates of the average IFMR parameters with error bars from the hierarchical model and comparisons with other results

	IFMR Intercept	IFMR Slope
Hierarchical Model	0.440 ± 0.140	0.090 ± 0.040
Kalirai et al. (2008)	0.394 ± 0.025	0.109 ± 0.007
Williams et al. (2009)	0.339 ± 0.015	0.129 ± 0.004

number of clusters included in our analysis is only five. By contrast, Kalirai et al. (2008) and Williams et al. (2009) employed 13 and 11 stellar clusters, respectively. So they have more data to constrain the IFMR parameters, which leads to the narrow error bars in their studies. 2.) In my analysis, clusters have different IFMR parameters and I am reporting the mean IFMR of five clusters. However, Kalirai et al. (2008) and Williams et al. (2009) assume all clusters have the same IFMR. This is a much simple model that can be fit with less data. 3.) Kalirai et al. (2008) and Williams et al. (2009) used linear regression technique to fit the initial and final masses of WDs in their clusters, so their estimates of IFMR parameters are mainly subject to uncertainties of WDs' initial and final masses. Their estimates are also indirectly affected by the ages and distances of clusters. However, we utilise a Bayesian hierarchical model, which take account of uncertainties of all parameters such as ages, distances, metallicities, etc.

Lastly we emphasis that the aim of this research is to develop statistical techniques that improve the estimates of IFMR parameters for star clusters. So we concentrate on the statistical side of the work. The astrophysical results are still preliminary in terms of the number of clusters and model assumptions.

5.6 Sensitivity Analysis

In this section, we present the sensitivity analysis in the hierarchical analysis of IFMR parameters.

5.6.1 Sensitivity to Prior Distribution

Here we investigate whether the hierarchical analysis in Eq. 5.7 is sensitive to the prior distribution on $\mathbf{\Gamma}$. We use the marginally non-informative prior distribution proposed by Huang and Wand (2013), i.e.,

$$\mathbf{\Gamma} \mid \lambda_1, \lambda_2 \sim \text{Inverse Wishart} \left(2\nu \begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{pmatrix}, \nu + 1 \right),$$

$$\lambda_1, \lambda_2 \sim \text{Inverse Gamma}(1/2, 1/\Lambda),$$

λ_1 and λ_2 are hyper-parameters, and they independently follow the same inverse gamma distribution with its first parameter fixed at $1/2$ and second parameter $1/\Lambda$ a small positive number, i.e., large positive Λ . Huang and Wand (2013) suggests that $\nu = 2$ leads to a marginal uniform distribution for correlation ρ and arbitrarily large positive Λ leads to arbitrarily weakly informative prior distributions for σ_1 and σ_2 . Because $\nu = 2$ is necessary to have a marginally non-informative prior distribution on ρ , so in this hierarchical analysis, we take $\nu = 2$. As for Λ , we choose four large values: $10^3, 10^4, 10^5, 10^6$ and fit the hierarchical model with these values, then compare the MCMC draws of IFMR parameters of the five included clusters.

Figs 5.13–5.15 present the QQ plots of IFMR parameters from the hierarchical fits when Λ takes different values. All points in these QQ plots lie close to the 45° red line, meaning that MCMC draws of IFMR parameters are essentially the same. Though there are some small deviations from the red line, they are mainly caused by Monte Carlo errors. So we conclude that the hierarchical result is not sensitive to the choice of value of Λ provided that $\Lambda \geq 10^3$.

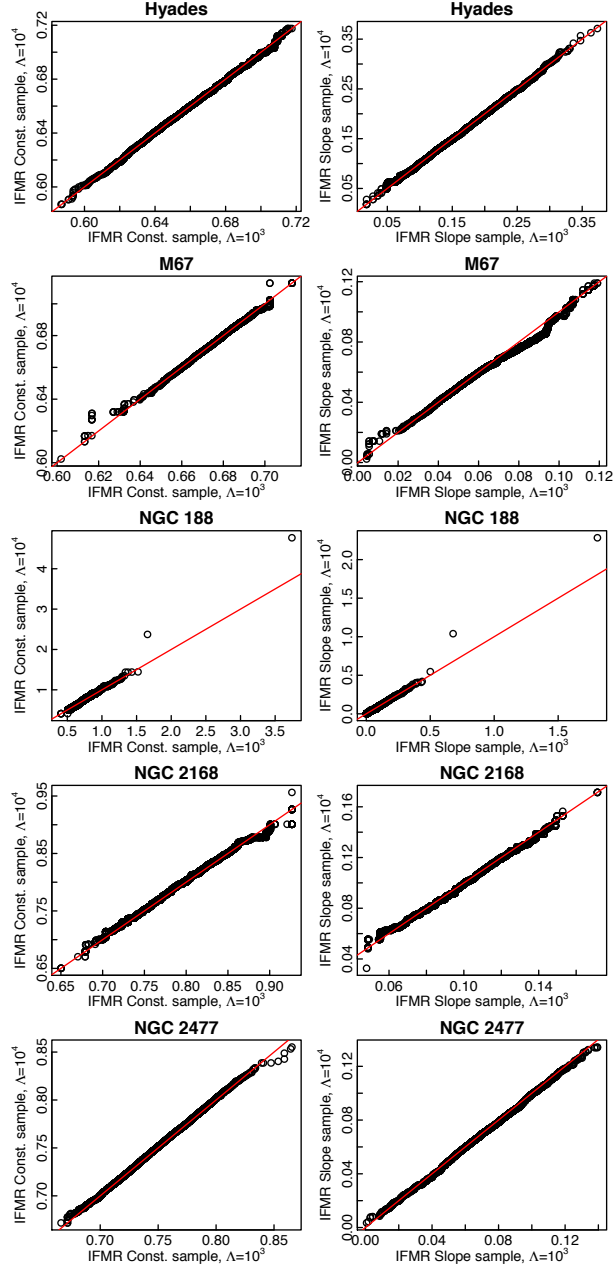


Figure 5.13: Quantile-quantile plots of MCMC samples of IFMR parameters of five clusters. The x-axis and y-axis are quantiles corresponding to fits with $\Lambda = 10^3$ and $\Lambda = 10^4$, respectively. The red line represents that quantiles from samples in x and y axes are equal.

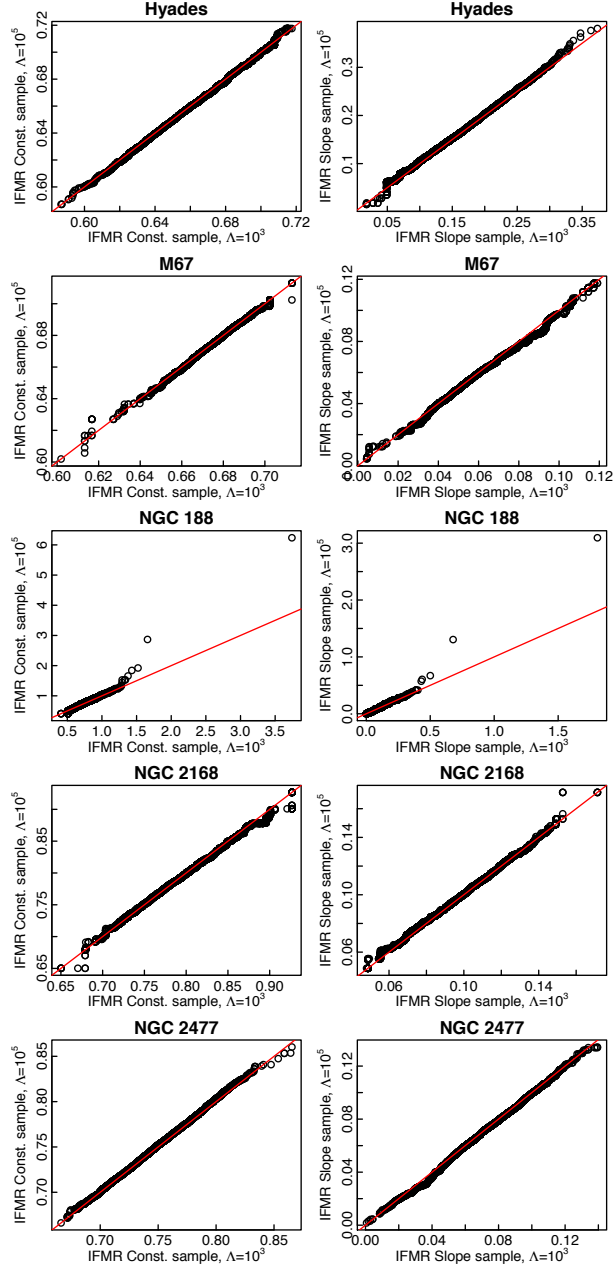


Figure 5.14: Quantile-quantile plots of MCMC samples of IFMR parameters of five clusters. The x-axis and y-axis are quantiles corresponding to fits with $\Lambda = 10^3$ and $\Lambda = 10^5$, respectively. The red line represents that quantiles from samples in x and y axes are equal.

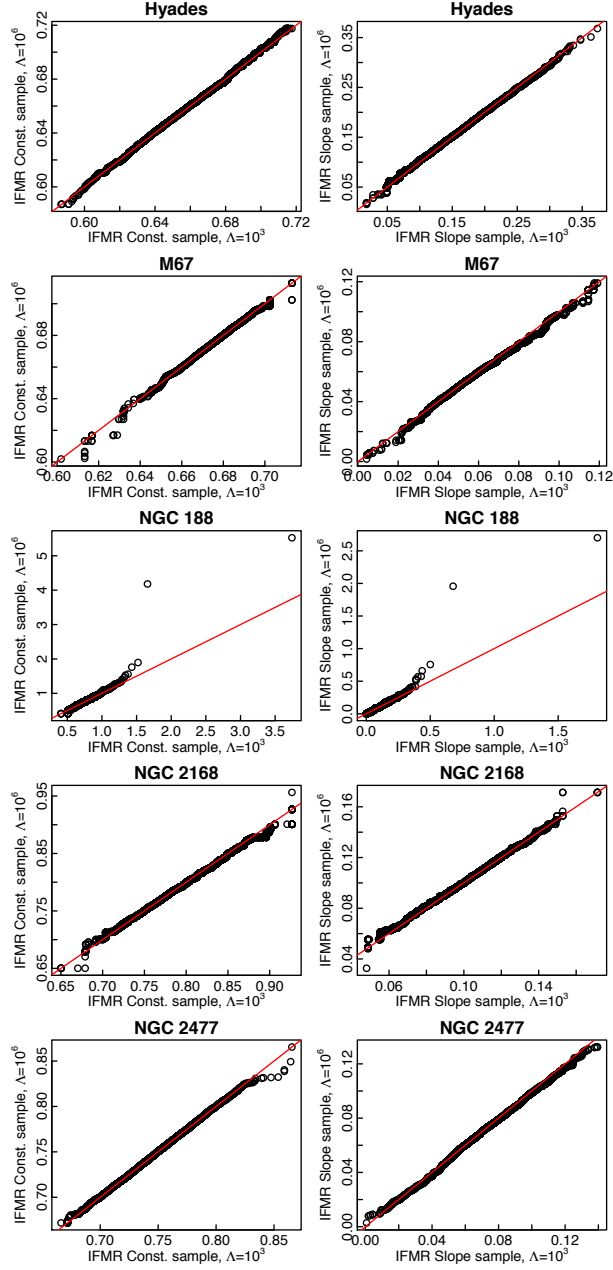


Figure 5.15: Quantile-quantile plots of MCMC samples of IFMR parameters of five clusters. The x-axis and y-axis are quantiles corresponding to fits with $\Lambda = 10^3$ and $\Lambda = 10^6$, respectively. The red line represents that quantiles from samples in x and y axes are equal.

5.6.2 Sensitivity to Membership of WDs in M67

In Section 5.5, Table 5.11 presents the hierarchical and case-by-case estimates of initial and final masses and membership probabilities for 35 WDs in M67, among which nine WDs have posterior membership probabilities less than 0.5 and are classified as non-members or field stars. The other 26 WDs are inferred as members of M67. In this section we investigate whether the membership of these nine WDs affects the case-by-case posterior distribution of cluster M67.

In Section 5.5, when performing the case-by-case analysis on the cluster M67 with BASE-9, we set the prior membership probabilities of all WDs based on other research (e.g., Bellini et al., 2010a,b; Williams et al., 2013; Barnes et al., 2016). For clarity, we call this analysis the original fit. Then we fit M67 under two other circumstances: 1.) Case I: setting all WDs to have a 100% prior probability of being a member in M67, and 2.) Case II: letting nine non-members in the original fit (Table 5.11) have prior membership probabilities being 0 and members have prior probabilities being 1. In other words, the first case forces all WDs in M67 to be its members whereas the second case removes nine non-members and sets other 26 as its members.

Table 5.14 presents the point estimates (with error bars) of cluster parameters of M67 under three settings. The contraction in the last two columns are IFMR shallowest constant and slope, respectively. From this table, the case-by-case estimates of IFMR slope vary under different settings. In Case I when all WDs are forced to be members of M67, the IFMR slope is the steepest, while in the original fit, IFMR is shallowest. The estimates of other parameters are similar under three settings except the age. The estimate of age from the original fit is consistent to that from the Case II, i.e., when the nine inferred non-member WDs from the original fit are excluded. The age estimate from Case I is younger than the others. However, the

Table 5.14: Estimates with error bars of cluster parameters (age, distance modulus, metallicity, absorption, IFMR constant and slope) for M67 under three cases.

Settings	Age (Gyr)	$m - M_V$	[Fe/H] ¹	Abs. ¹	IFMR Cons.	IFMR Slo.
Original	3.90 ± 0.02	9.85 ± 0.01	-0.03	0.14	0.68 ± 0.01	0.05 ± 0.01
Case I	3.07 ± 0.03	9.86 ± 0.02	-0.01	0.15	0.67 ± 0.01	0.10 ± 0.01
Case II	3.89 ± 0.02	9.86 ± 0.01	-0.04	0.16	0.68 ± 0.01	0.07 ± 0.01

¹: The standard errors for estimates of metallicity and absorption under three fits are the same, 0.01.

age estimates from the original fit and Case I are close to results obtained through other approaches, such as the spectroscopy (Williams et al., 2013) and proper motion (Bellini et al., 2010b).

Fig. 5.16 presents the CMD plots of the original and two other pilot fits. The black lines are from the fitted model under the original fit, and the red and blue lines are from models under Case I and II, respectively. From this plot, both fits from Case I (red line) and II (blue line) miss the main sequence turnoff stars, sub-giant branch and base of the red giant branch. By contrast, the model under the original fit (black line) matches both parts of the cluster well. So the original fit is more reliable than the other two fits.

In summary, the membership of these WDs affect the posterior distribution of cluster parameters, yet our further analysis supports the original fit because it best matches the photometric data among those three fits.

5.7 Concluding Remarks

The IFMR is a mapping between the ZAMS mass of a star to its WD mass and is vital for the determination of parameters in stellar evolution (Weidemann, 2000; Williams et al., 2004; Meng et al., 2008; Williams et al., 2009). Many attempts have been made to study the IFMR using data from different star clusters, leading to numerous

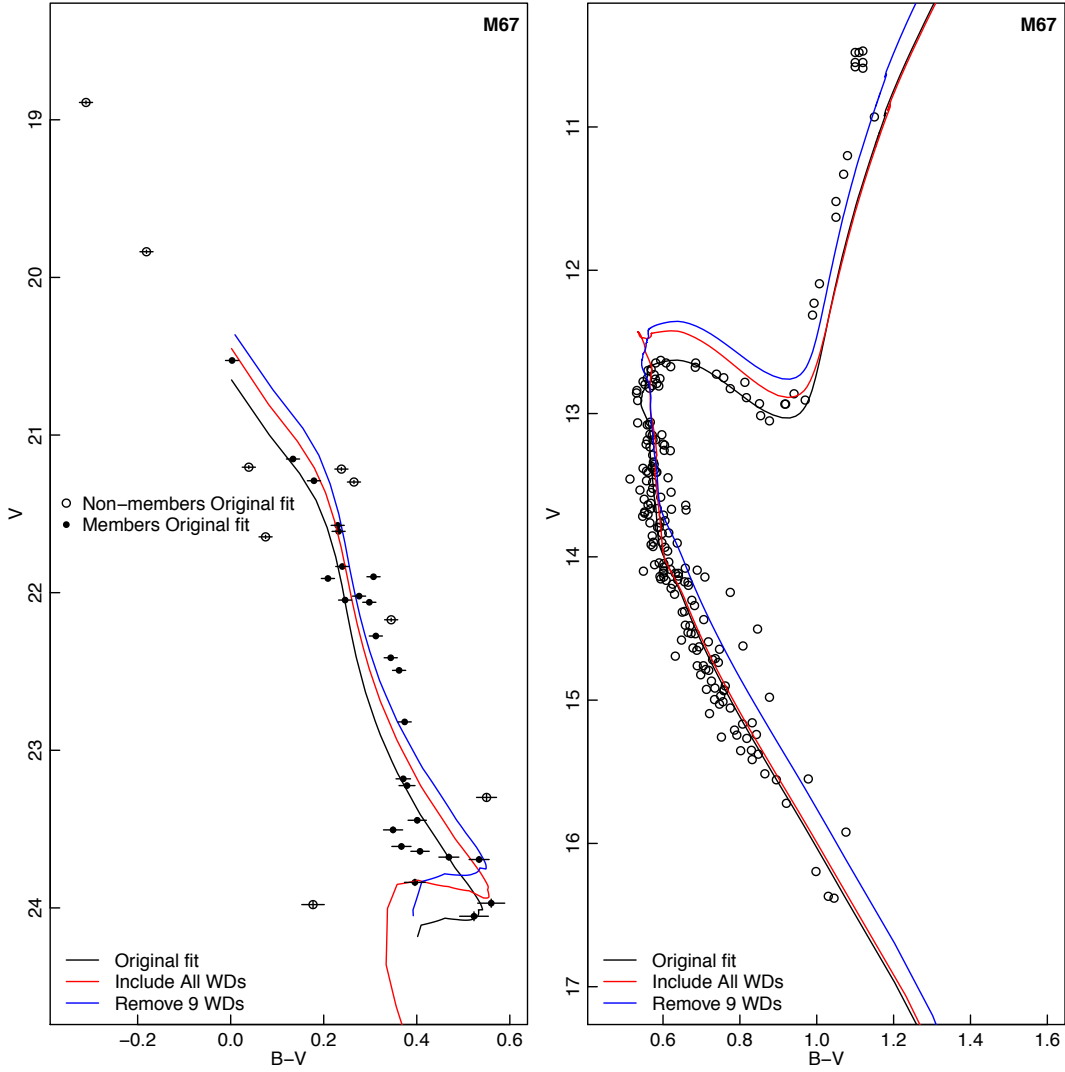


Figure 5.16: Colour-magnitude diagram (CMD) plots of three fits. The black lines are from the original fit, red lines from the fit that includes all WDs in M67, blue lines from the fit that removes nine non-members. In the left panel, hollow points represent non-member WDs from the original fit and solid points are members. Observed error bars are added to the WDs.

versions of the IFMR (See e.g., Kalirai et al., 2005, 2008; Catalan et al., 2008; Salaris et al., 2009). Stein et al. (2013) proposed a principled Bayesian analysis for the IFMR parameters of a star cluster and implemented this method with a black-box package, called BASE-9. This package can fit one star cluster in a Bayesian way and return MCMC samples for stellar parameters including IFMR parameters.

In statistics, Bayesian hierarchical models (See e.g., Gelman, 2006a; Gelman and Hill, 2007; Cressie et al., 2009; Licquia and Newman, 2015) pool multiple objects (star clusters in this case) into a comprehensive model, so that they can borrow strength from each other and eventually obtain improved estimates on object-level parameters (IFMRs of clusters). We attempt to use the power of Bayesian hierarchical models to obtain improved estimates of IFMR parameters.

However, the stellar evolution models tends to be complicated and computer-based, so embedding them into a Bayesian hierarchical model is a computational challenge. In this research we developed a two-stage algorithm that fits the Bayesian hierarchical model of multiple star clusters when the MCMC samples of all clusters are available from the black-box code. We have conducted simulation studies that show the advantages of our techniques. Also we have analysed five clusters via the Bayesian hierarchical model and our method can correct the unrealistic estimate of IFMR parameters from the black-box code for the cluster NGC 188. The unrealistic estimate is caused by narrow range of initial masses of WDs in NGC 188. So one particular situation where hierarchical models are helpful is when one or more objects have insufficient data, then the commonly used case-by-case analysis will produce result with large uncertainty.

In statistics, we have achieved our target, but this research still has limitations based on the astrophysical inputs. We did not know whether the two-stage algorithm would work with 2-dimensional object-level parameters. This example to IFMRs

is meant to a proof of concept for the more difficult computational problem. All astronomical results in this chapter are not definitive and they vary depending upon the models inside the black-box code and other assumptions. First, we assumed that IFMR parameters from different clusters follow a bivariate normal distribution. However, this assumption might be too idealised. We stress that Bayesian hierarchical models always require a population distribution on all objects and the shape of the distribution of IFMR parameters across all clusters is not available. We therefore use the bivariate normal distribution as a starting point. If a case can be made for a different distribution, the statistical algorithm developed in this chapter will work as long as the case-by-case results are valid. Many studies have shown that the metallicity and age have an effect to IFMR parameters of a cluster (See, e.g., Kalirai et al., 2005; Catalan et al., 2008; Meng et al., 2008; Zhao et al., 2012). The Bayesian hierarchical model in Eq. (5.7) can be readily adapted to investigate the effect of these factors. Due to the limited number of star clusters we have, however, we do not take the effect of age or metallicity into account.

Second, even though our research is based on the BASE-9 package, our statistical techniques in this chapter can be employed with other black-box packages as long as they produce MCMC samples from the case-by-case fits. Different underlying stellar evolution models will affect the results of case-by-case fits, hence they will impact the performance of hierarchical fits. However, this is irrelevant to the performance of our computational algorithm.

Third, the data analysis in this chapter is preliminary in the scientific sense that it is just a test to show that our statistical algorithm works in this situation when the case-by-case fits from BASE-9 are good. In the future, once sufficient data are available, we plan to use a flexible distribution for the population of IFMR parameters and perform a more wide-ranging analysis.

6 Conclusion & Future Work

6.1 Conclusion

Bayesian hierarchical models combine multiple objects (i.e., units of data) and produce partially pooled (shrinkage) estimates that have better statistical properties such as smaller mean squared errors (MSE) and better predictive accuracy (Gelman, 2006a; Morris and Lysy, 2012; Gelman et al., 2014; Katahira, 2016). In practice hierarchical models can be difficult to implement in complex situations that require sophisticated computer code to fit even a single object. The primary contribution of my thesis is to develop efficient algorithms for fitting hierarchical models in this situation, treating the computer code for fitting a single object as a black box that we do not open and do not need to alter. To give context to my methods, I develop them using several examples from astronomy. These examples are not meant to be definitive in terms of their astronomical results, but are meant to be testbeds for proofs of concept of my new algorithms.

I employed my methods to three astronomy problems: the published distance moduli to the Large Magellanic Cloud (LMC), the ages of a group of Galactic halo WDs candidates, and the initial-final mass relations (IFMRs) of a few star clusters. I have done simulation studies under these three settings, and I find that the root mean squared error (RMSE) of hierarchical estimates for object-level parameters re-

duce by 54.3% for the LMC project, 75.0% for the WD project, and 82.0% for the IFMR project, compared with the estimates from case-by-case analyses. My two-stage sampler algorithm uses case-by-case samples from the black-box and it only takes an additional 10 minutes to produce the shrinkage estimates under the hierarchical models, only increasing the computing time by 3.3% and 0.5% for the WD and IFMR projects, respectively. Because the LMC project uses a Gaussian-Gaussian hierarchical model that is straightforward, its computing time is not important. The interval estimates produced by hierarchical models perform also well, with their actual coverage probabilities close to the nominal level.

In summary, I have developed efficient statistical algorithms that fit Bayesian hierarchical models by using case-by-case fits from a black-box code and produce estimates of good properties, costing a manageable amount of time. The three astrophysical projects are preliminary and tentative, but are not meant to give definitive answers for astronomical problems. I include the LMC project to show how to combine case-by-case fits into a hierarchical model in a simple but realistic setting. The WD hierarchical model is a prototype analysis for Gaia, which will measure exquisite distances to tens of thousands of WDs. The IFMR project is used to illustrate how to combine multiple clusters into a hierarchical model and improve estimates of IFMR parameters when some of them has insufficient data to constrain the IFMR parameters.

6.2 Future Work

For the ages of Galactic halo WDs and IFMRs of star clusters projects, we can pool more objects in the hierarchical model and obtain better estimates for more objects. In the near future, we expect not only better calibrated photometry for many more WDs, but also highly informative priors on distance and population membership from

the Gaia satellite’s exquisite astrometry. We look forward to pooling many more WDs in our hierarchical model in order to derive an accurate and precise Galactic halo age distribution. Similarly, after having photometry for more star clusters, we can pool them in the Bayesian hierarchical model and use our algorithm to fit the hierarchical model, leading to more reliable estimates of IFMR parameters.

There are also a few statistical ideas to explore in the future. Firstly, one requirement of the two-stage algorithms may be relaxed. For now we assume that the case-by-case samples are stationary. However, this assumption can be relaxed to the condition that the case-by-case samples cover the case-by-case posterior distribution. Secondly, we can explore new approaches to give better proposals for the hierarchical posterior density and make the algorithms more efficient. Thirdly, we can develop some techniques which verifies that samples from the two-stage sampler is consistent with the posterior distribution. For now, the high-dimensional integrals hinder the evaluation of marginal posterior distributions.

References

- Alves, D. R. (2004). A review of the distance and structure of the large magellanic cloud. *New Astronomy Reviews*, 48(9):659–665.
- Bai, Y., Craiu, R. V., and Narzo, A. F. D. (2011). Divide and conquer: A mixture-based approach to regional adaptation for mcmc. *Journal of Computational and Graphical Statistics*, 20(1):63–79.
- Bailer-Jones, C. A. (2011). Bayesian time series analysis of terrestrial impact cratering. *Monthly Notices of the Royal Astronomical Society*, 416(2):1163–1180.
- Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311.
- Barnes, S. A., Weingrill, J., Fritzewski, D., Strassmeier, K. G., and Platais, I. (2016). Rotation periods for cool stars in the 4 gyr old open cluster m67, the solar–stellar connection, and the applicability of gyrochronology to at least solar age. *The Astrophysical Journal*, 823(1):16–31.
- Bellini, A., Bedin, L., Pichardo, B., Moreno, E., Allen, C., Piotto, G., and Anderson, J. (2010a). Absolute proper motion of the galactic open cluster m 67. *Astronomy and Astrophysics*, 513(1):51–56.

- Bellini, A., Bedin, L., Piotto, G., Salaris, M., Anderson, J., Brocato, E., Ragazzoni, R., Ortolani, S., Bonanos, A., and Platais, I. (2010b). The end of the white dwarf cooling sequence in m67. *Astronomy and Astrophysics*, 513(2):50–61.
- Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.
- Bergeron, P., Wesemael, F., and Beauchamp, A. (1995). Photometric calibration of hydrogen- and helium-rich white dwarf models. *Publications of the Astronomical Society of the Pacific*, 107(717):1047–1054.
- Bernardinelli, L. and Montomoli, C. (1992). Empirical bayes versus fully bayesian analysis of geographical variation in disease risk. *Statistics in medicine*, 11(8):983–1007.
- Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147.
- Bhardwaj, A., Kanbur, S. M., Macri, L. M., Singh, H. P., Ngeow, C.-C., Wagner-Kaiser, R., and Sarajedini, A. (2016). Large magellanic cloud near-infrared synoptic survey. ii. the wesenheit relations and their application to the distance scale. *The Astronomical Journal*, 151(4):88.
- Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.
- Browne, W. and Goldstein, H. (2002). An introduction to bayesian multilevel hierarchical modelling using mlwin. Technical report, Institute of Education, London.

- Browne, W. J., Draper, D., et al. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis*, 1(3):473–514.
- Carlin, B. P. and Louis, T. A. (2000). Empirical bayes: past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289.
- Casella, G. (2001). Empirical bayes gibbs sampling. *Biostatistics*, 2(4):485–500.
- Catalan, S., Isern, J., García-Berro, E., and Ribas, I. (2008). The initial–final mass relationship of white dwarfs revisited: effect on the luminosity function and mass distribution. *Monthly Notices of the Royal Astronomical Society*, 387(4):1693–1706.
- Chan, K. and Ledolter, J. (1995). Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252.
- Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics*, 40(2):136–157.
- Clementini, G., Gratton, R., Bragaglia, A., Carretta, E., Di Fabrizio, L., and Maio, M. (2003). Distance to the large magellanic cloud: The rr lyrae stars. *The Astronomical Journal*, 125(3):1309.
- Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.
- Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

- Craiu, R. V., Rosenthal, J., and Yang, C. (2009). Learn from thy neighbor: Parallel-chain and regional adaptive mcmc. *Journal of the American Statistical Association*, 104(488):1454–1466.
- Crandall, S. and Ratra, B. (2015). Non-gaussian error distributions of lmc distance moduli measurements. *The Astrophysical Journal*, 815(2):87.
- Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3):553–570.
- Cui, W. and George, E. I. (2008). Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900.
- da Silva, L., Girardi, L., Pasquini, L., Setiawan, J., Von Der Lühe, O., De Medeiros, J., Hatzes, A., Döllinger, M., and Weiss, A. (2006). Basic physical parameters of a selected sample of evolved stars. *Astronomy and Astrophysics*, 458(2):609–623.
- Dame, K., Gianninas, A., Kilic, M., Munn, J. A., Brown, W. R., Williams, K. A., von Hippel, T., and Harris, H. C. (2016). New halo white dwarf candidates in the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 463(3):2453–2464.
- Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27(3):567–578.
- Daniels, M. J. and Kass, R. E. (1999). Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263.

- Daniels, M. J. and Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184.
- Datta, G. S. (1996). On priors providing frequentist validity of bayesian inference for multiple parametric functions. *Biometrika*, 83(2):287–298.
- Datta, G. S. and Ghosh, J. K. (1995). On priors providing frequentist validity for bayesian inference. *Biometrika*, 82(1):37–45.
- De Grijs, R., Wicker, J. E., and Bono, G. (2014). Clustering of local group distances: publication bias or correlated measurements? i. the large magellanic cloud. *The Astronomical Journal*, 147(5):122–135.
- DeGennaro, S. (2009). *White Dwarfs and the Ages of Stellar Populations*. PhD thesis, The University of Texas at Austin.
- DeGennaro, S., von Hippel, T., Jefferys, W. H., Stein, N., van Dyk, D., and Jeffery, E. (2009). Inverting color-magnitude diagrams to access precise star cluster parameters: A new white dwarf age for the hyades. *The Astrophysical Journal*, 696(1):12–23.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1):1–38.
- Di Benedetto, G. (1997). Improved calibration of cosmic distance scale by cepheid pulsation parallaxes. *The Astrophysical Journal*, 486(1):60–74.
- Dotter, A., Chaboyer, B., Jevremović, D., Kostov, V., Baron, E., and Ferguson, J. W. (2008). The dartmouth stellar evolution database. *Astrophysical Journal Supplement Series*, 178(1):89–101.

- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1):1–26.
- Efron, B. (1996). Empirical bayes methods for combining likelihoods. *Journal of the American Statistical Association*, 91(434):538–550.
- Efron, B. and Morris, C. (1972). Limiting the risk of bayes and empirical bayes estimators: Part ii: The empirical bayes case. *Journal of the American Statistical Association*, 67(337):130–139.
- Efron, B. and Morris, C. (1973). Stein’s estimation rule and its competitors: an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.
- Efron, B. and Morris, C. (1977). *Stein’s Paradox in Statistics*. Scientific American offprints. W.H. Freeman.
- Feast, M. (2005). Local distance indicators. In *Symposium-International Astronomical Union*, volume 201, pages 181–189. Cambridge University Press.
- Feast, M. W. and Catchpole, R. M. (1997). The cepheid period-luminosity zero-point from hipparcos trigonometrical parallaxes. *Monthly Notices of the Royal Astronomical Society*, 286(1):L1–L5.
- Fitzpatrick, E. L., Ribas, I., Guinan, E. F., DeWarf, L. E., Maloney, F. P., and Massa, D. (2002). Fundamental properties and distances of the large magellanic cloud from eclipsing binaries. ii. hv 982. *The Astrophysical Journal*, 564(1):260–273.
- Fort, G., Moulines, E., et al. (2003). Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259.

- Gelman, A. (2006a). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48:241–251.
- Gelman, A. (2006b). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.
- Gelman, A. (2008). Rejoinder. *Bayesian Analysis*, 3(3):467–477.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.
- Gelman, A. et al. (2005). Analysis of variance why it is more important than ever. *The annals of statistics*, 33(1):1–53.
- Gelman, A. and Hill, J. (2007). Data analysis using regression and multi-level/hierarchical models. *Cambridge: Cambridge University Press*.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.
- Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of markov chain monte carlo*, pages 163–174.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6(6):721–741.
- Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood. *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156–163.

- Gianninas, A., Curd, B., Thorstensen, J. R., Kilic, M., Bergeron, P., Andrews, J. J., Canton, P., and Agüeros, M. (2015). Ultracool white dwarfs and the age of the galactic disc. *Monthly Notices of the Royal Astronomical Society*, 449(4):3966–3980.
- Gieren, W. P., Fouqué, P., and Gómez, M. (1998). Cepheid period-radius and period-luminosity relations and the distance to the large magellanic cloud. *The Astrophysical Journal*, 496(1):17–30.
- Girardi, L., Bressan, A., Bertelli, G., and Chiosi, C. (2000). Evolutionary tracks and isochrones for low-and intermediate-mass stars: From 0.15 to 7 $m_{\odot}$, and from $z = 0.0004$ to 0.03. *Astronomy and Astrophysics Supplement Series*, 141(3):371–383.
- Groenewegen, M. A. T. (2000). Lmc and smc cepheids: combining ogle with denis 2mass infrared data. *Astronomy and Astrophysics*, 363:901–916.
- Gustafson, P., Hossain, S., and MacNab, Y. C. (2006). Conservative prior distributions for variance parameters in hierarchical models. *Canadian Journal of Statistics*, 34(3):377–390.
- Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.
- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.
- Hernandez, X. and Valls-Gabaud, D. (2008). A robust statistical estimation of the basic parameters of single stellar populations - I. Method. *Monthly Notices of the Royal Astronomical Society*, 383:1603–1618.

- Hernandez, X., Valls-Gabaud, D., and Gilmore, G. (1999). Deriving star formation histories: inverting hertzsprung-russell diagrams through a variational calculus maximum likelihood method. *Monthly Notices of the Royal Astronomical Society*, 304(4):705–719.
- Hills, S., von Hippel, T., Courteau, S., and Geller, A. M. (2015). Bayesian investigation of isochrone consistency using the old open cluster ngc 188. *The Astronomical Journal*, 149(3):94–101.
- Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2):439–452.
- Inno, L., Bono, G., Matsunaga, N., Fiorentino, G., Marconi, M., Lemasle, B., Da Silva, R., Soszyński, I., Udalski, A., and Romaniello, M. (2016). The panchromatic view of the magellanic clouds from classical cepheids. i. distance, reddening, and geometry of the large magellanic cloud disk. *The Astrophysical Journal*, 832(2):176–195.
- Inno, L., Matsunaga, N., Bono, G., Caputo, F., Buonanno, R., Genovali, K., Laney, C., Marconi, M., Piersimoni, A., and Primas, F. (2013). On the distance of the magellanic clouds using cepheid nir and optical-nir period-wesenheit relations. *The Astrophysical Journal*, 764(1):84–96.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.
- Jao, W.-C., Mason, B. D., Hartkopf, W. I., Henry, T. J., and Ramos, S. N. (2009). Cool subdwarf investigations. ii. multiplicity. *The Astronomical Journal*, 137(4):3800–3808.

- Jaynes, E. T. (1983). Concentration of distributions at entropy maxima. *ET Jaynes: Papers on probability, statistics, and statistical physics*, pages 315–336.
- Jeffery, E. J., von Hippel, T., DeGennaro, S., van Dyk, D. A., Stein, N., and Jefferys, W. H. (2011). The white dwarf age of ngc 2477. *The Astrophysical Journal*, 730(1):35–44.
- Jeffreys, H. (1961). *Theory of probability (3rd ed.)*. Oxford University Press.
- Kalirai, J. S., Hansen, B. M., Kelson, D. D., Reitzel, D. B., Rich, R. M., and Richer, H. B. (2008). The initial-final mass relation: direct constraints at the low-mass end. *The Astrophysical Journal*, 676(1):594–609.
- Kalirai, J. S., Richer, H. B., Reitzel, D., Hansen, B. M. S., Rich, R. M., Fahlman, G. G., Gibson, B. K., and von Hippel, T. (2005). The initial-final mass relationship: Spectroscopy of white dwarfs in ngc 2099 (m37). *The Astrophysical Journal Letter*, 618:L123–L127.
- Kass, R. E. and Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):717–726.
- Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370.
- Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, 73:37–58.
- Kilic, M., Brown, W. R., Prieto, C. A., Kenyon, S., Heinke, C. O., Agüeros, M. A., and Kleinman, S. (2012). The elm survey. iv. 24 white dwarf merger systems. *The Astrophysical Journal*, 751(2):141–153.

- Kilic, M., Munn, J. A., Williams, K. A., Kowalski, P., von Hippel, T., Harris, H. C., Jeffery, E. J., DeGennaro, S., Brown, W. R., and McLeod, B. (2010). Visitors from the halo: 11 gyr old white dwarfs in the solar neighborhood. *The Astrophysical Journal*, 715(1):L21–L25.
- Kilic, M., Thorstensen, J. R., Kowalski, P., and Andrews, J. (2012). 11–12 gyr old white dwarfs 30 pc away. *Monthly Notices of the Royal Astronomical Society*, 423(1):L132–L136.
- Laney, C. and Stobie, R. (1994). Cepheid period–luminosity relations in k, h, j and v. *Monthly Notices of the Royal Astronomical Society*, 266(2):441–454.
- Leman, S. C., Chen, Y., and Lavine, M. (2009). The multiset sampler. *Journal of the American Statistical Association*, 104(487):1029–1041.
- Levine, R. A. and Casella, G. (2003). Implementing matching priors for frequentist inference. *Biometrika*, 90(1):127–137.
- Licquia, T. C. and Newman, J. A. (2015). Improved estimates of the milky way’s stellar mass and star formation rate from hierarchical bayesian meta-analysis. *The Astrophysical Journal*, 806(1):96–117.
- Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.
- Liu, J. S., Liang, F., and Wong, W. H. (2000). The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134.
- Mandel, K. S., Wood-Vasey, W. M., Friedman, A. S., and Kirshner, R. P. (2009).

- Type ia supernova light-curve inference: Hierarchical bayesian analysis in the near-infrared. *The Astrophysical Journal*, 704(1):629–651.
- March, M., Karpenka, N. V., Feroz, F., and Hobson, M. (2014). Comparison of cosmological parameter inference methods applied to supernovae light curves fitted with salt-ii. *Monthly Notices of the Royal Astronomical Society*, 437(4):3298–3311.
- Meibom, S., Grundahl, F., Clausen, J. V., Mathieu, R. D., Frandsen, S., Pigulski, A., Narwid, A., Steslicki, M., and Lefever, K. (2009). Age and distance for the old open cluster ngc 188 from the eclipsing binary member v 12. *The Astronomical Journal*, 137(6):5086.
- Meng, X., Chen, X., and Han, Z. (2008). Initial-final mass relationship for stars of different metallicities. *Astronomy and Astrophysics*, 487(2):625–635.
- Meng, X.-L. and Zaslavsky, A. M. (2002). Single observation unbiased priors. *Annals of Statistics*, 30(5):1345–1375.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.
- Miller, G. E. and Scalo, J. M. (1979). The initial mass function and stellar birthrate in the solar neighborhood. *The Astrophysical Journal Supplement Series*, 41:513–547.
- Montgomery, M., Klumpe, E., Winget, D., and Wood, M. (1999). Evolutionary calculations of phase separation in crystallizing white dwarf stars. *The Astrophysical Journal*, 525(1):482–491.
- Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

- Morris, C. N. and Lysy, M. (2012). Shrinkage estimation in multilevel normal models. *Statistical Science*, 27(1):115–134.
- Muraveva, T., Clementini, G., and Palmer, M. (2014). Improving the cosmic distance ladder. distance and structure of the large magellanic cloud. *European Astronomical Society Publications Series*, 68:275–278.
- Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380.
- Nylander, J. A., Wilgenbusch, J. C., Warren, D. L., and Swofford, D. L. (2007). Awty (are we there yet?): a system for graphical exploration of mcmc convergence in bayesian phylogenetics. *Bioinformatics*, 24(4):581–583.
- O’Malley, E. M., Von Hippel, T., and Van Dyk, D. A. (2013). A bayesian approach to deriving ages of individual field white dwarfs. *The Astrophysical Journal*, 775(1):1–11.
- Panagia, N. (1998). New distance determination to the lmc. *Memorie della Societa Astronomica Italiana*, 69:225–235.
- Park, T., van Dyk, D. A., and Siemiginowska, A. (2008). Searching for narrow emission lines in x-ray spectra: Computation and methods. *The Astrophysical Journal*, 688(2):807–825.
- Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.
- Perryman, M. A., Lindegren, L., Kovalevsky, J., Hoeg, E., Bastian, U., Bernacca,

- P., Cr    , M., Donati, F., Grenon, M., Grewing, M., et al. (1997). The hipparcos catalogue. *Astronomy and Astrophysics*, 323:L49–L52.
- Perryman, M. A. C., Brown, A. G. A., Lebreton, Y., Gomez, A., Turon, C., Cayrel de Strobel, G., Mermilliod, J. C., Robichon, N., Kovalevsky, J., and Crifo, F. (1998). The hyades: distance, structure, dynamics, and age. *Astronomy and Astrophysics*, 331:81–120.
- Persaud, B., Lan, B., Lyon, C., and Bhim, R. (2010). Comparison of empirical bayes and full bayes approaches for before–after road safety evaluations. *Accident Analysis & Prevention*, 42(1):38–43.
- Pietrzy  ski, G. and Gieren, W. (2002). The araucaria project: Deep near-infrared survey of nearby galaxies. i. the distance to the large magellanic cloud from k-band photometry of red clump stars based on observations obtained with the new technology telescope at the european southern observatory as part of project 69. d-0352. *The Astronomical Journal*, 124(5):2633–2638.
- Pietrzy  ski, G., Graczyk, D., Gieren, W., Thompson, I., Pilecki, B., Udalski, A., Soszy  ski, I., Koz  owski, S., Konorski, P., Suchomska, K., et al. (2013). An eclipsing-binary distance to the large magellanic cloud accurate to two per cent. *Nature*, 495(7439):76–79.
- Podariu, S., Souradeep, T., Gott III, J. R., Ratra, B., and Vogeley, M. S. (2001). Binned cosmic microwave background anisotropy power spectra: Peak location. *The Astrophysical Journal*, 559(1):9–22.
- Pont, F. and Eyer, L. (2004). Isochrone ages for field dwarfs: method and application to the age-metallicity relation. *Monthly Notices of the Royal Astronomical Society*, 351:487–504.

-
- Potekhin, A. and Chabrier, G. (2013). Equation of state for magnetized coulomb plasmas. *Astronomy and Astrophysics*, 550:43–58.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Renedo, I., Althaus, L. G., Bertolami, M. M., Romero, A. D., Córscico, A. H., Rohrmann, R. D., and García-Berro, E. (2010). New cooling sequences for old white dwarfs. *The Astrophysical Journal*, 717(1):183–195.
- Robin, A. C., Reylé, C., Derriere, S., and Picaud, S. (2003). A synthetic view on structure and evolution of the milky way. *Astronomy and Astrophysics*, 409(2):523–540.
- Romaniello, M., Salaris, M., Cassisi, S., and Panagia, N. (2000). Hubble space telescope observations of the large magellanic cloud field around sn 1987a: Distance determination with red clump and tip of the red giant branch stars. *The Astrophysical Journal*, 530(2):738–743.
- Salaris, M., Althaus, L. G., and García-Berro, E. (2013). Comparison of theoretical white dwarf cooling timescales. *Astronomy and Astrophysics*, 555:96–103.
- Salaris, M., Serenelli, A., Weiss, A., and Bertolami, M. M. (2009). Semi-empirical white dwarf initial-final mass relationships: a thorough analysis of systematic uncertainties due to stellar evolution models. *The Astrophysical Journal*, 692(2):1013–1032.
- Sarajedini, A., Grocholski, A. J., Levine, J., and Lada, E. (2002). K-band red clump distances to the large magellanic cloud clusters hodge 4 and ngc 1651. *The Astrophysical Journal*, 124(5):2625–2632.

- Schuurman, N., Grasman, R., and Hamaker, E. (2016). A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate behavioral research*, 51(2-3):185–206.
- Scott, J. G., Berger, J. O., et al. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.
- Shariff, H., Dhawan, S., Jiao, X., Leibundgut, B., Trotta, R., and van Dyk, D. (2016). Standardizing type ia supernovae optical brightness using near-infrared rebrightening time. *Monthly Notices of the Royal Astronomical Society*, 463(4):4311–4316.
- Si, S., van Dyk, D. A., and von Hippel, T. (2017a). Sensitivity analysis of hierarchical models for the ages of galactic halo white dwarfs. In Tremblay, P.-E., Gaensicke, B., and Marsh, T., editors, *20th European White Dwarf Workshop*, volume 509 of *Astronomical Society of the Pacific Conference Series*, pages 69–72.
- Si, S., van Dyk, D. A., von Hippel, T., Robinson, E., Webster, A., and Stenning, D. (2017b). A Hierarchical Model for the Ages of Galactic Halo White Dwarfs. *Monthly Notices of the Royal Astronomical Society*, 468(4):4374–4388.
- Sollima, A., Carballo-Bello, J., Beccari, G., Ferraro, F., Pecci, F. F., and Lanzoni, B. (2010). The fraction of binary systems in the core of five galactic open clusters. *Monthly Notices of the Royal Astronomical Society*, 401(1):577–585.
- Stein, N. M., van Dyk, D. A., von Hippel, T., DeGennaro, S., Jeffery, E. J., and Jefferys, W. H. (2013). Combining computer models to account for mass loss in stellar evolution. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(1):34–52.

- Stenning, D., Wagner-Kaiser, R., Robinson, E., Van Dyk, D., Von Hippel, T., Sarajedini, A., and Stein, N. (2016). Bayesian analysis of two stellar populations in galactic globular clusters. i. statistical and computational methods. *The Astrophysical Journal*, 826(1):41–56.
- Sung, H. and Bessell, M. (1999). Ubv_i ccd photometry of m35 (ngc 2168). *Monthly Notices of the Royal Astronomical Society*, 306(2):361–370.
- Taylor, B. (2006). The benchmark cluster reddening project. ii. a reddening value for m67. *The Astronomical Journal*, 133(2):370–386.
- Tiao, G. C. and Box, G. E. P. (1973). Some comments on bayes estimators. *The American Statistician*, 27(1):12–14.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.
- van Dyk, D. A., Degennaro, S., Stein, N., Jefferys, W. H., and von Hippel, T. (2009). Statistical analysis of stellar evolution. *Annals of Applied Statistics*, 3:117–143.
- van Dyk, D. A. and Meng, X.-L. (2010). Cross-fertilizing strategies for better em mountain climbing and da field exploration: A graphical guide book. *Statistical Science*, 25(4):429–449.
- van Dyk, D. A. and Park, T. (2008). Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796.
- van Leeuwen, F., Feast, M., Whitelock, P., and Yudin, B. (1997). First results from hipparcos trigonometrical parallaxes of mira-type variables. *Monthly Notices of the Royal Astronomical Society*, 287(4):955–960.

- VandenBerg, D. A. and Stetson, P. (2004). On the old open clusters m67 and ngc 188: Convective core overshooting, color-temperature relations, distances, and ages. *Publications of the Astronomical Society of the Pacific*, 116(825):997–1011.
- von Hippel, T., Jefferys, W. H., Scott, J., Stein, N., Winget, D., DeGennaro, S., Dam, A., and Jeffery, E. (2006). Inverting color-magnitude diagrams to access precise star cluster parameters: A bayesian approach wiyn open cluster study xxiii. *The Astrophysical Journal*, 645(2):1436–1447.
- von Hippel, T., Robinson, E., Jeffery, E., Wagner-Kaiser, R., DeGennaro, S., Stein, N., Stenning, D., Jefferys, W. H., and van Dyk, D. (2014). Bayesian analysis for stellar evolution with nine parameters (base-9): User’s manual. *ArXiv e-prints*.
- von Hippel, T. and Sarajedini, A. (1998). Wiyn open cluster study. i. deep photometry of ngc 188. *The Astronomical Journal*, 116(4):1789–1800.
- Wagner-Kaiser, R., Mackey, D., Sarajedini, A., Chaboyer, B., Cohen, R. E., Yang, S.-C., Cummings, J. D., Geisler, D., and Grocholski, A. J. (2017). Exploring the nature and synchronicity of early cluster formation in the large magellanic cloud: Ii. relative ages and distances for six ancient globular clusters. *Monthly Notices of the Royal Astronomical Society*, 471(1):3347–3358.
- Wei, G. C. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man’s data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704.
- Weidemann, V. (2000). Revision of the initial-to-final mass relation. *Astronomy and Astrophysics*, 363:647–656.
- Williams, K. A., Bolte, M., and Koester, D. (2004). An empirical initial-final mass

- relation from hot, massive white dwarfs in ngc 2168 (m35). *The Astrophysical Journal Letters*, 615(1):L49–L52.
- Williams, K. A., Bolte, M., and Koester, D. (2009). Probing the lower mass limit for supernova progenitors and the high-mass end of the initial-final mass relation from white dwarfs in the open cluster m35 (ngc 2168). *The Astrophysical Journal*, 693(1):355–369.
- Williams, K. A., Howell, S. B., Liebert, J., Smith, P. S., Bellini, A., Rubin, K. H., and Bolte, M. (2013). Time-resolved spectroscopy of the polar eu cancri in the open cluster messier 67. *The Astronomical Journal*, 145(5):129–137.
- Winget, D. and Kepler, S. (2008). Pulsating white dwarf stars and precision asteroseismology. *Annual Review of Astronomy and Astrophysics*, 46:157–199.
- Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 11(1):95–103.
- Yi, S., Demarque, P., Kim, Y.-C., Lee, Y.-W., Ree, C. H., Lejeune, T., and Barnes, S. (2001). Toward better age estimates for stellar populations: The y2 isochrones for solar mixture. *Astrophysical Journal Supplement Series*, 136(2):417–437.
- Zhao, J. K., Oswalt, T. D., Willson, L. A., Wang, Q., and Zhao, G. (2012). The initial-final mass relation among white dwarfs in wide binaries. *The Astrophysical Journal*, 746(2):144–154.

A Case-by-case Fits of WDs

In this section, we present the case-by-case fits of 10 WDs used in the data analysis section in Chapter 4. Figs A.1 – A.10 display the two-dimensional scatterplots and histograms of stellar parameters for 10 WDs, respectively.

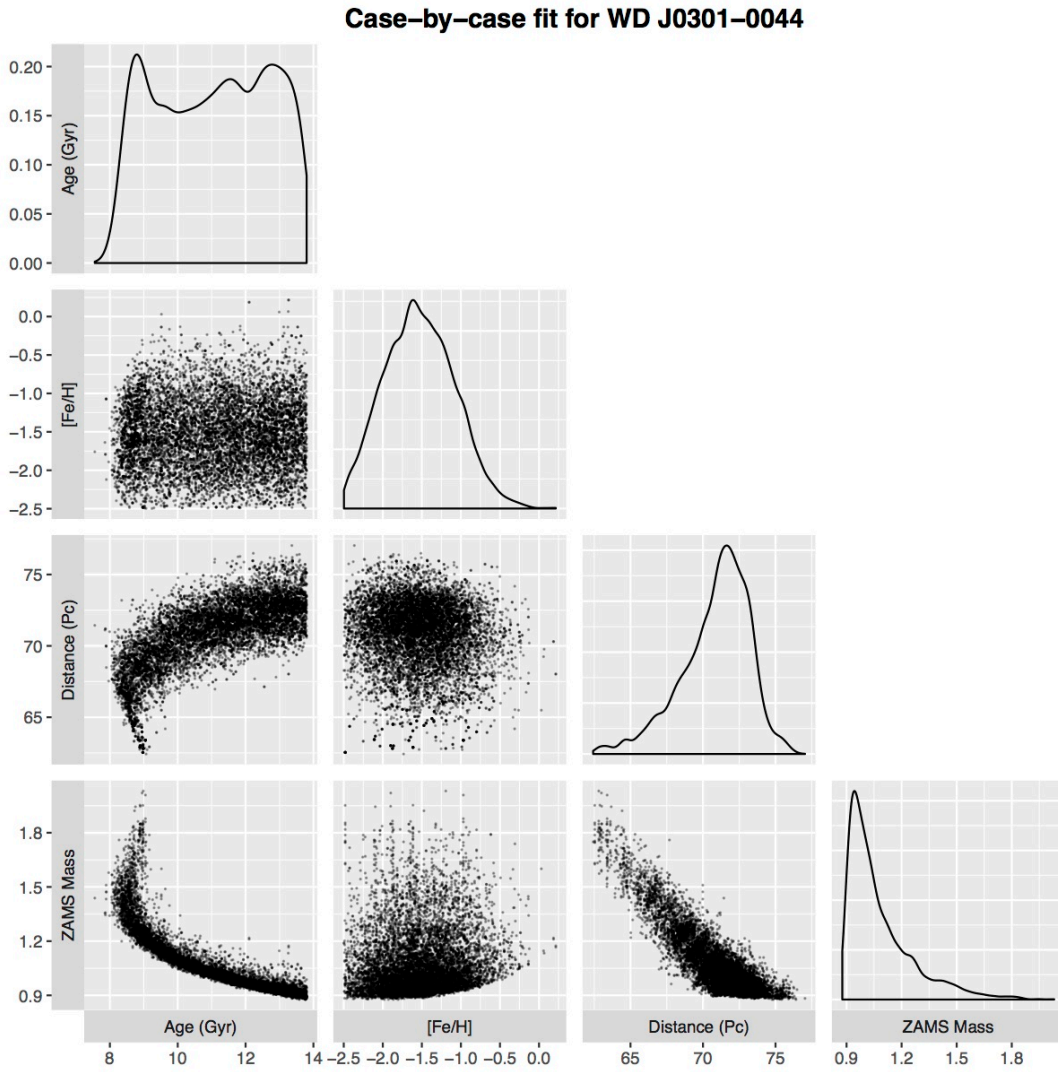


Figure A.1: The scatterplots of parameters of WD J0301-0044 from the case-by-case fit.

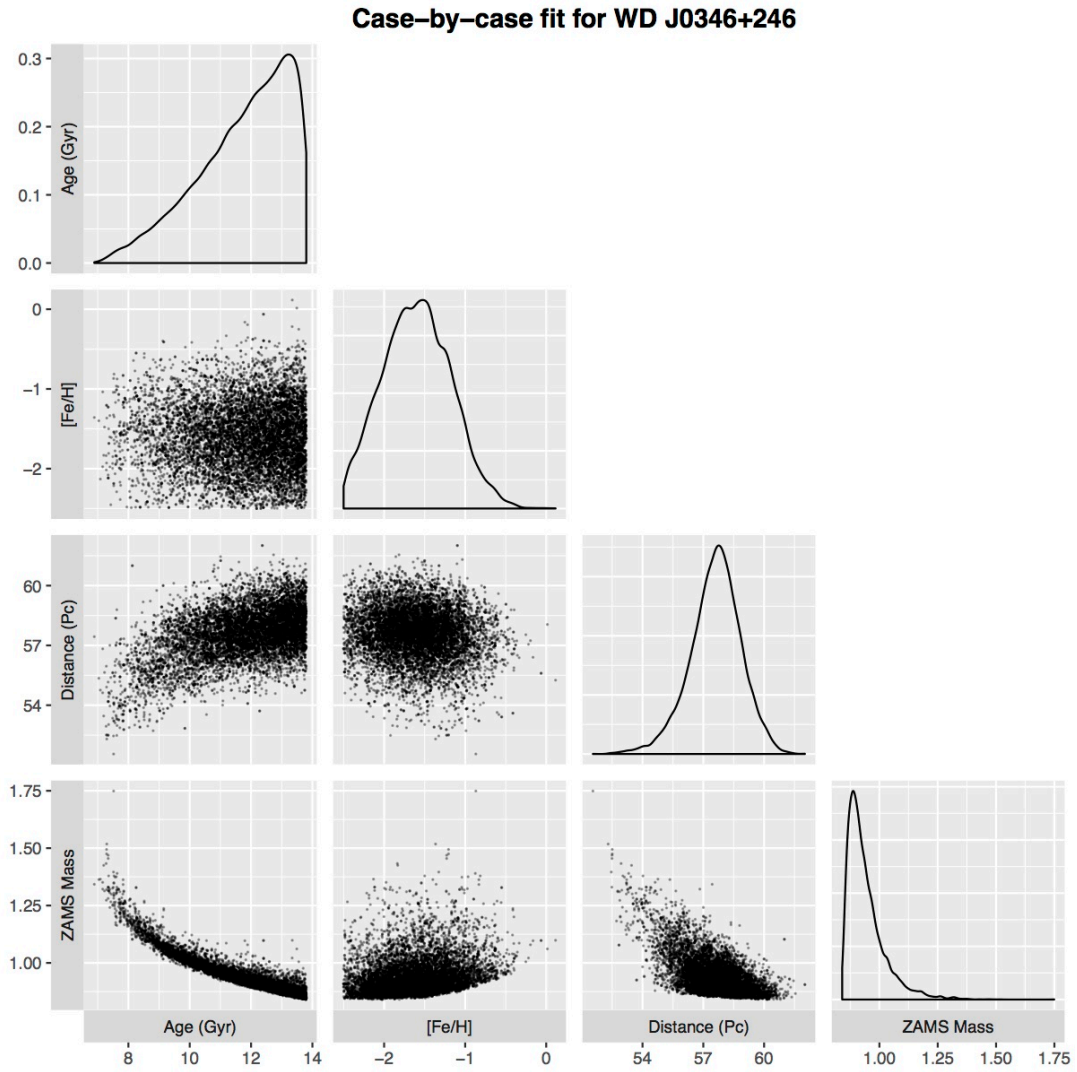


Figure A.2: The scatterplots of parameters of WD J0346+246 from the case-by-case fit.

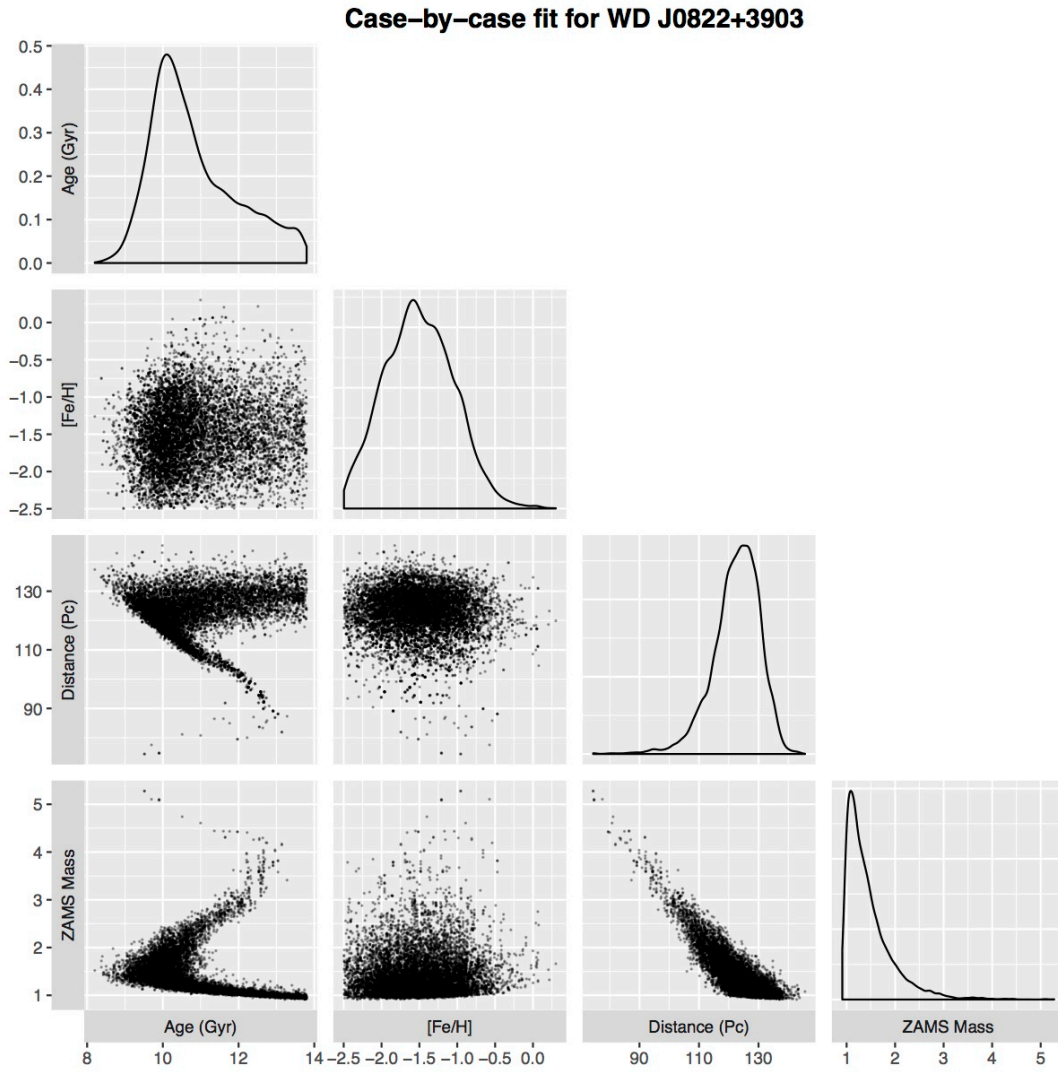


Figure A.3: The scatterplots of parameters of WD J0822+3903 from the case-by-case fit.

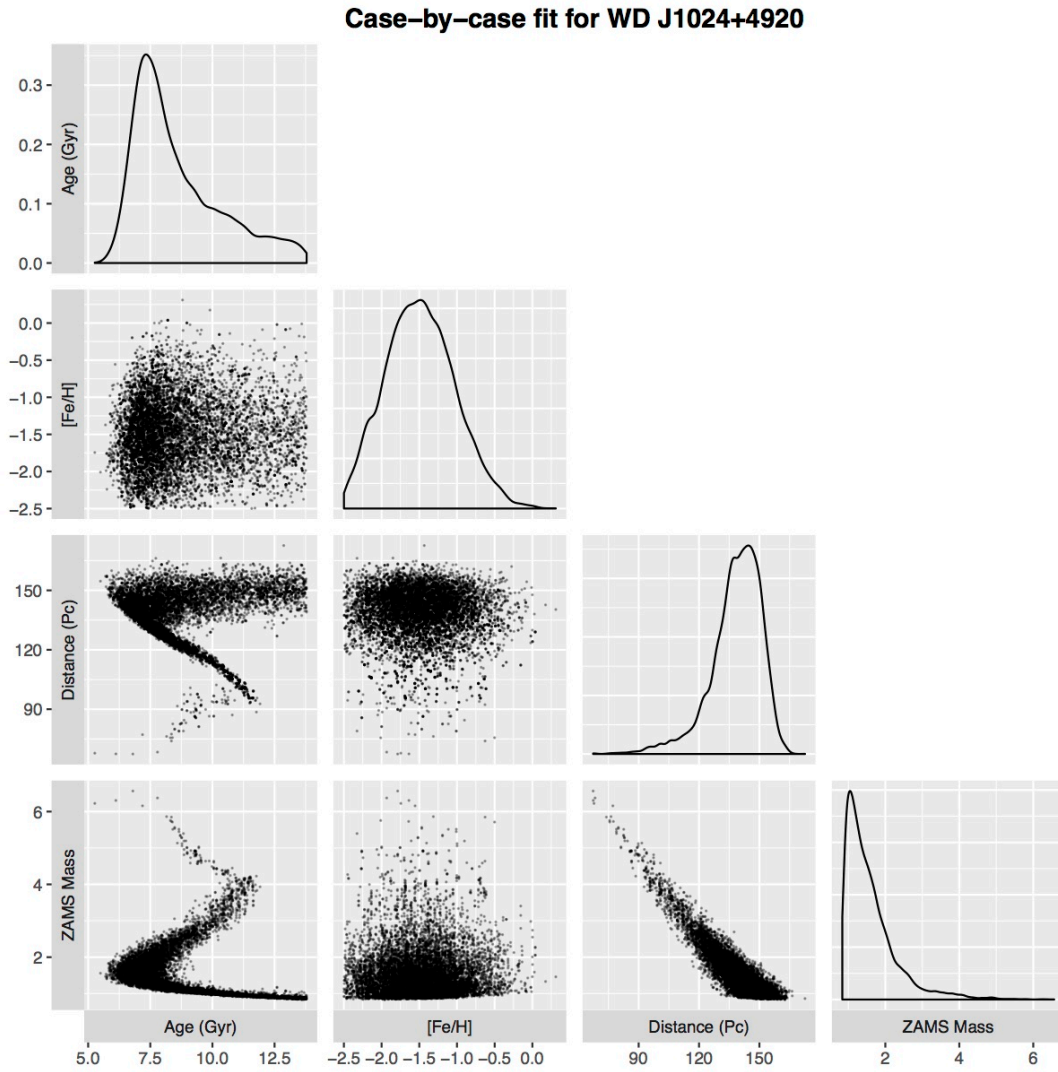


Figure A.4: The scatterplots of parameters of WD J1024+4920 from the case-by-case fit.

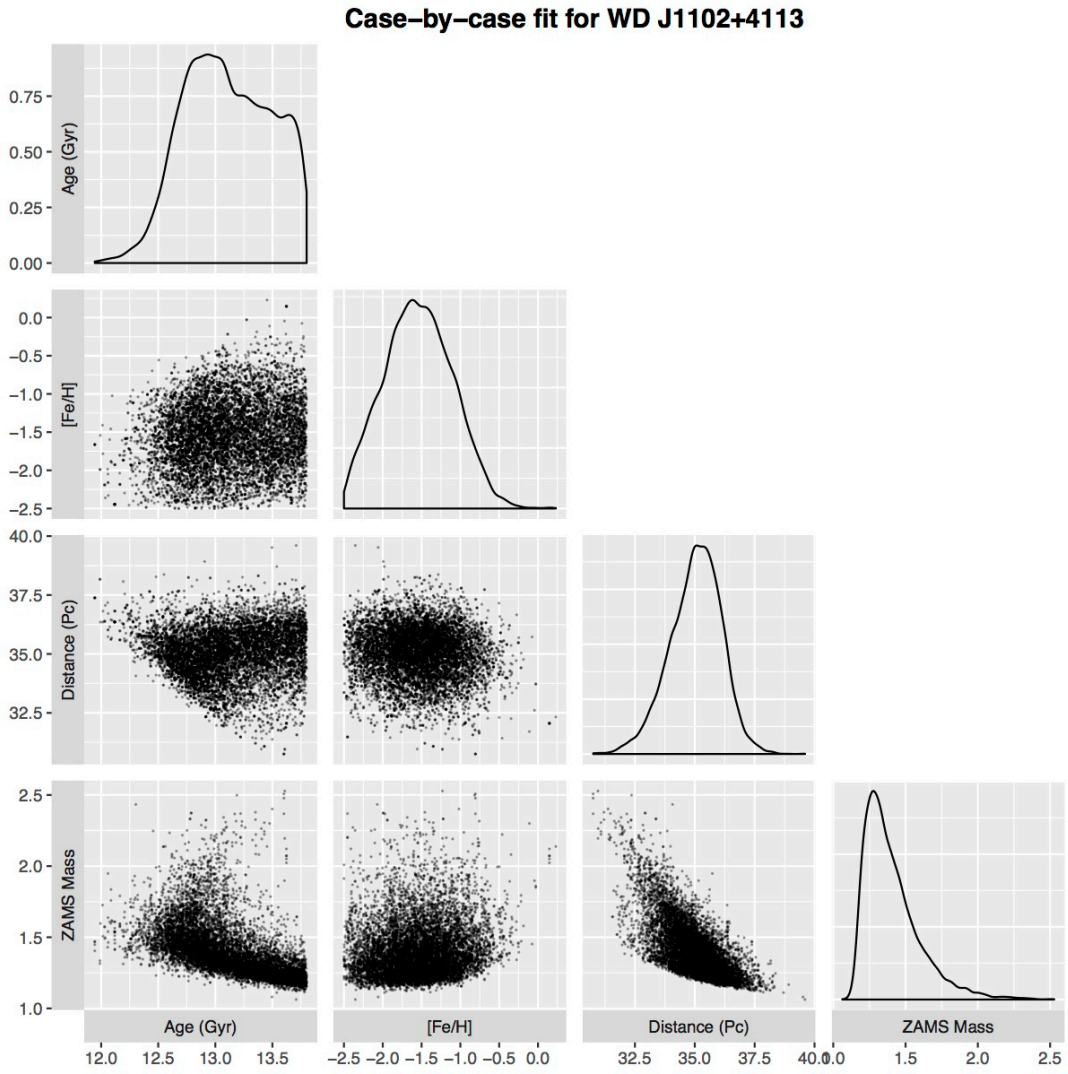


Figure A.5: The scatterplots of parameters of WD J1102+4113 from the case-by-case fit.

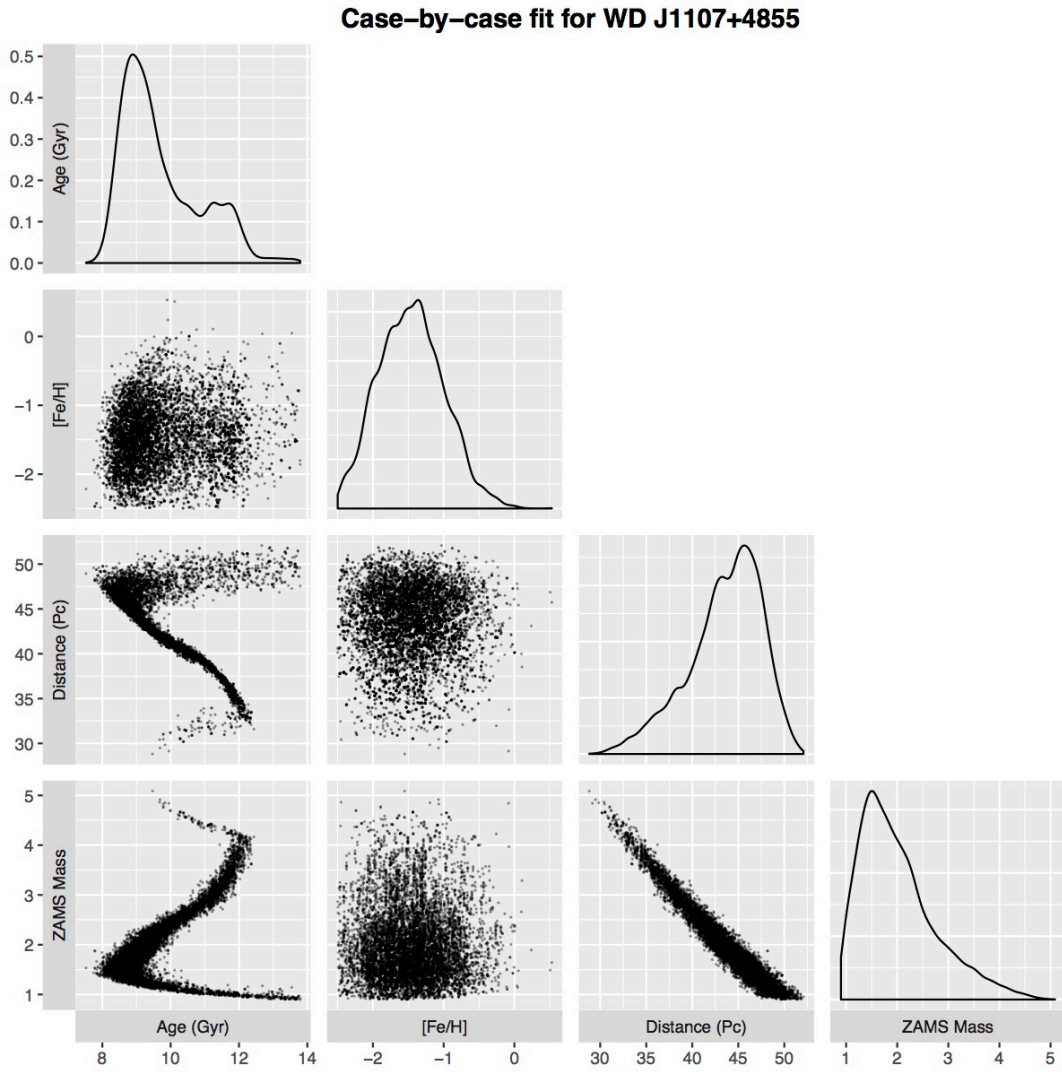


Figure A.6: The scatterplots of parameters of WD J1107+4855 from the case-by-case fit.

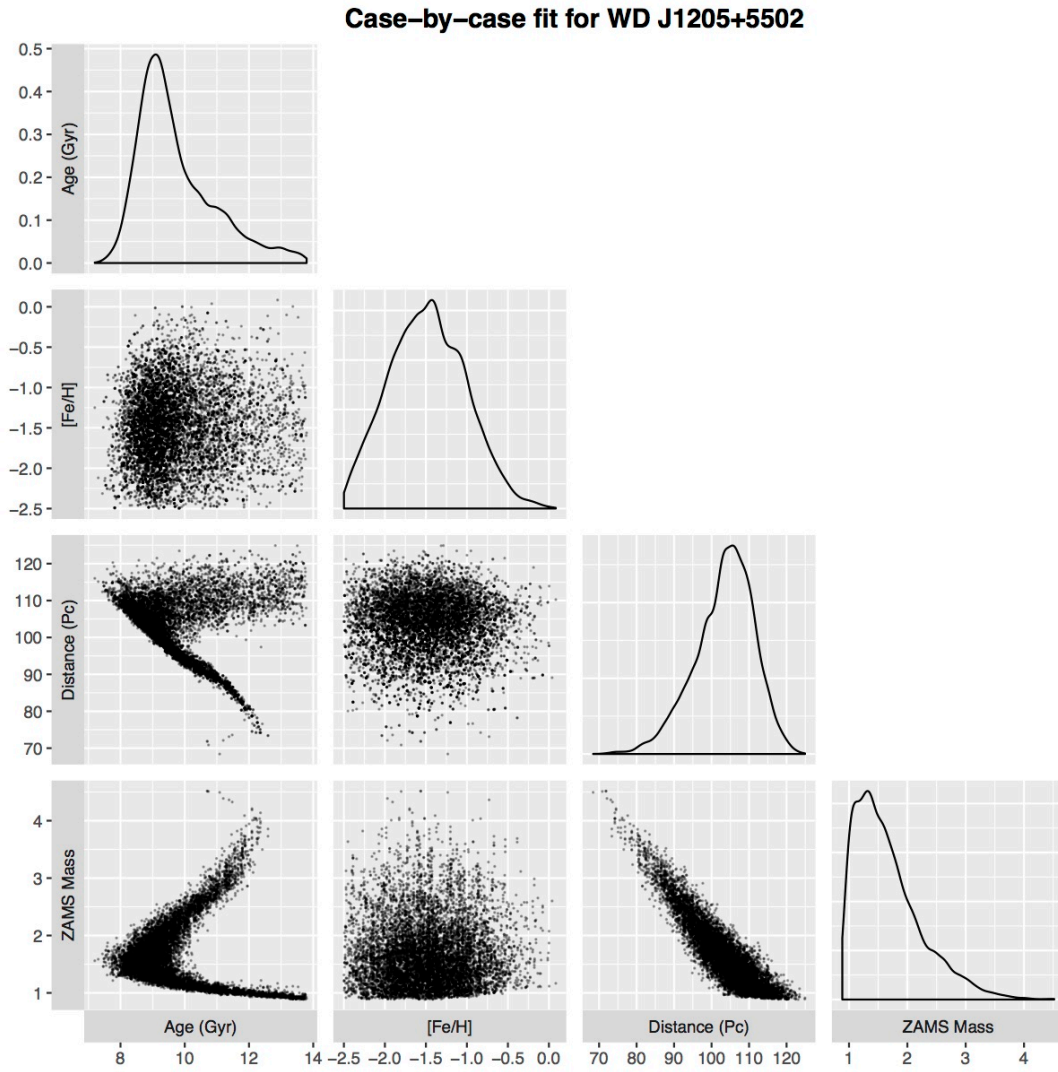


Figure A.7: The scatterplots of parameters of WD J1205+5502 from the case-by-case fit.

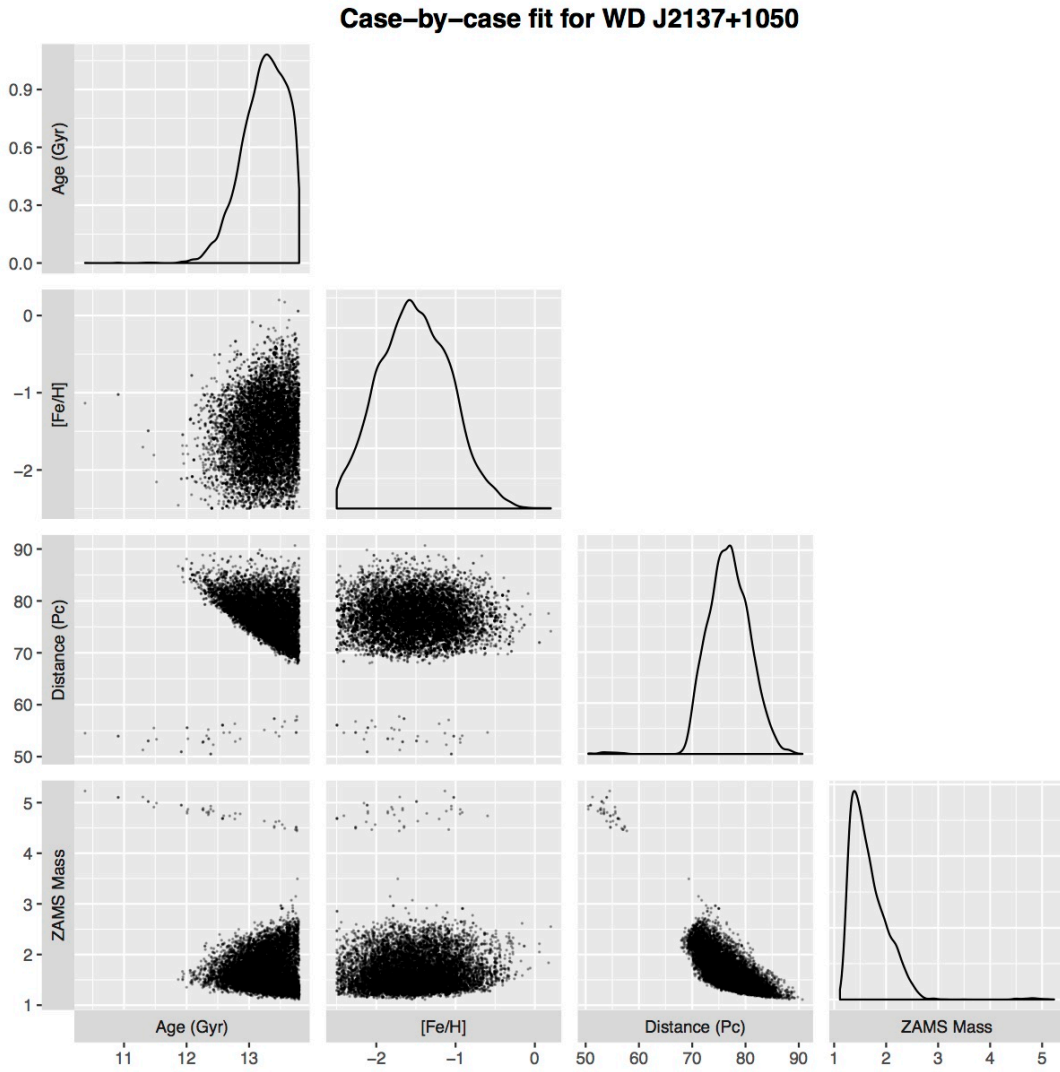


Figure A.8: The scatterplots of parameters of WD J2137-1050 from the case-by-case fit.

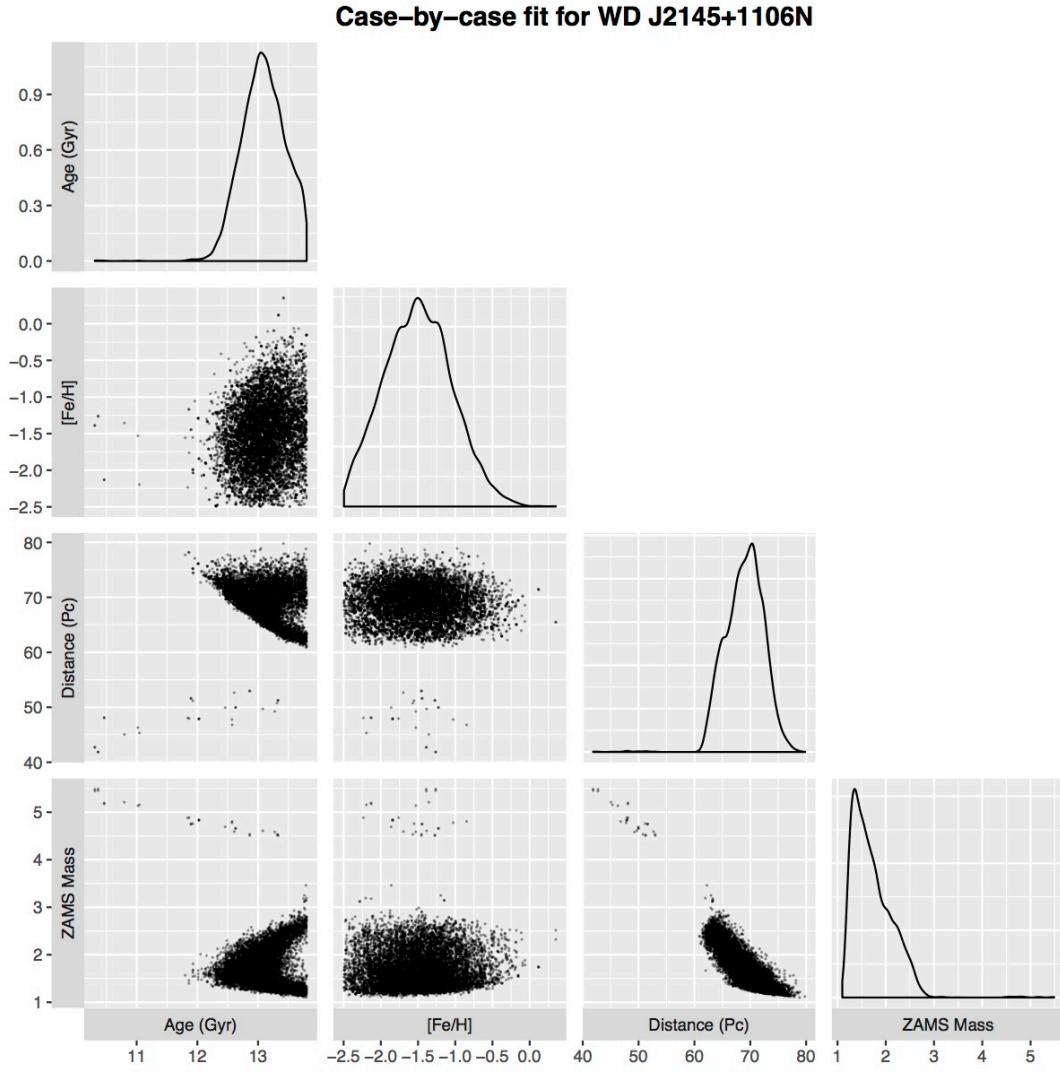


Figure A.9: The scatterplots of parameters of WD J2145+1106N from the case-by-case fit.

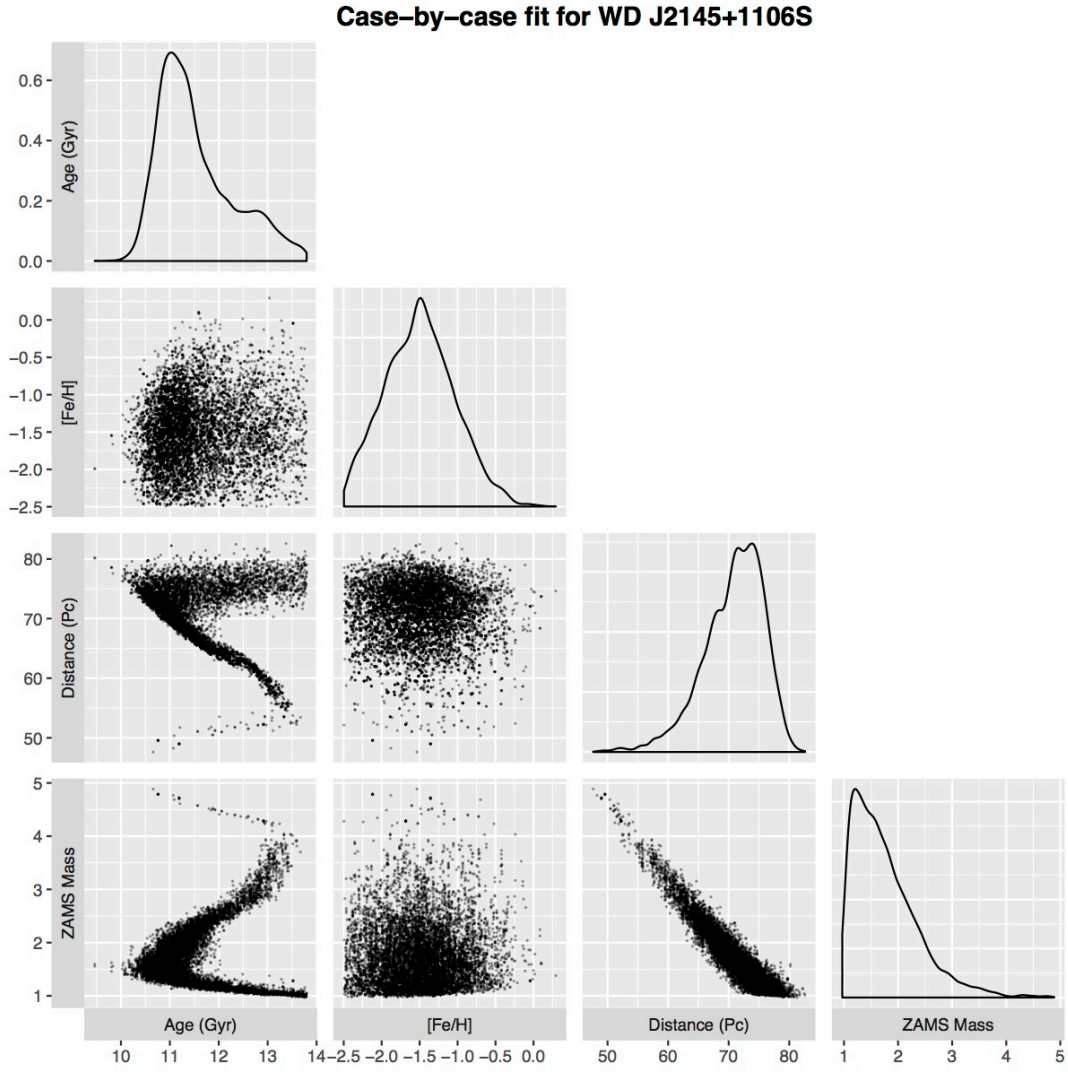


Figure A.10: The scatterplots of parameters of WD J2145+1106S from the case-by-case fit.

B Figures and Tables for the IFMR Project

B.1 Photometry for WDs in Clusters

In this section, Tables B.1 – B.3 give photometry for WDs of five clusters used in the data analysis section in Chapter 5.

B.2 Cluster-specific Results

In this section, we present the case-by-case (cluster-specific) fits of five clusters used in the data analysis section in Chapter 5. Figs B.1 – B.5 display the two-dimensional scatterplots and histograms of cluster parameters for five clusters, respectively.

Table B.1: Photometry for WDs in the Hyades.

Cluster	WD	$U \pm \sigma_U$	$B \pm \sigma_B$	$V \pm \sigma_V$
Hyades	HZ14	9.155 ± 0.032	10.195 ± 0.03	10.351 ± 0.029
	VR16	9.298 ± 0.114	10.274 ± 0.114	10.373 ± 0.113
	HZ7	9.969 ± 0.025	10.859 ± 0.022	10.903 ± 0.016
	VR7	10.129 ± 0.049	10.977 ± 0.041	11.004 ± 0.035
	HZ4	11.136 ± 0.054	11.811 ± 0.054	11.725 ± 0.054
	LB227	11.034 ± 0.015	11.752 ± 0.015	11.697 ± 0.014

References: DeGennaro et al. (2009); Stein et al. (2013)

Table B.2: Photometry for WDs in the M67.

Cluster	WD	$B \pm \sigma_B$	$V \pm \sigma_V$
M67	WD 1	23.854 ± 0.01	23.505 ± 0.018
	WD 2	22.293 ± 0.01	22.047 ± 0.01
	WD 3	22.119 ± 0.01	21.91 ± 0.01
	WD 4	21.721 ± 0.01	21.646 ± 0.01
	WD 5	19.656 ± 0.01	19.837 ± 0.01
	WD 6	22.074 ± 0.01	21.834 ± 0.01
	WD 7	21.843 ± 0.01	21.611 ± 0.01
	WD 8	21.469 ± 0.01	21.29 ± 0.01
	WD 9	21.286 ± 0.01	21.152 ± 0.01
	WD 10	23.603 ± 0.01	23.224 ± 0.014
	WD 11	22.206 ± 0.01	21.899 ± 0.01
	WD 12	21.454 ± 0.01	21.216 ± 0.01
	WD 13	20.53 ± 0.01	20.527 ± 0.01
	WD 14	23.194 ± 0.01	22.82 ± 0.01
	WD 15	22.587 ± 0.01	22.275 ± 0.01
	WD 16	24.227 ± 0.01	23.693 ± 0.019
	WD 17	22.359 ± 0.01	22.061 ± 0.01
	WD 18	22.855 ± 0.01	22.493 ± 0.01
	WD 19	23.552 ± 0.01	23.181 ± 0.012
	WD 20	21.243 ± 0.01	21.204 ± 0.01
	WD 21	18.579 ± 0.01	18.89 ± 0.01
	WD 22	24.53 ± 0.013	23.97 ± 0.026
	WD 23	24.576 ± 0.013	24.053 ± 0.028
	WD 24	21.563 ± 0.01	21.298 ± 0.01
	WD 25	23.977 ± 0.01	23.61 ± 0.018
	WD 26	24.147 ± 0.01	23.678 ± 0.019
	WD 27	22.757 ± 0.01	22.413 ± 0.01
	WD 28	24.048 ± 0.01	23.641 ± 0.017
	WD 29	23.845 ± 0.01	23.444 ± 0.017
	WD 30	24.156 ± 0.01	23.979 ± 0.022
	WD 31	22.298 ± 0.01	22.022 ± 0.01
	WD 32	24.234 ± 0.01	23.838 ± 0.02
	WD 33	21.803 ± 0.01	21.573 ± 0.01
	WD 34	22.517 ± 0.01	22.172 ± 0.01
	WD 35	23.849 ± 0.01	23.299 ± 0.019

References: Bellini et al. (2010a,b)

Table B.3: Photometry for WDs in the NGC188, NGC2168 and NGC2477.

Cluster	WD	$U \pm \sigma_U$	$B \pm \sigma_B$	$V \pm \sigma_V$	$I \pm \sigma_I$
NGC188	WD 1	...	...	22.651 ± 0.054	22.496 ± 0.124
	WD 2	...	...	23.35 ± 0.119	23.538 ± 0.386
	WD 3	...	...	23.49 ± 0.06	23.413 ± 0.185
	WD 4	...	...	23.508 ± 0.105	23.416 ± 0.559
	WD 5	...	...	23.669 ± 0.045	23.176 ± 0.149
	WD 6	...	...	24.206 ± 0.131	23.756 ± 0.228
	WD 7	...	...	24.439 ± 0.074	24.078 ± 0.255
	WD 8	...	...	24.469 ± 0.054	23.72 ± 0.195
	WD 9	...	...	25.261 ± 0.12	24.422 ± 0.297
NGC2168	WD 1	19.858 ± 0.019	20.953 ± 0.02	20.989 ± 0.019	...
	WD 2	20.532 ± 0.028	21.63 ± 0.031	21.569 ± 0.032	...
	WD 3	20.506 ± 0.02	21.364 ± 0.022	21.216 ± 0.02	...
	WD 4	18.764 ± 0.017	19.937 ± 0.018	20.065 ± 0.017	...
	WD 5	18.568 ± 0.017	19.735 ± 0.017	19.863 ± 0.016	...
	WD 6	20.339 ± 0.025	21.348 ± 0.028	21.303 ± 0.026	...
	WD 7	20.826 ± 0.023	21.759 ± 0.028	21.701 ± 0.027	...
	WD 8	19.727 ± 0.024	20.746 ± 0.024	20.785 ± 0.022	...
	WD 9	18.72 ± 0.018	19.665 ± 0.017	19.657 ± 0.016	...
	WD 10	20.479 ± 0.026	21.488 ± 0.029	21.398 ± 0.026	...
	WD 11	20.28 ± 0.025	21.65 ± 0.031	21.631 ± 0.029	...
	WD 12	19.558 ± 0.019	20.645 ± 0.019	20.719 ± 0.019	...
	WD 13	20.243 ± 0.02	21.17 ± 0.022	21.175 ± 0.02	...
NGC2477	WD 1	...	...	23.108 ± 0.008	22.816 ± 0.018
	WD 2	...	...	23.689 ± 0.01	23.237 ± 0.025
	WD 3	...	...	23.566 ± 0.011	23.282 ± 0.024
	WD 4	...	...	23.972 ± 0.016	23.638 ± 0.034
	WD 5	...	...	23.957 ± 0.015	23.586 ± 0.031
	WD 6	...	...	23.963 ± 0.015	23.576 ± 0.03
	WD 7	...	...	23.904 ± 0.012	23.461 ± 0.031

References for NGC188: von Hippel and Sarajedini (1998); Meibom et al. (2009)

References for NGC2168: Sung and Bessell (1999); Williams et al. (2004)

References for NGC2477: Jeffery et al. (2011); Stein et al. (2013)

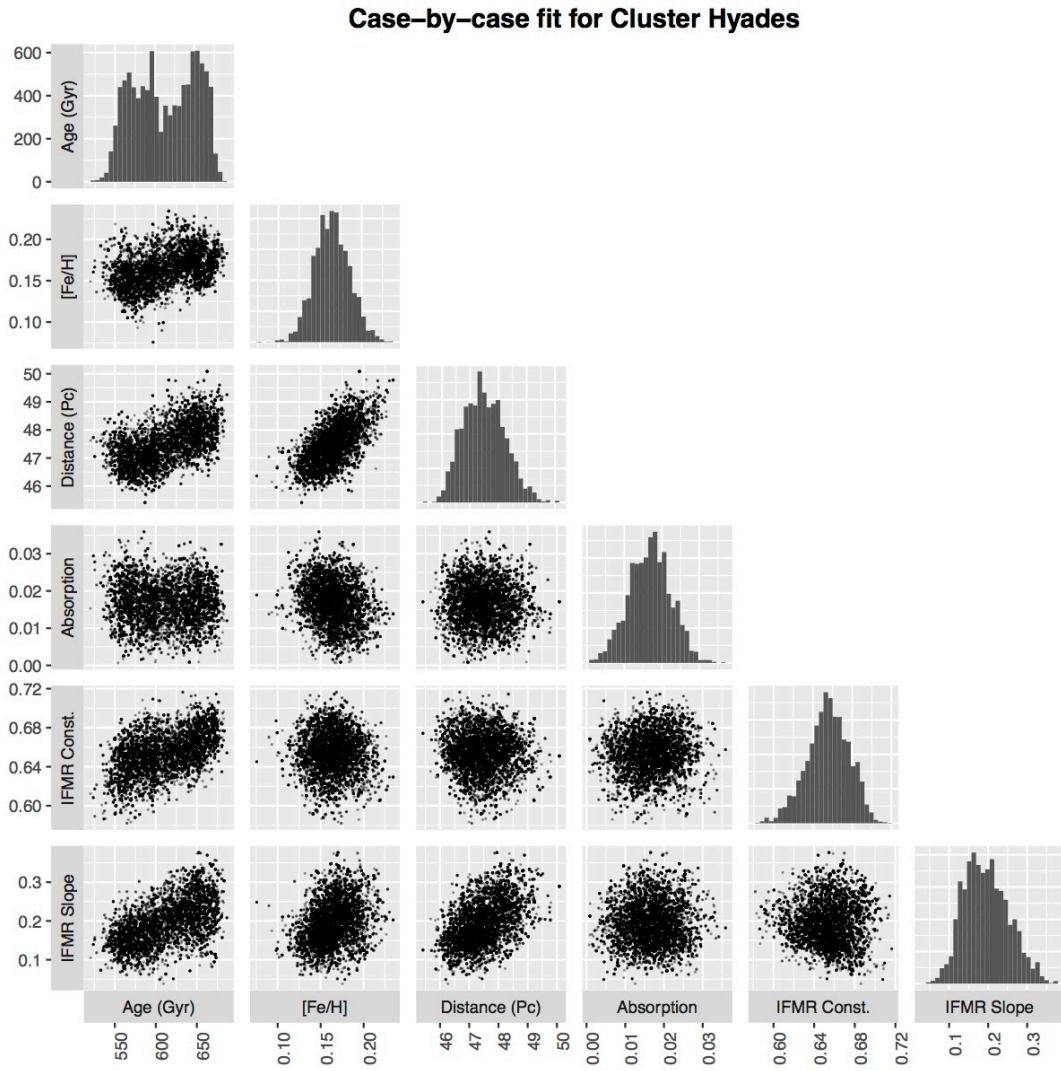


Figure B.1: The scatterplots of Hyades parameters from the case-by-case fit.

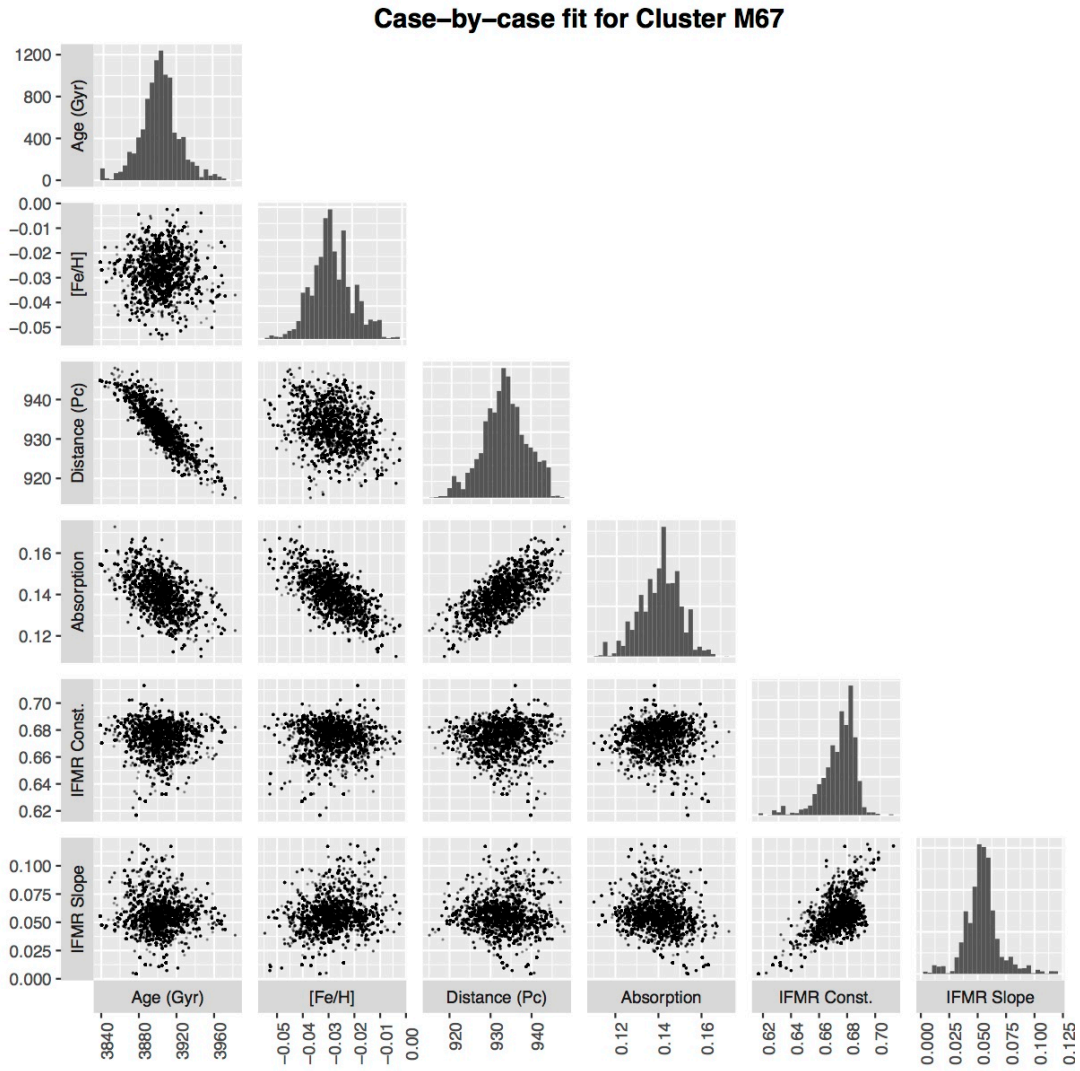


Figure B.2: The scatterplots of M67 parameters from the case-by-case fit.

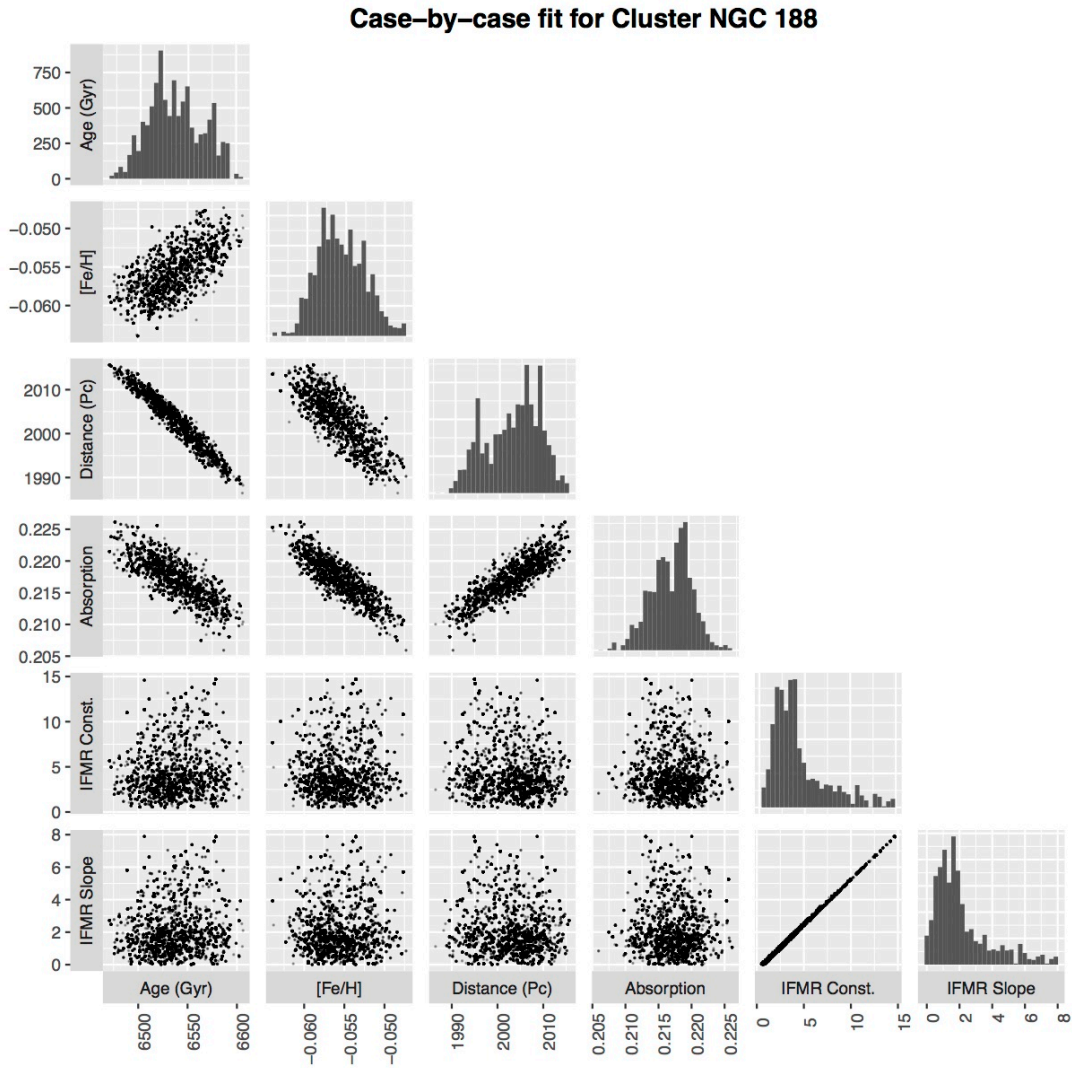


Figure B.3: The scatterplots of NGC 188 parameters from the case-by-case fit.

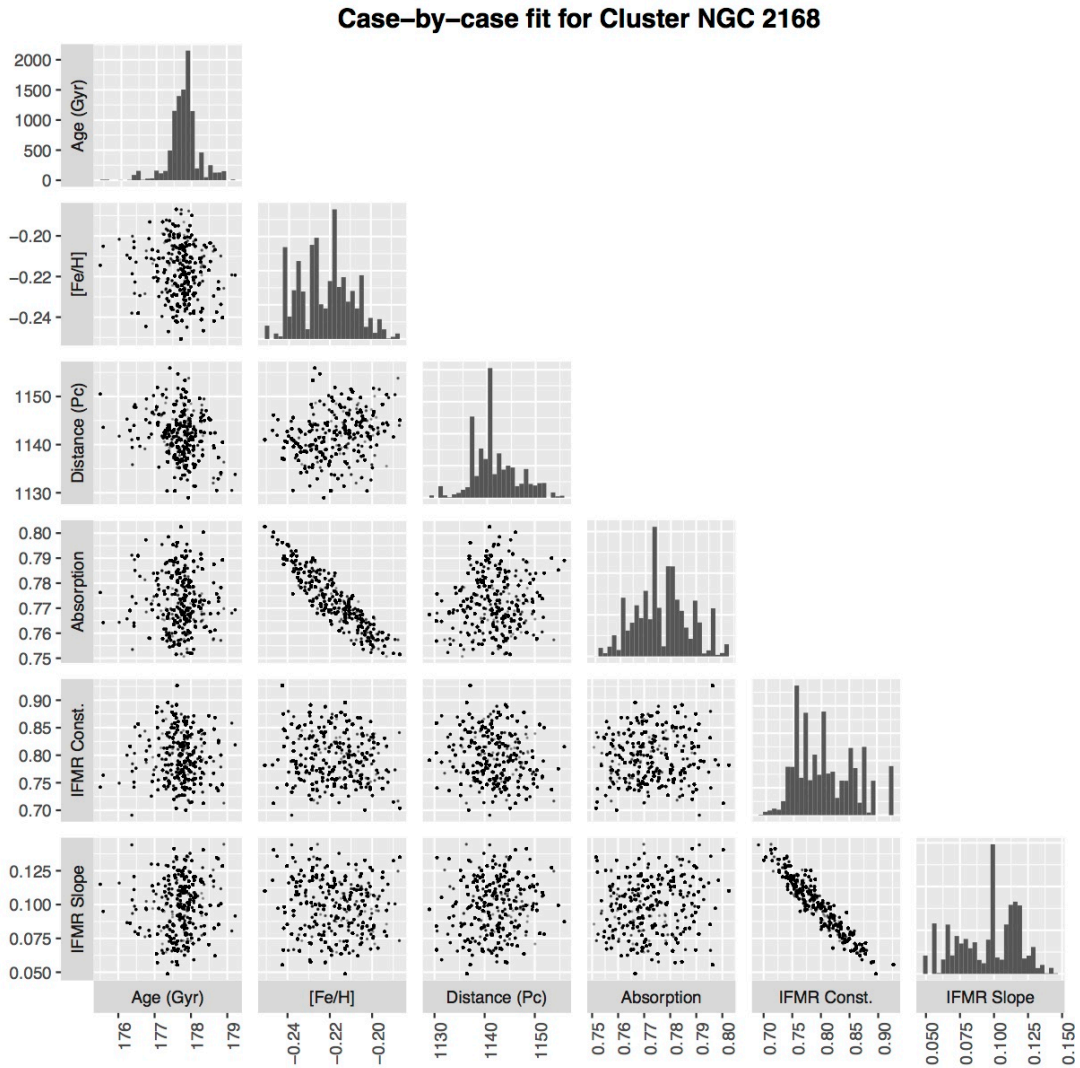


Figure B.4: The scatterplots of NGC 2168 parameters from the case-by-case fit.

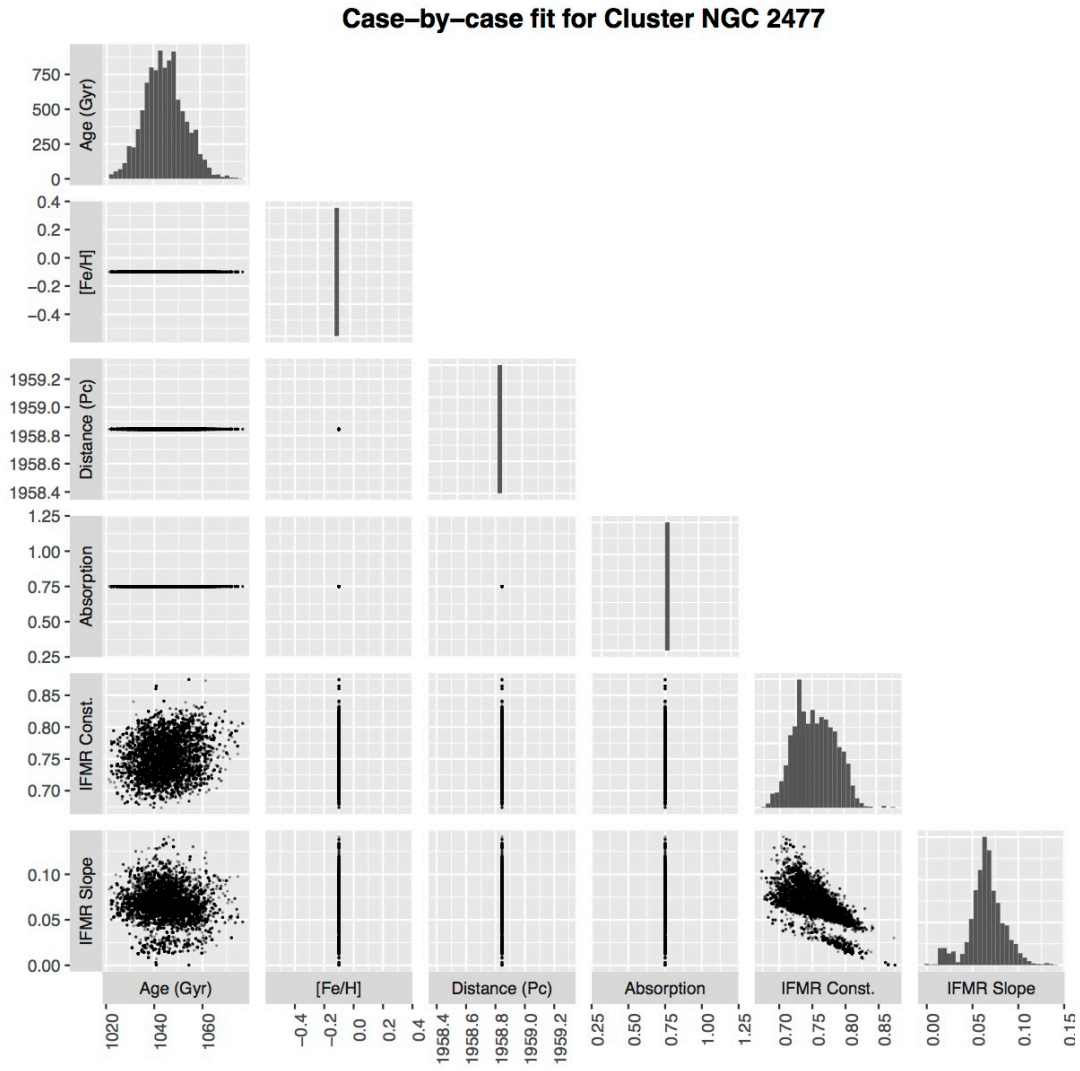


Figure B.5: The scatterplots of NGC 2477 parameters from the case-by-case fit.

Acronyms

CI Confidence Interval.

CMD colour-magnitude diagram.

CP Coverage Probability.

EB empirical Bayes.

EB-inv Empirical Bayes with inverse transformation.

EB-log Empirical Bayes with log transformation.

FB Fully Bayesian.

FB-Gibbs Fully Bayesian with Gibbs Sampler.

FB-TS Fully Bayesian with Two-Stage Sampler.

IFMR initial-final mass relation.

IMF Initial Mass Function.

IS Importance Sampling.

LMC Large Magellanic Cloud.

MA Metropolis Algorithm.

MAE Mean Absolute Error.

MAP Maximum A Posterior.

MCMC Markov chain Monte Carlo.

MHA Metropolis-Hastings Algorithm.

MLE Maximum Likelihood Estimate.

MS main-sequence.

RG red giant.

RMSE Root Mean Squared Error.

SIR Sampling Importance Resampling.

WD white dwarf.

ZAMS Zero-Age Main Sequence.

Glossary

EM algorithm Short for Expectation-Maximisation algorithm. It is a statistical algorithm to find modal estimates.

Empirical Bayes An approach for fitting Bayesian models.

Fully Bayesian An approach for fitting Bayesian models.

Large Magellanic Cloud A satellite galaxy of the Milky Way.

MCEM algorithm Short for Monte Carlo Expectation-Maximisation algorithm. It is the Monte Carlo version of the EM algorithm.

MCMC Short for Markov chain Monte Carlo, a collection of computational algorithms in statistics that simulate samples from unnormalised probability density functions.

MLE Short for Maximum Likelihood Estimate, a point estimation principle in statistics that maximises the likelihood function.

Zero-age Main Sequence mass The mass of a star when it starts to be a main sequence star.