

IMPERIAL

**BIAS-CORRECTED SHARPE RATIOS:
CLOSING THE GAP BETWEEN
IN-SAMPLE & OUT-OF-SAMPLE**

JOSEPH MULLIGAN

A thesis submitted under the supervision of
Prof. Antoine Jacquier
Prof. Johannes Muhle-Karbe

Department of Mathematics
Imperial College London
March 2026

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London

Abstract

This thesis investigates how we should correct in-sample Sharpe ratios in practice. It is well observed that out-of-sample Sharpe ratios are typically lower than their in-sample equivalents, sometimes significantly so. This phenomenon has historically been studied from multiple testing perspectives (“p-hacking”) or attributed to alpha decay (opportunities being arbitrated away). Existing literature on fund and factor performance has separately investigated how complexity and statistical bias can cause this effect within their respective domains. This work unifies these strands, offering both theoretical and practical results to aid in the correction of over-optimistic in-sample Sharpe ratios.

First, we address the problem from a structural perspective. By adopting a linear model for returns and a classical Markowitz portfolio construction, we derive theoretical expressions for the in- and out-of-sample mean and variance of the proposed trading strategy. We explicitly quantify the inflation experienced by the in-sample Sharpe ratio as the number of parameters in the model increases, and demonstrate how this is mitigated through additional data points and stronger signals. These theoretical results are validated through both simulation and empirical studies.

Second, we shift to a statistical approach, where the construction of the backtest is unknown. We propose a Bayesian method to estimate the true Sharpe ratio of new candidate investments by modelling the observed distribution of in- and out-of-sample Sharpe ratios of suitably homogeneous comparable investments. This method accounts for both selection bias (a form of overfitting) and alpha decay, yielding significantly more accurate estimates of out-of-sample performance. This is evidenced through an in-depth application of the model to a large dataset of hedge fund returns, where it significantly improves upon naive baselines.

Acknowledgements

I would like to thank Professors Antoine Jacquier and Johannes Muhle-Karbe for their constant support throughout the production of this thesis. It has been said that the output of a PhD is the student, and I am certain that I have both Antoine and Johannes to thank for the development I have experienced during this process. Their complementary approaches were central in pushing me forward on multiple fronts, and I thoroughly enjoyed learning from them both over the course of the PhD.

Qube Research & Technologies were instrumental in allowing me to pursue this opportunity, and for that I am thankful. It has been my great pleasure to discuss the ideas contained within this thesis with the researchers at QRT, and to discuss all manner of other topics too. I learned a great deal from the experience, and was honoured to get to know so many friendly, funny, and formidable people over the last few years.

To my parents, thank you, for everything. I could not fit on a hundred pages everything I am thankful to you for, your endless support in every regard has made it possible for me to be here today writing this. You inspire me every day. To my siblings, cousins, grandmothers, aunties, and uncles, you bring so much into my life, and I'm sure you'll all be glad to hear that yes, I am finally done being a student. To Caitlin, thank you for your unwavering love, support and laughter, you are a wonder, and bring light into every day. I have been fortunate to have many wonderful friends during this journey, both old and new, thank you to you all, you make it all so much more fun. Finally, I would be remiss not to make note of my wonderful colleagues in the Weeks office over the last four years; I look forward to seeing the fantastic things you all achieve in the years to come.

Declaration of Originality

I hereby declare that the work presented in this thesis is my own, unless otherwise explicitly stated, and that all content resulting from the collaboration with others has been appropriately acknowledged.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Funding Statement

This research was supported by Qube Research & Technologies and the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1).

Disclaimer

Unless specifically indicated otherwise, the views and opinions expressed in this communication are those of the author(s) and do not reflect the official policy or position of [Qube Research & Technologies Ltd]. The information provided herein is for informational purposes only and should not be construed as advice, financial or otherwise, or opinion of [Qube Research & Technologies Ltd]. [Qube Research & Technologies Ltd] has neither approved nor disapproved of such content and makes no representations as to the accuracy, completeness, or reliability of any information contained herein.

Recipients are advised not to rely solely on this communication for any purpose. [Qube Research & Technologies Ltd] shall not be liable for any errors or omissions, nor for any damages resulting from reliance on this information.

Contents

Abstract	i
Acknowledgements	ii
Declarations	iii
List of Figures	vii
List of Tables	xii
1 Introduction	1
1.1 Background & Motivation	1
1.1.1 Multiple Testing Approaches	3
1.1.2 Model Complexity	6
1.1.3 Cross-Sectional Methods	8
1.2 Thesis Outline	12
Appendices	14
1.A Application of Harvey, Liu, and Zhu to the Open Source Asset Pricing Data	14
1.B Sharpe Ratios Are Angles	18
1.C Overview of Implementability, Transaction Costs & Capacity Literature	19
2 In-Sample and Out-of-Sample Sharpe Ratios for Linear Predic- tive Models	23
2.1 Introduction	24
2.2 Model Setup	28
2.2.1 The True Sharpe Ratio	29
2.2.2 The In-Sample Sharpe Ratio	31
2.2.3 The Out-of-Sample Sharpe Ratio	33
2.3 Main Results	33
2.3.1 Accuracy of the Approximation of the In-Sample Variance . .	35

2.3.2	The Replication Ratio	37
2.3.3	Unknown Covariances	42
2.3.4	Autoregressive Signals	44
2.4	Comparison with the Literature	46
2.5	Predicting Commodity Futures	48
2.5.1	Gaussian and iid Signals and Residuals	51
2.5.2	AR(1) Signals and t -distributed Residuals and Shocks	52
2.6	Empirical Study: Goyal, Welch, and Zafirov Dataset	53
2.7	Conclusion	57
	Appendices	59
2.A	Derivation of Equation (2.2.2)	59
2.B	General Case Result of Proposition 2.3.1	60
2.C	Proof of Proposition 2.B.1	62
2.C.1	In-sample expected average	63
2.C.2	In-sample expected variance	65
2.C.3	Out-of-sample expected average	74
2.C.4	Out-of-sample expected variance	75
2.D	Monte Carlo Details	78
2.E	Additional Values for Section 2.5 and Section 2.6	79
3	Correcting In-Sample Sharpe Ratios for Selection Bias and Alpha Decay	83
3.1	Introduction	84
3.2	Bayesian Model: Theory & Implementation	89
3.2.1	Preliminaries	89
3.2.2	Model Specification	91
3.2.3	Model Application	94
3.3	Empirical Case Study: Hedge Fund Returns	99
3.4	Conclusion	114
	Appendices	115
3.A	Extensions, Other Cases & Additional Theory	115
3.A.1	Asymptotic Distributions of the Sample Sharpe Ratio	115
3.A.2	Other Types of Selection Bias	119
3.B	Implementation & Robustness	123
3.B.1	The Necessity of Joint Fitting, Alpha Decay & Parameter Recovery	123
3.B.2	Numerical Implementation Details	128

3.B.3	Hybrid Parameter Estimation	130
3.B.4	Effect of Correlated Sharpe Ratios	133
3.C	Supplementary Material for Section 3.3	137
3.C.1	Filter Criteria	137
3.C.2	Deriving Independent Returns	138
3.C.3	Fitting Approach	143
3.C.4	Fund Clustering	143
3.C.5	Comparator Models & Metrics	150
3.C.6	Walkforward Analysis	155
3.C.7	Additional Figures & Tables	156
3.D	Illustrative Examples	162
3.E	Comparison to Chen and Zimmermann	169
3.F	Proofs	179
3.F.1	Proof of Proposition 3.A.1: Asymptotic Variance for the Gen- eral Likelihood	179
4	Conclusion	185
	Bibliography	187

List of Figures

1.A.1	Distribution of in- and out-of-sample t -statistics from the Open-Source Asset Pricing Dataset with our pre-processing approach in Appendix 3.E.	15
1.A.2	Replication of Figure 3 from Harvey, Liu, and Zhu using the Open-Source Asset Pricing Dataset and our pre-processing approach in Appendix 3.E.	16
1.A.3	Amount of explained reduction in performance between the in-sample and out-of-sample t -statistic.	16
1.A.4	Number of implied tests required for the Bonferroni correction to adjust the in-sample t -stat to match the out-of-sample t -stat.	16
1.A.5	Distribution of out-of-sample p -values under various corrections, and the accompanying distribution of in- and out-of-sample period lengths.	17
1.A.6	Distribution of p -values using the observed out-of-sample Sharpe ratios and the maximum of the in- and out-of-sample period lengths to convert this to a t -statistic, and then p -value.	18
1.B.1	The Sharpe ratio for two points.	19
2.1.1	The replication ratio for the case study in Section 2.5 with varying backtest length.	27
2.3.1	Values of N, K, T_{IS} where $\tilde{c}_2 > 0$	36
2.3.2	Magnitude of the total errors $\sum_i \varepsilon_i$ as a percentage of the in-sample variance $\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{IS}}]]$	37
2.3.3	Difference between the expected Sharpe Ratio $\mathbb{E}[\text{SR}]$ and the ratio of the expected mean and variance $\mathbb{E}[\widehat{\mathbb{E}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_i}]] / \sqrt{\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_i}]]}$, i.e. the error due to the missing convexity adjustment for both in-sample and out-of-sample Sharpe ratios.	38
2.3.4	Replication ratio $\frac{\text{SR}_{\text{EOOS}}}{\text{SR}_{\text{EIS}}}$, SR_{EIS} and SR_{EOOS} in the 1-asset, 1-signal case as in (2.3.3), with varying SR, T_{IS}	40

2.3.5	Replication ratio $\frac{SR_{E\text{OOS}}}{SR_{EIS}}$ for increasing assets N and signals K , and fixed in-sample Sharpe ratio $SR_{EIS} = 3$ with $T_{IS} = 252 \cdot 100$. Large T_{IS} is required to keep SR_{EIS} fixed at 3 as p, m increase, with smaller T_{IS} we cannot always find β such that $SR_{EIS} = 3$	41
2.3.6	Replication ratio, IS Sharpe ratio and OOS Sharpe ratio as assets increase under different correlations	42
2.3.7	In- and out-of-sample Sharpe ratio and replication ratio using either the true or estimated covariance matrices $\Sigma_\epsilon, \hat{\Sigma}_\epsilon$ to form $\hat{\mathbf{w}}_t$ in Monte Carlo simulations.	44
2.3.8	Impact of s_t following an AR(1) process. The top row shows the replication ratio, and the bottom row shows the simple difference of this ratio vs $\phi = 0$, corresponding to the iid case.	46
2.4.1	Comparison of the replication ratio for Kan, Wang, and Zheng and ours with $T = T_{IS} = T_{OOS} = 2520$ and $SR = 1.5$ for varying N . For ours we set $K = 10$ or $K = 1$	47
2.5.1	The expected in-sample Sharpe ratio, expected out-of-sample Sharpe ratio and replication ratio for the experiments in Section 2.5.1. . . .	52
2.5.2	The expected in-sample Sharpe ratio, expected out-of-sample Sharpe ratio and replication ratio for the experiments in Section 2.5.2. . . .	53
2.6.1	Empirical and implied analytical replication ratio by number of signals K , backtest length T_{IS} and implied true Sharpe ratio SR (annualised).	57
3.1.1	Performance deterioration between in-sample and out-of-sample performance in different domains. Sources are: Thematic ETFs & Sector ETFs (estimated), ETFs, Alternative Beta Strategies, Hedge Funds, Factors, Anomalies and Quantitative Strategies.	85
3.1.2	Average pre- and post-listing Sharpe ratios over time, with a 15-year rolling lookback, and the average Bayesian posterior mean for each date.	87
3.3.1	Cumulative (normalised) average returns vs months pre- and post-listing split into the three identified performance periods.	102
3.3.2	Empirical distribution of pre- and post-listing Sharpe ratios by period. Empirical means are marked with dashed lines and annotated with their value, and the fitted marginal likelihood from the Bayesian model is shown with a solid line.	103

3.3.3	Unadjusted, bias adjusted, and bias & decay adjusted prior densities for the three periods.	104
3.3.4	Bayesian estimated haircuts by period, with the posterior estimated either with pre-listing data alone, or with pre- and post-listing data.	106
3.3.5	Improvement in statistical metrics relative to a naive baseline for the models described in Appendix 3.C.5.	108
3.3.6	Sharpe ratios for the posterior based trading strategy for the Naive, James–Stein and Bayesian models by Sharpe ratio cutoff, with the probability threshold set at 50%.	112
3.3.7	Sharpe ratios for the posterior based trading strategy for the Naive, James–Stein and Bayesian models by probability threshold, with the Sharpe ratio cutoff set at 0.5.	112
3.3.8	Improvement over the benchmark for the Bayesian strategy by Sharpe ratio cutoff and probability threshold.	113
3.3.9	Sharpe ratios for the predictive based trading strategy for the Naive, James–Stein and Bayesian models by quantile threshold.	114
3.A.1	Distributions of autocorrelation ϕ , skewness $\gamma_{3,r}$ and excess kurtosis $\gamma_{4,r} - 3$ for the hedge fund returns used in Section 3.3.	117
3.A.2	In- and Out-of-Sample Sharpe ratio distributions as functions of the autocorrelation of returns ϕ , skewness of returns $\gamma_{3,r}$ and kurtosis of returns $\gamma_{4,r}$	118
3.B.1	Indistinguishable marginal likelihoods: an unbiased gamma prior, $\Gamma(1, 10)$ versus biased selection with a t prior, $lst(-0.07, 0.02, 7.58)$ with selection parameters $\kappa = 0, \ell = 28$ (see (3.2.1)).	123
3.B.2	Comparing fitting with in-sample data only, in- and out-of-sample data, and in- and out-of-sample data without allowing for alpha decay.	125
3.B.3	Kullback–Leibler divergence between the fitted and true models. The plot shows how fit quality varies with sample size when using only in-sample data versus using both in- and out-of-sample data.	127
3.B.4	Relative errors in recovered parameters for the two fitting methods. The plots show how the bias and variance of the parameter estimates change with the number of samples used for fitting.	127
3.B.5	Average Mean Absolute Error of the posterior mean $\mathbb{E}[\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}]$ vs the true Sharpe ratio SR per fit using in-sample only or in- and out-of-sample data.	128
3.B.6	The impact of correlation, controlled by ρ , on the estimation of SR_0 and ζ_0^2	137

3.C.1	Proportion of funds with non-zero loading on the top 10 factors, and the proportion of funds that load on either of the two market factors. BAB denotes “Betting Against Beta” and TS Mom. denotes “Time Series Momentum”.	140
3.C.2	Pairwise correlations between fund returns, before and after removal of systematic factor contributions.	141
3.C.3	Distribution of Sharpe ratios before and after removal of systematic factor contributions, split into the pre- and post-listing period for each fund.	142
3.C.4	Cumulative P&L of inverse volatility weighted combination of raw and non-systematic returns.	142
3.C.5	Explained variance by number of components for the FAMD analysis.	144
3.C.6	Silhouette score by number of clusters.	145
3.C.7	Number of funds per cluster, and per period	146
3.C.8	Top features for cluster selection, sorted by mean SHAP Value	146
3.C.9	Cluster composition by style.	147
3.C.10	Pre- and Post-Listing Sharpe ratio distribution by cluster.	148
3.C.11	Median pre- to post-listing Sharpe ratio haircuts by cluster, decomposed into selection bias and alpha decay.	150
3.C.12	Improvement in statistical metrics relative to a naive baseline when fitted with a walkforward approach.	157
3.C.13	Average AUM in USD vs months pre- or post-listing in commercial databases.	157
3.C.14	Average AUM in USD over time.	158
3.C.15	Total AUM in USD, and total fund count, by style over time.	158
3.C.16	Total AUM in USD, and total fund count, by listing status over time.	159
3.C.17	Cumulative distribution of period lengths.	159
3.C.18	Fitted parameters for the walkforward analysis. Solid lines show the fitted parameters for the 1996 start date, and dashed lines for the 2009 start date. The y-axis is truncated for each plot for clarity.	161
3.D.1	Prior distribution of true Sharpe ratios under unbiased, fixed, hierarchical and random selection.	163
3.D.2	Densities for the hyperprior and conditional prior of the hierarchical case. Note these are not bias adjusted.	164
3.D.3	Densities for the in- and out-of-sample Sharpe ratios in the unbiased, fixed, hierarchical and random cases.	164

3.D.4	Posteriors for the fixed, hierarchical, random and unbiased cases; $\widehat{\text{SR}}_{\text{IS}} = 1.5$, $T_{\text{IS}} = 252 \cdot 3$	165
3.D.5	Replication ratios for unbiased/random selection bias vs fixed selection bias.	166
3.D.6	Replication ratio $\mathbb{E}[\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}]/\widehat{\text{SR}}_{\text{IS}}$ by variance split ω	167
3.D.7	Replication ratio by SR_0 and ς_0^2	168
3.D.8	Replication ratio by κ and ℓ	168
3.E.1	Cumulative long-short returns for the cash flow yield variance predictor using the original or Markowitz construction.	171
3.E.2	In- and post-sample long-short t -statistics for original and Markowitz portfolios, dashed lines denote means.	172
3.E.3	Model fit (marginal likelihoods) for in-sample only model (with $\text{SR}_0 = 0$) versus in- and pre-publication model.	174
3.E.4	Unbiased and bias adjusted priors for in-sample only model (with $\text{SR}_0 = 0$) versus in- and pre-publication model.	175
3.E.5	Percent change from the in-sample Sharpe ratio to the posterior mean $\mathbb{E}[\text{SR} \mid \widehat{\text{SR}}_{\text{IS},i}]/\widehat{\text{SR}}_{\text{IS},i} - 1$ for in-sample only model (with $\text{SR}_0 = 0$) versus in- and post-sample model.	176
3.E.6	Distributions of in-sample Sharpe ratios, posterior means, decayed posterior means, and the out-of-sample Sharpe ratios, for the pre-publication and post-sample models.	177
3.E.7	Attribution of average adjustments into selection bias, alpha decay and noise for the pre-publication and post-sample models.	178

List of Tables

2.5.1	Commodities selected for testing from the LME (London Metal Exchange), ICE (Intercontinental Exchange), NYMEX (New York Mercantile Exchange), COMEX (Commodity Exchange) and NYBOT (New York Board of Trade).	49
2.E.1	Simulation Results comparing Analytical and Estimated values under Gaussian IID and AR(1)+t assumptions for Section 2.5. Sharpe ratios are annualised assuming 252 trading days.	80
2.E.2	Statistics from resampling study, comparing analytical (implied) and empirical values for Section 2.6. Sharpe ratios are annualised from monthly values.	82
3.3.1	Fitted parameters for the Bayesian model by period. Annualised equivalents are added in (brackets) for SR_0 , ς_0 and κ	103
3.3.2	Unadjusted, bias adjusted and bias & decay adjusted expected true Sharpe ratios $\mathbb{E}[SR]$ for the three periods.	105
3.A.1	Prior, selection, sample and decay parameters for the experiments of Appendix 3.A.1.	119
3.A.2	Bias adjusted priors and likelihoods.	122
3.B.1	Prior, selection, sample and decay parameters for the experiments of Appendix 3.B.1.	125
3.B.2	Fitted parameters from fitting with in-sample data only, in- and out-of-sample data, and in- and out-of-sample data without allowing for alpha decay.	125
3.B.3	Prior, selection, sample and decay parameters for the subsampling experiments using the random selection bias of Appendix 3.B.1. . .	126
3.B.4	Mapping between model parameters and empirical moments used in GMM estimation. When out-of-sample data is unavailable, all OOS moments are omitted. $\gamma_4(\cdot)$ denotes the kurtosis.	133
3.C.1	Qualities supplied to FAMD, grouped by theme. All performance qualities are computed using both the raw and non-systematic returns.	144
3.C.2	Cluster composition. Note most clusters are primarily US firm domicile, bar cluster 22 which is European.	147

3.C.3	Posterior parameter estimates by cluster and period. SR_0, ς_0 and κ are reported in annualised terms, divide by $\sqrt{12}$ to yield the original parameter estimates.	149
3.C.4	Fitted parameters by model and period. For SR_0, ς_0 and κ , annualised equivalents are shown in parentheses. A dash (–) indicates that the parameter is not identified in the corresponding model. . .	160
3.D.1	Prior, selection and sample parameters for the experiments of Appendix 3.D.	162
3.E.1	Fitted parameter values for Appendix 3.E. Note that the parameter values are in monthly terms.	173

1 | Introduction

1.1	Background & Motivation	1
1.2	Thesis Outline	12
	Appendices	14
1.A	Application of Harvey, Liu, and Zhu to the Open Source Asset Pricing Data	14
1.B	Sharpe Ratios Are Angles	18
1.C	Overview of Implementability, Transaction Costs & Capacity Literature	19

1.1 Background & Motivation

In this thesis we explore one of the most frustrating features of investing: the difference between in- and out-of-sample performance. Specifically, we study the problem from both a structural perspective (how much does overfitting in a tractable model cause in-sample performance to be inflated relative to out-of-sample performance?) and a more behavioural one (how does searching for higher Sharpe ratios lead to statistical biases in the measurement of performance?).

Trying to predict the future is a long-standing problem in finance. In 1933 Alfred Cowles [Cow33] analysed the forecasts of 45 analysts who attempted to predict either specific stocks which would outperform, or the general direction of the market overall. He found that there was no evidence they had any skill at all. In his article, he describes some analysts as seeming to channel the “Delphic Oracle”: the ancient Greek priestess famous for her often cryptic prophecies.

Relevant to our problem of interest, the Oracle at Delphi was once subject to an ill-designed trial plagued by multiple testing issues. In the histories of Herodotus c. 425 BCE [HWD08], we are told the story of Croesus and the oracles [Tho16]. Croesus wished to know which of the oracles had genuine access to the “mind and

will of the gods”. To that end, he sent envoys to all the major oracles of the Greek world, and on a fixed day they were asked: “What is Croesus, son of Alyattes, king of the Lydians, doing at this moment?” The Delphic Oracle pronounced (in hexameter verse) that he was cooking lamb and tortoise together in a bronze pot, which, indeed, he was. Convinced of the oracular abilities of the Pythia, he went on to wage war against the Persians following her prediction that this would lead to the destruction of a mighty empire. Unfortunately, he was defeated by the Persians and later burned alive. While the stakes are not quite so high in finance, we should still be careful when measuring the accuracy of questionable predictors.

Historically, particularly within the finance literature, work on the issue of inflated in-sample statistics has primarily focused on multiple testing: the invalidation of standard statistical tests due to the testing of multiple hypotheses. But, recently, work has begun to be done on approaching this issue from new perspectives. Kan, Wang, and Zheng [KWZ24] study the in- and out-of-sample Sharpe ratio of the Markowitz portfolio constructed using historical (in-sample) estimates of the mean and covariance. In Chapter 2 we extend this approach to the classical setting of a linear predictive model, for example as used by Gârleanu and Pedersen [GP13] in their work on optimal dynamic portfolio construction. This allows us to study how the number of signals used to create a prediction, and their relative strength, affect the deterioration between in-sample and out-of-sample performance.

The other avenue which has seen recent advances is the study of why researchers perform multiple tests. Chen and Zimmermann [CZ20] propose a model of publication bias where studies with low t -statistics are not published (the “file drawer problem” [Ros79]). In Chapter 3 we also model the bias of in-sample statistics (specifically, Sharpe ratios rather than t -statistics) as a function of selection bias, but importantly extend the model with alpha decay and provide the significant additional methodological advances required to apply the model in practice and yield useful estimates for investment decisions. This more general approach than the pre-

ceding chapter allows us to tackle a wider array of problems, and has applications across the finance industry.

Before we cover these new techniques, we should review the substantive existing literature, and discuss why extensions to these approaches are required.¹

1.1.1 Multiple Testing Approaches

Research into performance measurement in the finance literature generally follows two main threads: (i) estimating mutual and hedge fund returns and alpha; and (ii) measuring the statistical significance of published asset pricing research, with a subset focused on multiple testing in quantitative trading strategies.

Harvey and Liu have produced significant work in both avenues. Their best-known contribution is on multiple testing in the asset pricing literature [HLZ16]. They collect a dataset of reported t -statistics for 316 published cross-sectional asset pricing factors and study how the significance threshold, adjusted for multiple testing, should rise as more factors are published. Specifically, they convert the published t -statistics into p -values, apply a Bonferroni [Bon35], Holm [Hol79] or Benjamini–Hochberg–Yekutieli (BHY) [BH95; BY01] correction, and convert the results back into t -statistics. Based on this analysis, they recommend raising the significance hurdle to at least 3 for newly published factors.²

For the academic asset pricing literature, this seems a reasonable suggestion. However, in our primary application, quantitative trading, statistical significance is typically not the primary concern. Instead, Sharpe ratios [Sha66; Sha94] are the preferred quantity of interest. The main reason for this is quite simple, Sharpe ratios provide a much more direct measure of profitability. Trading strategies typically target a fixed amount of volatility per day, month or year. Therefore, we can

¹ Note that in this thesis we do not deal with implementability, transaction costs or capacity. However, these are important topics in the context of in- and out-of-sample performance, so a brief overview of some relevant literature is provided in Appendix 1.C.

² We apply their approach to the Open Source Asset pricing dataset [CZ22] in Appendix 1.A.

quickly estimate the annual profit generated by a strategy using just its measured Sharpe ratio and the volatility target. t -statistics do not have this property. Due to the dependence on the number of observations, we could have a strategy with a statistically significant t -statistic while having an economically insignificant Sharpe ratio. The inverse is possible (spurious short term outperformance), but much less likely.

Harvey and Liu [HL15] propose a solution to this difference in preferred metrics by leveraging the link between the Sharpe ratio and the t -statistic:³

$$t = \text{SR}\sqrt{T},$$

where SR is the non-annualised (per-period) Sharpe ratio and T is the number of observations. We can therefore apply the multiple testing approach of [HLZ16] to Sharpe ratios: first, scale them into t -statistics to derive p -values; second, apply a multiple testing correction; and finally, reverse the procedure to calculate a haircut Sharpe ratio.

For the multiple testing adjustments above, the number of tests conducted is a primary input for the calculation of the adjusted p -values. In a similar vein, Bailey and López De Prado [BL14] investigate how the maximum observed Sharpe ratio increases as the number of samples drawn increases. Via a classic result from extreme value theory [EKM97], they have that the expected maximum observed Sharpe ratio after M trials is

$$\mathbb{E}\left[\max_{i \in [1, M]} \{\widehat{\text{SR}}_i\}\right] \approx \mathbb{E}[\widehat{\text{SR}}_i] + \sqrt{\mathbb{V}[\widehat{\text{SR}}_i]} \left((1 - \gamma) \Phi^{-1}\left(1 - \frac{1}{M}\right) + \gamma \Phi^{-1}\left(1 - \frac{1}{M}e^{-1}\right) \right),$$

where $\mathbb{E}[\widehat{\text{SR}}_i]$ and $\mathbb{V}[\widehat{\text{SR}}_i]$ is the (fixed) expectation and variance for the drawn sample Sharpe ratios, γ is the Euler–Mascheroni constant, Φ^{-1} is the percentage point function for the normal distribution and e is Euler’s number.⁴

³ This link also connects the Sharpe ratio to geometry; see appendix 1.B.

⁴ This has the unfortunate effect of mixing the distribution of the true unobserved Sharpe ratio with the sampling distribution of the Sharpe ratio. For a discussion on the difference between these two, see Section 3.2.1.

In a later paper, Bailey, Borwein, and Prado [BBP16] define the “Probability of Backtest Overfitting” (PBO): The probability that the optimal in-sample strategy will perform worse than the median of the M tested configurations out-of-sample,

$$\text{PBO} := \sum_{m=1}^M \mathbb{P} \left[\overline{\text{rank}}_m < \frac{M}{2} \mid \text{rank}_m = M \right] \mathbb{P}[\text{rank}_m = M]$$

where $\overline{\text{rank}}_m$ represents the ranking of strategy m out-of-sample, and $\text{rank}_m = M$ denotes the condition that strategy m is optimal in-sample. They propose a method they call Combinatorially Symmetric Cross-Validation (CSCV) to estimate this. The method involves splitting the data of all trials into S periods and recombining these to form $\binom{S}{S/2}$ pairs of in- and out-of-sample sets, which are denoted C_S . For each pair $j \in C_S$, denoting the index of the best strategy in-sample by m^* , the normalised rank of this strategy out-of-sample is computed,

$$\overline{\text{norm. rank}}_j = \frac{\overline{\text{rank}}_{m^*}^j}{M + 1},$$

and the PBO is estimated by the relative ranks less than 0.5:

$$\text{PBO} = \frac{1}{|C_S|} \sum_{j \in C_S} \mathbb{1}_{\overline{\text{norm. rank}}_j < 0.5}.$$

What do all the proposed methods above have in common? They require careful bookkeeping of the previously conducted tests. While suggestions have been made to move towards this goal [Lóp19], the full history of tested strategies is often unavailable, which makes application of any method that depends on it to be impractical for real world usage. Harvey and Liu [HL21b] have recently worked on estimating the total number of tests conducted in producing the set of published factors, but they concede that the model suffers from identification issues, and subjective choices must be made to come to any conclusion.

1.1.2 Model Complexity

This therefore motivates the search for other approaches which do not rely on this knowledge of unseen strategies. Instead, these methods typically focus on complexity. The first such contribution we wish to highlight is from Novy-Marx [Nov15] who studies how the Sharpe ratio of a portfolio constructed from the best k of K signals grows as a function of k and K . This can be seen as a more extreme case of the multiple testing adjustments previously reviewed. However, while historical iterations of a particular signal are often not tracked, if we wish to combine multiple signals into a portfolio we must clearly have them all to hand. As expected, as k and K increase, the observed t -statistics steadily increase, even when the signals being combined have no predictive power at all.

On a similar theme, Paulsen and Söhl [PS20] introduce the Sharpe Ratio Information Criterion (SRIC),

$$\text{SRIC} := \widehat{\text{SR}}_{\text{IS}} - \frac{K}{T_{\text{IS}} \widehat{\text{SR}}_{\text{IS}}},$$

where $\widehat{\text{SR}}_{\text{IS}}$ is the in-sample optimal Sharpe ratio, K is the number of free parameters in the model less one (for example, if you use a regression model with 5 signals, plus an intercept, $K = 6 - 1 = 5$), and T is the length of the backtest. They go on to prove that within their model the expected out-of-sample Sharpe ratio is equivalent to the expected SRIC. Therefore, the expected out-of-sample Sharpe ratio is the observed in-sample Sharpe ratio less a penalty for model complexity, which is reduced as either the length of the backtest increases or the in-sample Sharpe ratio increases.

Along a similar idea, Koshiyama and Firoozye [KF19] propose a ‘‘Covariance-Penalty Correction’’ to address backtesting overfitting. They work in a single asset setting where the next step returns are predicted by a linear model, and the held position (the model predicted return) is jointly Gaussian with the returns. They

derive the estimate for the out-of-sample Sharpe ratio in this setup to be:

$$\mathbb{E}[\widehat{SR}_{\text{OOS}}] \approx \widehat{SR}_{\text{IS}} - \frac{2 \operatorname{tr}(\mathbf{H})}{T_{\text{IS}}} \left(1 - \widehat{SR}_{\text{IS}}^2\right)^{3/2},$$

where $\mathbf{H} = \mathbf{S} (\mathbf{S}^\top \mathbf{S})^{-1} \mathbf{S}^\top \in \mathbb{R}^{T_{\text{IS}} \times T_{\text{IS}}}$ is the “smoothing” matrix, $\mathbf{S} \in \mathbb{R}^{T_{\text{IS}} \times K}$ is the vector of predictors for the returns $\mathbf{r} \in \mathbb{R}^{T_{\text{IS}}}$, T_{IS} is the number of observed periods and the in-sample Sharpe ratio is computed by the correlation $\widehat{\rho}$ between the position $\mathbf{W} = \mathbf{S}\boldsymbol{\beta}$ ($\boldsymbol{\beta}$ being the loading from the predictors to the returns) and the returns: $\widehat{SR}_{\text{IS}} = \frac{\widehat{\rho}}{\sqrt{1+\rho^2}}$.⁵ Once again, we see that the expected out-of-sample Sharpe ratio is the in-sample Sharpe ratio less a penalty for complexity ($\operatorname{tr}(\mathbf{H}) = K$ under OLS), with this reducing as T_{IS} and $\widehat{SR}_{\text{IS}}$ increase.

More recently, Meng, Cao, and Wang [MCW25] have applied Random Matrix Theory (RMT) in the pursuit of an estimator for the out-of-sample Sharpe ratio in the high dimensional setting using only in-sample data. Primarily, they focus on the case where the mean is known, and the covariance matrix must be estimated, but do provide a result for when the mean is also unknown. Specifically, they derive:

$$\widehat{SR}_{\text{OOS}}(\mathbf{Q}) = \underbrace{\left(1 - \frac{1}{T_{\text{IS}}} \operatorname{tr} \left(\widehat{\boldsymbol{\Sigma}} \left(\widehat{\boldsymbol{\Sigma}} + \mathbf{Q} \right)^{-1} \right) \right)}_{\text{Correction Factor}} \cdot \underbrace{\frac{\boldsymbol{\mu}^\top \left(\widehat{\boldsymbol{\Sigma}} + \mathbf{Q} \right)^{-1} \boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}^\top \left(\widehat{\boldsymbol{\Sigma}} + \mathbf{Q} \right)^{-1} \widehat{\boldsymbol{\Sigma}} \left(\widehat{\boldsymbol{\Sigma}} + \mathbf{Q} \right)^{-1} \boldsymbol{\mu}}}}_{\text{In-Sample SR Estimate}},$$

where T_{IS} is the number of observations, $\boldsymbol{\mu}$ is the vector of (known) means, $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix and $\mathbf{Q}$ is the regularisation matrix. The key benefit of their result is that they allow for a regularisation matrix $\mathbf{Q}$ to be added to the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$, and show how their theoretical results can be used to find the optimal $\mathbf{Q}^*$ which maximises the out-of-sample Sharpe ratio. This is related to work on using RMT in Machine Learning (ML) as described by Couillet and Liao [CL23], such as setting the optimal ridge regression parameter for minimal

⁵ Or, $\mathbf{H} = \mathbf{S} \left(\mathbf{S}^\top \mathbf{S} + \sigma_{K+1}^2 \mathbf{I}_{T_{\text{IS}}} \right)^{-1} \mathbf{S}^\top$, where σ_{K+1}^2 is the smallest singular value of the concatenated matrix $[\mathbf{r}, \mathbf{S}]$ when TLS is used rather than OLS.

out-of-sample MSE, but this is not highlighted by [MCW25].

Chapter 2 complements this burgeoning literature on corrections for the Sharpe ratio that focus on strategy complexity. We study the in-sample and out-of-sample Sharpe ratio when the returns are predicted using a linear model, and as in the above literature find that increasing complexity (in the form of either additional signals or assets) leads to inflated in-sample Sharpe ratios (and therefore, a greater reduction in the out-of-sample period), but this is mitigated by increased sample sizes (longer backtests) and stronger signals. A fuller discussion of the chapter’s contribution and positioning is provided in the chapter introduction and abstract.

1.1.3 Cross-Sectional Methods

The other avenue explored in the literature related to this thesis involves methods to reduce the impact of estimation error. These approaches rely primarily on shrinkage and learning from the cross-section of observed performance, typically via Bayesian methods. The fund literature in particular has a long history of applying these approaches to improve estimation of fund alphas.

Baks, Metrick, and Wachter [BMW01] utilise a Bayesian model to apply prior beliefs in the estimation of active mutual fund manager alpha. They seek to divide managers into “skilled” (managers who deliver positive alpha) and “unskilled” (managers who do not), and use this to estimate the posterior alpha for each manager,

$$\mathbb{E}[\alpha] = \mathbb{P}(Z = 1) \mathbb{E}[\alpha \mid Z = 1] + (1 - \mathbb{P}(Z = 1)) \mathbb{E}[\alpha \mid Z = 0],$$

where Z denotes the condition that a manager is skilled (1) or unskilled (0) and α is the returns $\mathbf{r}$ of the manager in excess of some factors $\mathbf{F}$. Their approach requires some subjective judgement with regard to the proportion of skilled and unskilled managers, and the probability that a manager returns significant alpha $> 25\text{bps/month}$. They discuss that their approach shrinks alphas towards the prior mean, effectively controlling for “luck” among fund managers.

Shortly following this, Pástor and Stambaugh [PS02] introduce the concept of “Seemingly Unrelated Assets” (SURA) as a methodology to improve estimation of mutual fund alphas, particularly in the face of short histories, by using returns on “non-benchmark assets” with longer histories (for example, factors and industry portfolios). They propose an extended regression against both benchmark assets (Fama–French factors [FF93]) and the non-benchmark assets:

$$r_{i,t} = \delta_i + \boldsymbol{\beta}_{i,N}^\top \mathbf{r}_{N,t} + \boldsymbol{\beta}_{i,B}^\top \mathbf{r}_{B,t} + u_{i,t},$$

and separately regress the non-benchmark assets against the benchmark assets:

$$\mathbf{r}_{N,t} = \alpha_N + \boldsymbol{\beta}_B \mathbf{r}_{B,t} + \epsilon_{N,t}.$$

They then use this to decompose the fund’s alpha,

$$\alpha_i = \delta_i + \boldsymbol{\beta}_{i,N}^\top \alpha_N,$$

into a purely idiosyncratic part δ_i and a part explained by the non-benchmark assets own alpha, $\boldsymbol{\beta}_{i,N}^\top \alpha_N$. Similar to [BMW01], this framework incorporates subjective prior beliefs. They assume the following prior for the alpha of the non-benchmark assets,

$$\alpha_N | \boldsymbol{\Sigma} \sim \mathcal{N} \left(0, \frac{\sigma_N^2}{s^2} \boldsymbol{\Sigma} \right), \quad \boldsymbol{\Sigma}^{-1} \sim W \left((2s^2 \mathbf{I}_m)^{-1}, m + 3 \right),$$

where s^2 is the average variance of the residuals $\epsilon_{N,t}$, m is the number of non-benchmark assets and σ_N^2 is a parameter which controls the “mispricing uncertainty”, i.e. the extent to which the Fama–French factors are believed to price the additional assets (e.g. whether Momentum’s alpha is truly zero). When the factors are assumed to price these assets perfectly ($\sigma_N = 0$), the estimator simplifies to δ , which they show is much more precise than a naive OLS estimate for α .

Kosowski, Naik, and Teo [KNT07] build on this work, applying the SURA approach to hedge fund data, which also suffers from the short history problem. They

show that this yields improved estimates of hedge fund alphas, resulting in a 5.5% increase in the alpha of the spread between the top and bottom decile hedge funds. Although we do not apply them in this thesis, bootstrap approaches are also popular in the finance literature for resolving issues with performing inference. In addition to the Bayesian analysis, the authors conduct a residual-only resampling bootstrap to answer the question: is hedge fund performance explained purely by luck? They find that no, it is not, validating the view that a subset of managers are skilled, and we should try to identify them (as we do in Section 3.3).

Chen, Cliff, and Zhao [CCZ17] (CCZ) use an Expectation-Maximization (E-M) approach to separate funds into different skill groups. They model the cross-sectional distribution of fund alphas as a mixture of J (unknown) Normal distributions, and treat the alpha α_i as a noisy observation with fund-specific standard error s_i , so that conditional on the fund being a member of group $z_i = j$,

$$\alpha_i \mid z_i = j \sim \mathcal{N}(\mu_j, s_i^2 + \sigma_j^2).$$

They select the number of groups J via BIC and compute the probability that fund i belongs to group j by

$$p_{i,j} = \frac{\pi_j f_j(\alpha_i)}{\sum_{k=1}^J \pi_k f_k(\alpha_i)},$$

where π_j denotes the proportion of funds in group j , and f_j is the fitted density of fund alphas for group j . For funds with larger standard error s_i , the probability that fund i belongs to group j is driven primarily by π_j (the overall number of funds estimated to exist in that group), while for small s_i , the funds own estimated α_i more heavily influences its group allocation.

Harvey and Liu [HL18] extend this framework by proposing a “Noise Reduced Alpha” (NRA) model which uses a structural random effects framework to jointly estimate the overall distribution of fund alphas (again a Gaussian mixture), and the individual fund alphas simultaneously (rather than the two-step procedure con-

ducted by CCZ where alphas are first estimated by OLS, and then fed through the E-M approach). They then use the model to estimate the posterior alpha for each fund,

$$\mathbb{E}(\alpha_i \mid \mathbf{R}_{\text{IS}}) = \sum_{j=1}^J \mathbb{E}(\alpha_i \mid z_i = j, \mathbf{r}_{\text{IS},i}) p_{i,j}$$

where $\mathbf{R}_{\text{IS}} \in \mathbb{R}^{N \times T_{\text{IS}}}$ is the returns for all funds in the in-sample period, $\mathbf{r}_{\text{IS},i}$ is the returns for fund i , $z_i = j$ is the condition that fund i is a member of group j , and $p_{i,j}$ is the probability of this condition being true. When comparing this forecast versus the observed alpha in the out-of-sample period, they find the error of their estimate is 2.7% vs 3.2% for CCZ (returns per annum).

Most recently, Jensen, Kelly, and Pedersen [JKP23] have applied Bayesian approaches to the evaluation of asset pricing factors. They consider a dataset of asset pricing signals across multiple regions, considering each signal-region pair as an individual factor, and are interested in the alpha of each factor relative to the market:

$$f_{t,i} = \alpha_i + \beta_i r_{t,m} + \epsilon_{t,i}.$$

They decompose the alpha of each of these factors into:

$$\alpha_i = \alpha_0 + \alpha_{c,j} + \alpha_{s,k} + \alpha_{e,l},$$

where α_0 is a common component across all factors (which they fix to zero), $\alpha_{c,j} \sim \mathcal{N}(0, \sigma_c^2)$ is the contribution from the factor style cluster the factor belongs to, $\alpha_{s,k} \sim \mathcal{N}(0, \sigma_s^2)$ is a signal specific contribution shared across regions, and $\alpha_{e,l} \sim \mathcal{N}(0, \sigma_e^2)$ is a factor unique idiosyncratic element. Under this model, the posterior true alpha for each factor is given by

$$\mathbb{E}(\boldsymbol{\alpha} \mid \hat{\boldsymbol{\alpha}}) = \left(\boldsymbol{\Omega} + T\boldsymbol{\Sigma}^{-1} \right)^{-1} \left(\boldsymbol{\Omega}^{-1} \mathbf{1}_K \alpha_0 + T\boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\alpha}} \right)$$

where T is the number of observed periods, $\hat{\boldsymbol{\alpha}}$ is the observed alpha estimate, $\boldsymbol{\Sigma}$ is

the covariance of the residuals $\epsilon_{i,t}$ across factors, and

$$\mathbf{\Omega} = \mathbf{M}\mathbf{M}^\top \sigma_c^2 + \mathbf{Z}\mathbf{Z}^\top \sigma_s^2 + \mathbf{I}_K \sigma_e^2 \in \mathbb{R}^{K \times K}$$

is the prior covariance matrix of the factors, with $\mathbf{M} \in \mathbb{R}^{K \times J}$ denoting the matrix of cluster memberships, and $\mathbf{Z} \in \mathbb{R}^{K \times S}$ indicating the signal that each factor is based on. The authors conclude that, under this approach, 82.4% of factors exhibit a statistically significant α (defined as a posterior 95% credible interval excluding zero). They argue that the large number of published factors allows for better estimation of each factor’s α , countering the view that such abundance should be a cause for concern regarding multiple testing.

This brings us to our contribution. Chapter 3 similarly leans on the observed cross-section of in-sample and out-of-sample performance to not only improve point estimates for Sharpe ratios through Bayesian shrinkage, but also to capture and correct selection bias and alpha decay. We propose a model for the joint distribution of historical and future performance, and show that fitting this model to the in- and out-of-sample data is essential, with in-sample data alone being insufficient to disentangle the twin drivers of performance deterioration. By applying this framework to a comprehensive dataset of hedge fund returns, we demonstrate that our “bias-aware” estimates yield significantly more accurate out-of-sample predictions and enable robust allocation decisions in regimes where naive estimators fail.

1.2 Thesis Outline

The remainder of this thesis is structured as follows.

Chapter 2 establishes the theoretical link between model complexity and in-sample performance inflation. By extending the framework of Kan, Wang, and Zheng [KWZ24], it provides insights into why highly complex strategies disappoint out-of-sample. This chapter is based on the preprint “In-Sample and Out-of-Sample Sharpe Ratios

for Linear Predictive Models”, written in collaboration with my supervisors Antoine Jacquier and Johannes Muhle-Karbe.

Chapter 3 addresses the in-sample/out-of-sample problem from a low information perspective when we do not know how the backtest was produced. It additionally corrects for both selection bias (overfitting) and alpha decay simultaneously. This chapter is based on the preprint “Bayes Is All You Need: Better Estimates of True Sharpe Ratios” [Mul+26], written in collaboration with my supervisors Antoine Jacquier and Johannes Muhle-Karbe, and Juha Joenväärä of Aalto University.

Chapter 4 concludes with some brief recommendations for practical steps to avoid these issues.

Throughout, significant content is delegated to chapter appendices in the interest of narrative flow. This includes theoretical details, extensions, additional studies and proofs.

Chapter Appendix

1.A Application of Harvey, Liu, and Zhu to the Open Source Asset Pricing Data

Chen and Zimmermann provide a useful dataset for investigating asset pricing studies [CZ22] containing reproduced returns for 212 predictors (of which we select 203 after dropping factors which have time series that end prior to 2023). In Chapter 3 we use this dataset to compare the results from [CZ20] against our approach (Appendix 3.E). As part of our preprocessing, we take a slightly different approach to construct the long-short portfolios. Specifically, we combine the top and bottom quantiles in a Markowitz [Mar52] fashion according to their in-sample mean and covariance. This significantly increases both the in-sample and out-of-sample performance for most predictors.

We investigate here how this affects the application of HLZ’s multiple testing approach. In fig. 1.A.1 we see that the in-sample t -statistics for the long short portfolios for most predictors are very large, with the modal value being over 5. The out-of-sample t -statistics are not so impressive, but most are still greater than 2.

The Holm and BHY predictors depend on the distribution of observed statistics. The natural question therefore is what is the multiple testing adjusted t -stat hurdle if looking at the in-sample data alone? This redoes the analysis conducted in [HLZ16, Fig. 3]. Specifically, in fig. 1.A.2 we plot the minimum acceptable t -statistic for a new predictor to be significant, given the existing universe of published predictors on that date.

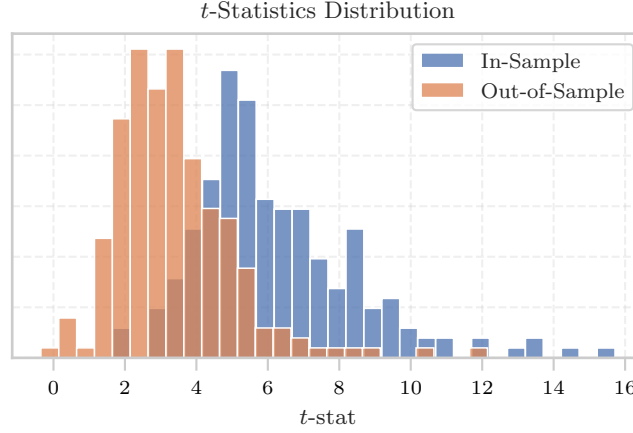


Figure 1.A.1: Distribution of in- and out-of-sample t -statistics from the Open-Source Asset Pricing Dataset with our pre-processing approach in Appendix 3.E.

The results differ significantly from the original analysis.⁶ Although the Bonferroni hurdle increases steadily with the number of factors, the Holm hurdle flatlines for most of the window. This is due to the majority of the portfolios in our dataset being highly significant, with an average t -statistic of 6.2 and a p -value of 0.001 in-sample. Similarly, the BHY correction shows only a mild increase, driven by the increasing count of published factors. This implies that little to no multiple testing adjustment is required on this dataset.

This is at odds with the observed degradation in out-of-sample performance. In fig. 1.A.3 we see that applying a Bonferroni correction in the style of [HL15]⁷ yields a 21% reduction to the median t -statistic, but the out-of-sample t -statistic is only 55% of the in-sample value, leaving 25% of the drop unexplained.

Although, this is predicated on the assumption that we have observed all tests (which is, of course, untrue). If we try to find the number of tests necessary for the in-sample t -stat to match its out-of-sample equivalent, we find that the median

⁶ To match [HLZ16] we have plotted the Market Beta (MRT), Earnings-Price (EP), Size (SMB), Book-to-Market (HML), Momentum (MOM), Liquidity (LIQ) and Idiosyncratic Volatility (IVOL) t -stats, but most are off the chart with our construction. Even with the most restrictive Bonferroni test, 92% of predictors are still significant. (98.5% and 96% for Holm and BHY respectively.)

⁷ Convert the t -statistic to a p -value, multiply by the number of tests (factors, $M = 203$) and then invert back to a t -statistic.

1.A Application of Harvey, Liu, and Zhu to the Open Source Asset Pricing Data

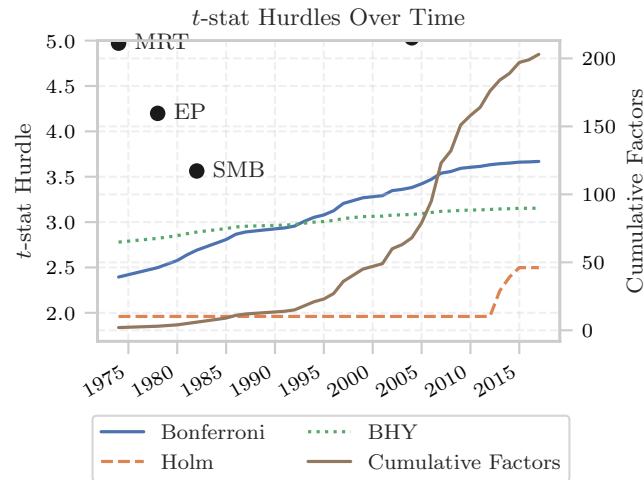


Figure 1.A.2: Replication of Figure 3 from Harvey, Liu, and Zhu using the Open-Source Asset Pricing Dataset and our pre-processing approach in Appendix 3.E.

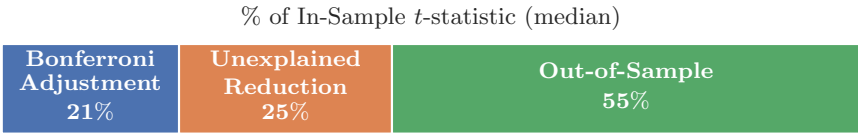


Figure 1.A.3: Amount of explained reduction in performance between the in-sample and out-of-sample t -statistic.

number of tests required would be 193, 095, which seems quite infeasible, indicating other approaches must be taken to explain the difference between in- and out-of-sample performance.

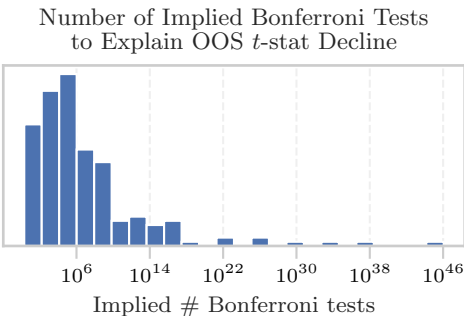


Figure 1.A.4: Number of implied tests required for the Bonferroni correction to adjust the in-sample t -stat to match the out-of-sample t -stat.

The final figure we wish to report is the number of significant predictors based solely on their out-of-sample performance. Although yes, the median out-of-sample Sharpe ratio is only 55% of its in-sample equivalent, we also see that most predictors still have t -stats > 2 out-of-sample. In fig. 1.A.5 we plot the empirical distribution of out-of-sample p -values. We see that for the adjusted p -values, more than 80% are significant at the 5% level, despite the average out-of-sample period being significantly shorter than the average in-sample period (which of course affects the t -statistic).

If we take the observed out-of-sample Sharpe ratio, and use the maximum of the in- or out-of-sample period lengths to turn this into a Sharpe ratio (i.e. if the predictor keeps performing as it has in the out-of-sample period for as long as it did in the in-sample period), then we see that this increases to more than 90% of predictors being significant at the 5% level, with almost 80% being significant at the same level after a BHY adjustment. While this isn't the focus of this thesis, this result does seem positive for the published asset pricing research.

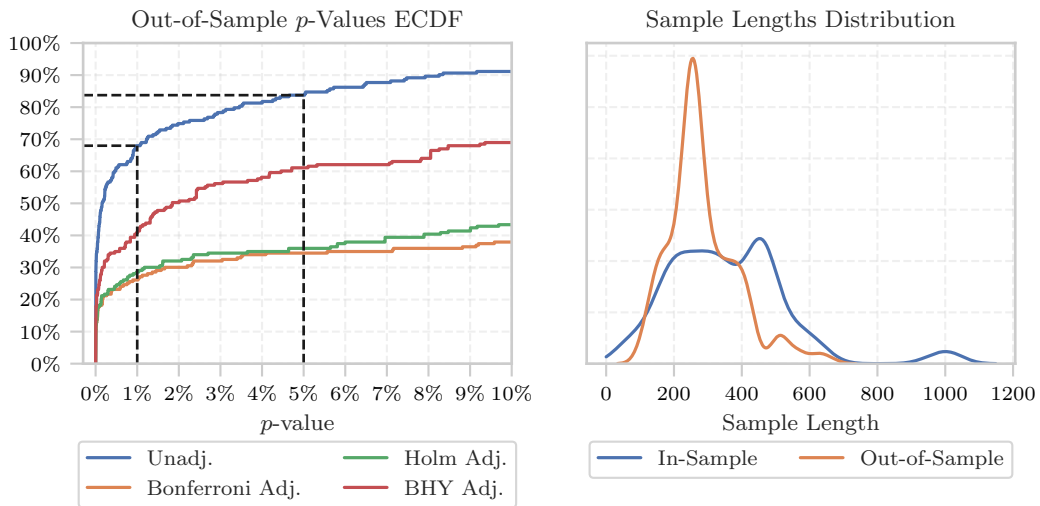


Figure 1.A.5: Distribution of out-of-sample p -values under various corrections, and the accompanying distribution of in- and out-of-sample period lengths.

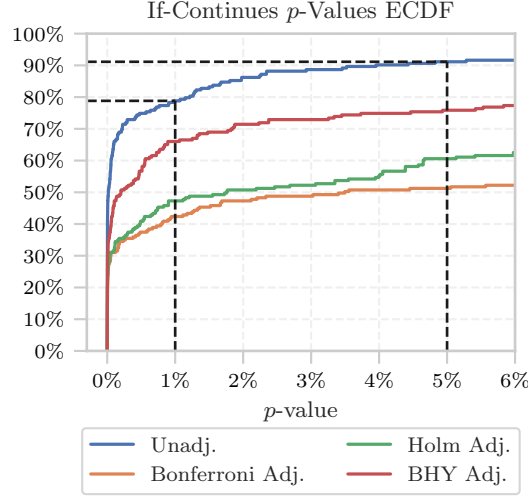


Figure 1.A.6: Distribution of p -values using the observed out-of-sample Sharpe ratios and the maximum of the in- and out-of-sample period lengths to convert this to a t -statistic, and then p -value.

1.B Sharpe Ratios Are Angles

Wood and Saville [WS02; WS05; SW91] make the connection between statistics and geometry. Specifically, the t -statistic for a sample can be viewed as an angle. Because the Sharpe ratio is just a scaled t -statistic, this holds true for the Sharpe ratio too. For visualisation purposes, imagine having just two data points, r_1 and r_2 which we stack into the vector $\mathbf{r} = [r_1 \ r_2]^\top$. Let $\hat{\mathbf{1}}$ be the unit vector, $\hat{\mathbf{1}} = \mathbf{1}/\sqrt{T}$, where $\mathbf{1} = (1, \dots, 1)^\top$. Defining ϑ as the angle between the vector $\mathbf{r}$ and the unit vector $\hat{\mathbf{1}}$,

$$\vartheta = \arccos \left(\frac{\hat{\mathbf{1}}^\top \mathbf{r}}{\|\mathbf{r}\|} \right),$$

we have that the Sharpe ratio can be expressed as

$$\text{SR} = \sqrt{\frac{T-1}{T}} \cot(\vartheta).$$

It's unclear if this interpretation is useful, but it is aesthetically pleasing, see fig. 1.B.1.

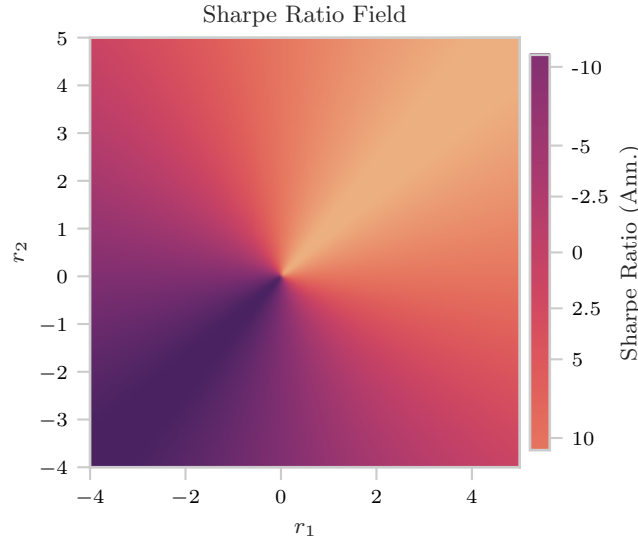


Figure 1.B.1: The Sharpe ratio for two points.

1.C Overview of Implementability, Transaction Costs & Capacity Literature

In addition to overfitting and selection bias, we identify three more key sources of reduced performance when moving from backtesting to live trading:

- **Implementability** - If the backtest isn't conducted carefully, the tested trading strategy may perform trades that aren't feasible in reality. This can come in many forms, for example it could be that the strategy trades heavily in an asset which has limited liquidity, or the signal relies on data which has been backfilled and was not actually available at the time it is timestamped for.
- **Transaction Costs** - In a similar vein, if transaction costs are not carefully modelled and considered when evaluating trading strategies, then the strategy is likely to perform extremely poorly when moving to live trading. In many cases this is recoverable, by slowing trading [GP13] we can trade a small amount of gross Sharpe ratio for significantly improved net Sharpe ratio, which is all we care about anyway.

- **Capacity** - Finally, and strongly related to the previous two, it is possible you do everything correctly on the previous two items when initially testing your strategy, but will later find out that as you attempt to scale the strategy to larger positions you are constrained by the available liquidity.

Primarily, these are mechanical and operational issues that can be mitigated in practice by careful point-in-time [Bre15] backtesting and modelling of realistic trading and associated costs [Web23]. With that said, there are a few papers on these topics that we think should be highlighted.

First, Hou, Xue, and Zhang [HXZ20] study the replicability of anomalies after removal of microcaps (which are infeasible to trade) from the tradeable universe. They find that 65% of anomalies fail to clear the 5% significance hurdle. While this has been critiqued by [CZ22] as being such a high percentage due to the set of anomalies the original authors chose to include in their analysis, [CZ22] also note that: “The typical mean return drops from around 60 bps per month to about 40 bps per month regardless of whether the adjustment is an NYSE only screen, a market equity screen, or the enforcement of value-weighting”.

Relatedly, Chen and Velikov find that after accounting for costs, post-publication decay, and the structural change in the markets in the last twenty years, the average published anomaly returns 4 basis points per month [CV23a].⁸ Conversely, Frazzini, Israel, and Moskowitz [FIM12] find that standard academic models significantly overestimate real-world trading fractions. They suggest that the total break-even capacity (the point at which trading costs fully erode profits) of Size, Value and Momentum exceeds \$2.7 trillion once sufficient cost optimisations have been implemented, while for comparison the naive variants would have a capacity of “only” \$435 billion.

⁸ We haven’t reproduced this exact result, but in our analysis of the Open Source Asset pricing dataset used by [CV23a] we found that performance could be significantly improved by constructing an improved long-short portfolio of the top bottom quantiles, so this figure could potentially be increased with some improved portfolio construction (à la Gârleanu and Pedersen [GP13]).

However, Novy-Marx and Velikov [NV16] do challenge this analysis, and find that while low turnover strategies do have large capacities (\$169 billion and \$50 billion for Size and Value respectively), mid turnover strategies like Momentum are much more constrained, around \$5 billion. Perhaps counterintuitively, they find that by restricting the tradeable universe to only large capitalisation stocks, the capacity of strategies actually increases. This is due to the strategy being forced to select more liquid assets (rather than loading on microcaps which have theoretically stronger signals), and thus having higher capacity. This aligns neatly with the findings of Cartea et al. [Car+25] who argue that liquidity constraints (the maximum volume available to an investor, regardless of cost) restrict trading due to the concentration of signal in stocks with low Daily Average Value (ADV). These two findings hint at the need to consider both liquidity and return predictability when performing portfolio construction.

From a theoretical perspective, Bonelli et al. [Bon+19] prove the link between signal dynamics and scale, showing the capacity is driven heavily by signal persistence, and therefore slower strategies (like Value) scale much better than fast ones (like STR).

Finally, Arnott et al. [Arn+19] further warn against the reduction in alpha out-of-sample due to crowding and failed diversification: as is often said “all correlations go to one in a crisis”.⁹ One remedy for this is often geographic diversification, however [Bec+16] show that not all factors work equally well in all markets, with Momentum, for example, not performing well in Japan, and Size, Illiquidity and Quality not replicating outside the US.

In conclusion, while the effects we study in this thesis (primarily overfitting, selection bias and alpha decay) are a vital part of the in- and out-of-sample puzzle, there are many more components which need to be executed correctly to yield

⁹ The earliest publicly available source I can find for this saying is a prophetic memo from Howard Marks to Oaktree Capital Management clients in March 2006 [Mar06], but he also notes it as a well known saying at that point.

profitable out-of-sample performance, and all must be treated with equal care and attention.

2 | In-Sample and Out-of-Sample Sharpe Ratios for Linear Predictive Models

2.1	Introduction	24
2.2	Model Setup	28
2.3	Main Results	33
2.4	Comparison with the Literature	46
2.5	Predicting Commodity Futures	48
2.6	Empirical Study: Goyal, Welch, and Zafirov Dataset	53
2.7	Conclusion	57
	Appendices	59
2.A	Derivation of Equation (2.2.2)	59
2.B	General Case Result of Proposition 2.3.1	60
2.C	Proof of Proposition 2.B.1	62
2.D	Monte Carlo Details	78
2.E	Additional Values for Section 2.5 and Section 2.6	79

Chapter Abstract

To begin our study of in- and out-of-sample Sharpe ratios, we examine how overfitting in a linear model inflates the in-sample performance estimate relative to the later out-of-sample observed performance. More specifically, we compute the in- and out-of-sample means and variances of the corresponding PnLs and use these to derive a closed-form approximation for the corresponding Sharpe ratios. We find that the out-of-sample replication ratio diminishes for complex strategies with many as-

sets and based on many weak rather than a few strong trading signals, and increases when more training data is used. The substantial quantitative importance of these effects is illustrated with a simulation case study for commodity futures following the methodology of Gârleanu and Pedersen [GP13], and an empirical case study using the dataset compiled by Goyal, Welch, and Zafirov [GWZ24]. This chapter is based on the pre-print paper “In-Sample and Out-of-Sample Sharpe Ratios for Linear Predictive Models” [MJM25].

2.1 Introduction

A standard approach to building systematic trading strategies is to first fit some predictive model to historical training data. For example, Gârleanu and Pedersen [GP13] regress price changes of commodity futures against past returns with various lookback periods. Using the forecasting model, one can then form a trading strategy, e.g., by combining forecasts for several assets using mean-variance analysis, and “backtest” its performance on historical data. The strategy’s “out-of-sample” performance on new testing data or in live trading will typically be worse than its “in-sample” performance on the training data for a number of reasons. But how much worse?

A number of empirical studies find that the out-of-sample performance of ETFs, anomalies and various other strategies tends to be significantly lower than the corresponding in-sample performance, with estimates ranging widely depending on the dataset [BL15; Wie+16; MP16; SLP17; FRT22].¹ One commonly used rule of thumb is to haircut the in-sample performance by 50% [HL15]. While such rules of thumb are useful, we seek a more systematic approach, enabling us to estimate the out-of-sample performance of different trading strategies based on their characteristics, and without having to resort to a new large-scale simulation study each time the

¹ More generally, as in many other quantitative disciplines [Ioa05], there is a controversial ongoing debate whether there is a “replication crisis” in financial economics [HLZ16; JKP23].

strategy is modified.

To this end, we focus on one crucial issue in this context: *overfitting*.² Kan, Wang, and Zheng [KWZ24] obtain analytical results in this spirit for the simplest setting where investment opportunities are constant, in that expected asset returns and covariances do not change over time. When these model parameters are estimated from a finite training dataset, Kan, Wang, and Zheng derive explicit formulas for the “replication ratio”, that is, the fraction of the in-sample Sharpe ratio that can be recovered out-of-sample.

In the present chapter, we extend this analysis to a setting closer to that of Gârleanu and Pedersen and many related papers, where trading strategies exploit several trading signals per asset and also take into account that these signals are not fully persistent but fluctuate over time. More specifically, we assume that expected returns are linear functions of iid trading signals drawn from a multivariate normal distribution. When the variance of this distribution tends to zero, the constant-parameter model of Kan, Wang, and Zheng is recovered as a special case. Conversely, our iid expected returns correspond to the limiting case of short-lived signals in the setting of Gârleanu and Pedersen.

The relationship between signals and expected returns can be estimated empirically by linear regression. For portfolios corresponding to this plug-in estimate, we compute the (unconditional) in- and out-of-sample means and variances of the P&L and derive closed-form approximations for the replication ratio, i.e., the fraction of the in-sample Sharpe ratio that can be achieved out-of-sample.

We find that as in Kan, Wang, and Zheng [KWZ24] the replication ratio is increasing in the size of the training dataset and the magnitude of the “true” Sharpe ratio of the strategy (i.e., the exact unconditional Sharpe ratio computed in the model rather than based on estimated parameters). Indeed, when the true Sharpe

² Another key driver of lower out-of-sample performance is *multiple testing* studied by [BL14; HL15]. While multiple testing may lead to overfitting, we focus on overfitting from a single test perspective.

ratio is low the strategies are more prone to overfitting, and small increases in the true Sharpe ratio yield significant improvements in the replication ratio. Once the true Sharpe ratio is sufficiently large, the model is more robust to overfitting and the replication ratio slowly converges towards 100%.

Conversely, the replication ratio is decreasing in the number of model parameters. In Kan, Wang, and Zheng [KWZ24] this quantity only depends on the number of assets; in our setting, it is determined both by the number of assets and by the trading signals per asset. The takeaway message is that complex strategies are more prone to overfitting, which needs to be balanced with sufficiently long backtesting periods.

The practical relevance of these theoretical results is in turn explored in a simulation study that follows the implementation in [GP13]. To wit, we fit a linear predictive model to a set of commodity futures forecast using momentum-style signals, and in turn use this as a simulation engine. We test our approach on simulations with both Gaussian and fat-tailed returns, and for both iid and autoregressive signals. Our analytical approximations provide very good predictions in the Gaussian iid setting where they were derived, but also perform encouragingly well with autoregressive signals driven by heavy-tailed noise, confirming the practicality of our results in realistic settings.

We observe that a ten-year backtest yields an expected replication ratio of just 30% (Figure 2.1.1). This steep haircut arises from the large number of parameters in the model; it can be reduced by imposing constraints on the model, such as regularisation or a panel regression as in the empirical section of [GP13].

To further confirm that our results are also relevant for real data, we test the method on a comprehensive dataset compiled by Goyal, Welch, and Zafirov [GWZ24], which provides over 100 years of monthly signals which have been published in the literature for predicting the equity premium. This allows us to deploy the methodology developed in the present chapter to subsets of the predictors and, in doing

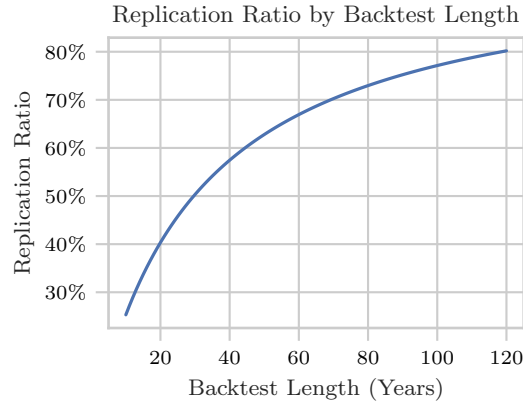


Figure 2.1.1: The replication ratio for the case study in Section 2.5 with varying backtest length.

so, control both complexity of the composite strategy and backtest length. Encouragingly, we recover the key relationships from our model in the data, in that more signals lead to more overfitting, whereas longer backtests and higher strength signals mitigate the issue.

In summary, our results suggest the following broad practical conclusions: be wary of low Sharpe ratio strategies, keep your models as simple as possible, and use the longest sensible backtest period available. While much of this advice aligns with practitioners' intuition, our results allow one to pinpoint how much of a haircut one should expect out-of-sample depending on the characteristics of the backtest, rather than broad rules of thumb. Additionally, estimating the out-of-sample haircut by simulation is often numerically slow. In contrast, our closed-form expressions permit accelerating this time-sensitive process by instantly computing the expected out-of-sample haircuts.

The remainder of this chapter is organised as follows. Section 2.2 sets up the model and differentiates between the true and the sample Sharpe ratios. Section 2.3 presents the main results and their interpretation. Section 2.4 compares our results to the existing literature on overfitting. Section 2.5 applies our results in a simulation case study using commodity futures, Section 2.6 applies the method to the Goyal,

Welch, and Zafirov dataset, and Section 2.7 concludes. Appendices 2.A and 2.C provide derivations of results used in the main text, Appendix 2.B provides the general case of the main result, and Appendix 2.D provides details of the Monte Carlo analysis we use to check the approximations used.

2.2 Model Setup

We consider a researcher who wants to fit a linear model to forecast asset returns and to trade based on these predictions. The researcher has access to some historical data observed at discrete times $\mathcal{T}_{\text{IS}} = \{0, \dots, T_{\text{IS}}\}$. We denote the returns per time step by $\mathbf{r}_{t+1} \in \mathbb{R}^N$ and the predictive signals by $\mathbf{s}_t \in \mathbb{R}^K$ for each $t \in \mathcal{T}_{\text{IS}}$. The researcher assumes that the relationship between the signals and the returns follows a linear model with parameter $\boldsymbol{\beta} \in \mathbb{R}^{N \times K}$:

$$\mathbf{r}_{t+1} = \boldsymbol{\beta} \mathbf{s}_t + \boldsymbol{\epsilon}_{t+1},$$

where $\boldsymbol{\epsilon}_{t+1} \in \mathbb{R}^N$ denotes the unexplained residual returns. We assume that the sequences $\{\mathbf{s}_t\}$ and $\{\boldsymbol{\epsilon}_{t+1}\}$ are independent of each other, and iid across time with distributions $\mathfrak{s}$ and $\mathfrak{e}$ respectively, which are assumed to be multivariate Gaussian:³

$$\mathfrak{s} \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \quad \text{and} \quad \mathfrak{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_e).$$

We note here that trading signals are typically assumed to be autoregressive as in [GP13]. However, imposing the iid assumption as in [DL18] makes the calculations below substantially more tractable, and we demonstrate using simulation experiments in Section 2.5 that the resulting formulas also provide very good approximations for autoregressive signals fitted using the methodology from the empirical part of [GP13].

³ We focus on the Gaussian case as it simplifies the exposition in many places. One can compute the true Sharpe ratio when the signals/noise are non-Gaussian, but we omit it here as the expression is verbose and broadly similar.

We impose without loss of generality that the signals are centred, except for a possible intercept term, i.e., $\boldsymbol{\mu}_s = (\mu_{s,1}, 0, \dots)$ where $\mu_{s,1} = 1$ if the model includes a constant (intercept) term, and $\mu_{s,1} = 0$ otherwise; removing the intercept simplifies many of the expressions, but we leave the expressions general as including a drift term may be desirable for some applications such as long-only equity portfolios.

Assuming $\boldsymbol{\beta}$ is known, the researcher can use the signal $\mathbf{s}_t$ at time t to accurately predict the expected returns $\mathbb{E}_t[\mathbf{r}_{t+1}] = \boldsymbol{\beta}\mathbf{s}_t$ conditional on the information at time t , and in turn trade a portfolio of the form $\mathbf{w}_t = \mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t$ where $\mathbf{Z} \in \mathbb{R}^{N \times N}$ is some symmetric weight matrix. For example, the standard (one-period) Markowitz portfolio is given by $\mathbf{w}_t = \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t$ where $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\mu}_t$ denote the covariance matrix and expected returns of the traded assets conditional on the information available at time t . In the present context, these are given by

$$\mathbb{E}_t[\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1}] = \boldsymbol{\beta}\mathbf{s}_t, \quad \text{Cov}_t[\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1}] = \boldsymbol{\Sigma}_\epsilon,$$

and thus the one-step Markowitz optimal portfolio is of the form $\mathbf{w}_t = \mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t$ with $\mathbf{Z} = \boldsymbol{\Sigma}_\epsilon^{-1}$. Somewhat less common but still relevant choices of $\mathbf{Z}$ include $\text{diag}(\boldsymbol{\Sigma}_\epsilon)^{-1}$, corresponding to an inverse variance allocation by asset, or $\mathbf{I}_N$, corresponding to a risk indifferent portfolio. The researcher can then test this portfolio against the returns to compute the Profit & Loss per time step,

$$\text{PnL}_t = \mathbf{w}_t^\top \mathbf{r}_{t+1}.$$

2.2.1 The True Sharpe Ratio

The Sharpe ratio typically refers to a summary statistic that compares the sample mean of a time series of returns to their sample standard deviation. In the model from the previous section, we can compare this to its “true” equivalent, defined as the ratio of the expected return and standard deviation of the daily P&L, averaged

over the distributions of the signal and noise terms:

$$\text{SR} := \frac{\mathbb{E}[\text{PnL}_t]}{\sqrt{\mathbb{V}[\text{PnL}_t]}} = \frac{\mathbb{E}[\mathbf{w}_t^\top \mathbf{r}_{t+1}]}{\sqrt{\mathbb{V}[\mathbf{w}_t^\top \mathbf{r}_{t+1}]}} = \frac{\mathbb{E}[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top (\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1})]}{\sqrt{\mathbb{V}[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top (\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1})]}}. \quad (2.2.1)$$

Using identities for quadratic expectations of Gaussian vectors, this can be computed as

$$\text{SR} = \frac{\text{tr}(\mathbf{G}\boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\mu}_s}{\sqrt{2 \text{tr}((\mathbf{G}\boldsymbol{\Sigma}_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\Sigma}_s \mathbf{G}\boldsymbol{\mu}_s + \text{tr}(\mathbf{F}\boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{F}\boldsymbol{\mu}_s}}, \quad (2.2.2)$$

where

$$\mathbf{F} := \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta} \quad \text{and} \quad \mathbf{G} := \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}.$$

See Appendix 2.A for the derivation of this expression. To provide some intuition for this result, suppose for simplicity that the signals include no intercept ($\boldsymbol{\mu}_s = 0$). Then, (2.2.2) simplifies to

$$\text{SR} = \frac{\text{tr}(\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s)}{\sqrt{2 \text{tr}((\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s)^2) + \text{tr}(\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s)}}.$$

In this form, clearly the expected return (numerator) increases with the Frobenius norm $\|\boldsymbol{\beta}\|_F$ (which measures the “predictability” of future asset returns), however the variance (denominator) also increases. For a single asset or signal, the Sharpe ratio is monotonically increasing in $\|\boldsymbol{\beta}\|_F$, but with multiple assets or signals the relationship is no longer obvious. We also see that the Sharpe ratio increases as the signal-to-noise ratio rises with smaller $\|\boldsymbol{\Sigma}_\epsilon\|_F$. Similarly to $\boldsymbol{\beta}$, the relationship with the signal variance $\boldsymbol{\Sigma}_s$ is generally not monotone, but for parameters similar to our simulation case study in Section 2.5 the Sharpe ratio typically increases when $\|\boldsymbol{\Sigma}_s\|_F$ increases relative to $\|\boldsymbol{\Sigma}_\epsilon\|_F$, again indicating an improvement in the signal-to-noise ratio of the model.

2.2.2 The In-Sample Sharpe Ratio

When a given predictive model is deployed in practice, its parameters β must be estimated. In the linear model we consider here, this can be done using ordinary least-squares (OLS) by stacking the returns across all time steps into the matrix $\mathbf{R} \in \mathbb{R}^{N \times T_{\text{Is}}}$ and the signals across all time steps into the matrix $\mathbf{S} \in \mathbb{R}^{K \times T_{\text{Is}}}$ and computing the OLS estimator

$$\hat{\beta} = \mathbf{R}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}. \quad (2.2.3)$$

Although this is the best estimator for β in many senses (unbiased, L^2 optimal, consistent, etc.), it is also naturally overfit to the particular realisations of the signal and returns in the in-sample period, which will clearly differ in the out-of-sample period.

Moreover, using this estimate for β for backtesting on the same in-sample dataset naturally induces “look-ahead bias”, as data from the full sample period has been used to estimate the model parameters which are then used to test the trading strategy on the same period.⁴

To determine the magnitude of these effects, we can also stack the residuals ϵ_{t+1} into the matrix $\mathbf{E} \in \mathbb{R}^{N \times T_{\text{Is}}}$ and with a substitution of the model for the returns into (2.2.3) we obtain

$$\hat{\beta} = \beta + \mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}.$$

This expression enables us to decompose the estimated regression parameter into the true regression parameter plus some sampling error. If the researcher trades the portfolio $\hat{\mathbf{w}}_t = \mathbf{Z}\hat{\beta}\mathbf{s}_t$ using this estimate, the corresponding P&L per time step is $\widehat{\text{PnL}}_t = \hat{\mathbf{w}}_t^\top \mathbf{r}_{t+1}$. The sample average and variance in the training data set in turn

⁴ The look-ahead bias could be avoided using a walk-forward regression as in [Jou+24].

are

$$\widehat{\mathbb{E}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}} \right] := \frac{1}{T_{\text{IS}}} \sum_{t=0}^{T_{\text{IS}}-1} \widehat{\text{PnL}}_t, \quad (2.2.4)$$

$$\widehat{\mathbb{V}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}} \right] := \frac{1}{T_{\text{IS}} - 1} \sum_{t=0}^{T_{\text{IS}}-1} \left(\widehat{\text{PnL}}_t - \widehat{\mathbb{E}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}} \right] \right)^2. \quad (2.2.5)$$

Together, these lead to the in-sample Sharpe ratio

$$\widehat{\text{SR}}_{\text{IS}} := \frac{\widehat{\mathbb{E}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}} \right]}{\sqrt{\widehat{\mathbb{V}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}} \right]}}.$$

The issue here is that $\widehat{\boldsymbol{\beta}}$ has been estimated using the same returns observations $\mathbf{r}_{t+1}$ which the model is then tested against, i.e. the model has been fit to the specific realisation of the noise in the historical period. Expanding out the sample P&L,

$$\begin{aligned} \widehat{\text{PnL}}_t &= \widehat{\mathbf{w}}_t^\top \mathbf{r}_{t+1} = (\boldsymbol{\Sigma}_\epsilon^{-1} \widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top (\boldsymbol{\beta} \mathbf{s}_t + \boldsymbol{\epsilon}_{t+1}) \\ &= (\boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{\beta} \mathbf{s}_t)^\top \boldsymbol{\beta} \mathbf{s}_t + (\boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{\beta} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \leftarrow \text{Truth} \\ &+ \left(\boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t \right)^\top \boldsymbol{\beta} \mathbf{s}_t \leftarrow \text{Misestimation} \\ &+ \left(\boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t \right)^\top \boldsymbol{\epsilon}_{t+1} \leftarrow \text{Overfitting} \end{aligned}$$

We observe that by using $\widehat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$ we obtain a combination of the true P&L and some additional P&L as a result of the estimation error in $\boldsymbol{\beta}$. In the in-sample period $\boldsymbol{\epsilon}_{t+1}$ is an element of $\mathbf{E}$, but in the out-of-sample period it is not ($\mathbf{E}$ is the matrix of stacked noise realisations from the in-sample period). The impact of this is that, in expectation, the misestimation term has expected value zero in both the in- and out-of-sample period due to $\mathbb{E}[\mathbf{E}] = \mathbf{0}$. The overfitting term however only has zero expectation in the out-of-sample period, leading to inflation in the expected return in the in-sample period, this is how the look ahead bias discussed earlier inflates the in-sample performance. However, in both periods the variance of the sample P&L is higher than if we had used the true estimator $\boldsymbol{\beta}$, as we will see in

our main result in Proposition 2.3.1.

2.2.3 The Out-of-Sample Sharpe Ratio

After estimating the predictive model on the training dataset, the researcher deploys the model out-of-sample at discrete trading times $\mathcal{T}_{\text{OOS}} = \{T_{\text{IS}}+1, \dots, T_{\text{IS}}+T_{\text{OOS}}+1\}$. The corresponding sample mean and variance then lead to the out-of-sample Sharpe ratio

$$\widehat{\text{SR}}_{\text{OOS}} := \frac{\widehat{\mathbb{E}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{OOS}}} \right]}{\sqrt{\widehat{\mathbb{V}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{OOS}}} \right]}}.$$

The difference, of course, is that in the future period $\widehat{\beta}$ is not fit to the particular observations of the returns and signals, and thus the out-of-sample Sharpe ratio will typically be lower than the in-sample Sharpe ratio previously calculated. We also see from the expressions derived below that the out-of-sample Sharpe ratio will be lower than the true Sharpe ratio due to misestimation of the regression parameter causing increased volatility over trading with the true β .

The questions therefore are: How much lower? And what can be done to mitigate this bias?

2.3 Main Results

The unconditional means of the in-sample and out-of-sample Sharpe ratio, $\mathbb{E}[\widehat{\text{SR}}_{\text{IS}}]$ and $\mathbb{E}[\widehat{\text{SR}}_{\text{OOS}}]$, cannot be computed in closed form. We can however compute the expected in-sample and out-of-sample expected return and variance. For clarity of exposition, we present here our main result for a special (but extremely common) case, taking $\mathbf{Z} = \Sigma_{\epsilon}^{-1}$, $\mu_s = \mathbf{0}$ and $\Sigma_s = \mathbf{I}_K$, i.e., the Markowitz portfolio for a set of centred standardised signals. The impact of having to estimate the covariance matrix when constructing the portfolio is discussed in Section 2.3.3. The general

version of this result is reported in Appendix 2.B.

Proposition 2.3.1. *The expected average in-sample $P\mathcal{E}L$ and out-of-sample $P\mathcal{E}L$ is given by*

$$\begin{aligned}\mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\widehat{\text{PnL}}_t \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= \text{tr}(\mathbf{\Gamma}) + \frac{KN}{T_{\text{IS}}}, \\ \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\widehat{\text{PnL}}_t \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] &= \text{tr}(\mathbf{\Gamma}).\end{aligned}$$

The expected variance of the in-sample and out-of-sample $P\mathcal{E}L$ is given by

$$\begin{aligned}\mathbb{E} \left[\widehat{\mathbb{V}} \left[\left(\widehat{\text{PnL}}_t \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= 2 \text{tr}(\mathbf{\Gamma}^2) + (c_1 + \tilde{c}_1) \text{tr}(\mathbf{\Gamma}) + c_2 + \tilde{c}_2 + \varepsilon, \\ \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left(\widehat{\text{PnL}}_t \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] &= 2 \text{tr}(\mathbf{\Gamma}^2) + c_1 \text{tr}(\mathbf{\Gamma}) + c_2,\end{aligned}$$

where

$$\begin{aligned}\mathbf{\Gamma} &= \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{\beta}, \quad c_1 = 1 + \frac{K+1}{T_{\text{IS}} - K - 1}, \quad c_2 = \frac{NK}{T_{\text{IS}} - K - 1}, \\ \tilde{c}_1 &= \frac{2K+5}{T_{\text{IS}} - K - 1} + \frac{2N(K^2 + K + 2T_{\text{IS}})}{T_{\text{IS}}(T_{\text{IS}} - K - 1)}, \\ \tilde{c}_2 &= \frac{NK(2N + K + T_{\text{IS}} + 4)}{T_{\text{IS}}(T_{\text{IS}} + 2)} - \frac{2N^2K^2}{T_{\text{IS}}^2(T_{\text{IS}} + 2)} - \frac{NK}{T_{\text{IS}} - K - 1}.\end{aligned}\tag{2.3.1}$$

and

$$\begin{aligned}\varepsilon &= \text{tr} \left(\text{Cov} \left[(\mathbf{S}\mathbf{S}^\top)^{-1}, \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right) + \text{tr} \left(\text{Cov} \left[(\mathbf{S}\mathbf{S}^\top)^{-1}, \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right) \\ &+ \text{tr} \left(\text{Cov} \left[(\mathbf{S}\mathbf{S}^\top)^{-1}, \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right).\end{aligned}$$

The proof relies on a combination of results on the expectations of Gaussian vectors or matrices, properties of projection matrices and matrix statistics tricks. For better readability, these lengthy computations are deferred to Appendix 2.C. The out-of-sample expected PnL and its variance as well as the in-sample expected PnL can be computed explicitly. The in-sample variance of the PnL is also almost fully

explicit, up to the ε term which does not have a closed-form expression. In practice, we approximate the in-sample variance by simply dropping this term as it is typically very small compared to the other terms. The accuracy of this approximation is verified in Section 2.3.1. Here, let us mention a few implications.

Remark 1 (Expected Mean Return). Examining the difference between the expected in-sample and out-of-sample returns, we see that the in-sample return is inflated by $\frac{KN}{T_{\text{IS}}}$. Thus, as the number of in-sample observations T_{IS} increases, our in-sample P&L will converge to our out-of-sample P&L. Conversely, as the number of signals K increases the overfitting bias becomes arbitrarily large. Increasing the number of assets N will also lead to further inflation of the in-sample P&L (see Figure 2.3.5).

Remark 2 (Expected Variance). Inspecting the expected variance, we see that $\tilde{c}_1 > 0$ provided $T_{\text{IS}} > K + 1$, which we would always expect to be the case, but whether $\tilde{c}_2 > 0$ depends on the interplay between the length T_{IS} of the backtest, as well as the number N and K of assets and signals. For a range of backtest lengths, Figure 2.3.1 displays the combinations of N, K for which $\tilde{c}_2 > 0$. Generally, when $K < N$ (with some sublinearity for large N) then $\tilde{c}_2 > 0$, and for large enough T_{IS} then K must be much larger than N for $\tilde{c}_2 < 0$. Overfitting β thus causes the in-sample variance to be marginally larger than the out-of-sample variance for small N , but for large N it can be significantly higher. Further, the out-of-sample variance can be much larger than the true variance, which causes the out-of-sample Sharpe to be lower than the true theoretical Sharpe ratio if β was known.

2.3.1 Accuracy of the Approximation of the In-Sample Variance

In our simulation study for commodity futures in Section 2.5, we find that the error term ε from Proposition 2.3.1 is negative but small relative to the other terms, with the estimated variance being 3% smaller than the analytical result with $\varepsilon = 0$. To support the robustness of this observation across a wide range of parameter values,

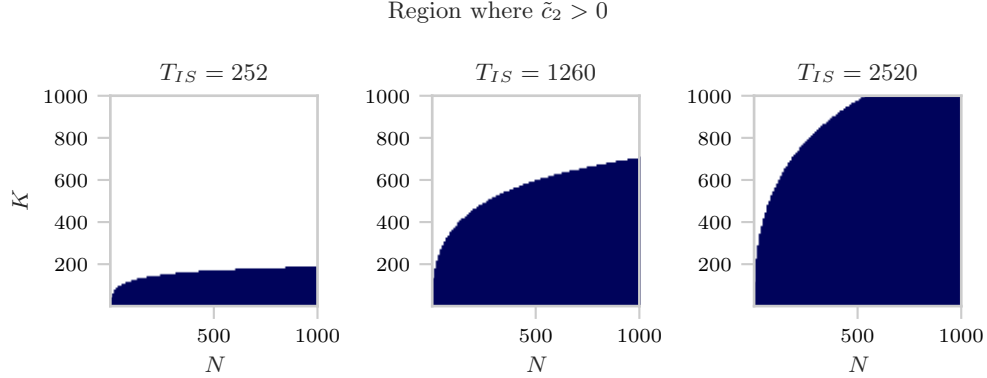


Figure 2.3.1: Values of N, K, T_{IS} where $\tilde{c}_2 > 0$.

we run separate simulations where the model parameters are selected randomly from distributions centred around the empirical point estimates (full details can be found in Appendix 2.D). This is distinct from the case study in Section 2.5, where all true parameters are calibrated to that particular dataset and held fixed. We find that for T_{IS} sufficiently larger than K , ε are small, less than 3% for $K/T_{IS} < 0.1$ as can be seen in Figure 2.3.2. This is equivalent to having at least ten data points per signal which in fact is a common rule of thumb [Har06, Sec. 4.4]. Furthermore, N has no strong effect on ε as these terms are driven by the covariance between the inverse Gram matrix $(\mathbf{S}\mathbf{S}^\top)^{-1}$ and the individual signal vectors $\mathbf{s}_t$, which becomes negligible as the number of signals K increases. Based on these observations, we henceforth disregard ε in our analysis.

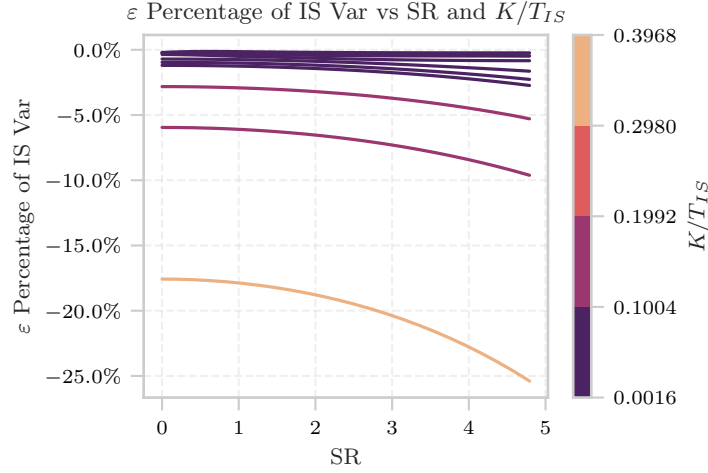


Figure 2.3.2: Magnitude of the total errors $\sum_i \varepsilon_i$ as a percentage of the in-sample variance $\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]$.

2.3.2 The Replication Ratio

In view of Proposition 2.B.1 and the discussion in Section 2.3.1 we propose the following approximations to the expected in- and out-of-sample Sharpe ratios:

$$\begin{aligned} \mathbb{E}[\widehat{\text{SR}}_{\text{IS}}] &\approx \text{SR}_{\text{EIS}} := \frac{\mathbb{E}[\widehat{\mathbb{E}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]]}{\sqrt{\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]}}}, \\ \mathbb{E}[\widehat{\text{SR}}_{\text{OOS}}] &\approx \text{SR}_{\text{EOOS}} := \frac{\mathbb{E}[\widehat{\mathbb{E}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{OOS}}}]]}{\sqrt{\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{OOS}}}]}}}. \end{aligned} \tag{2.3.2}$$

Here we take $\varepsilon = 0$ in $\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]$ as discussed in Section 2.3.1. Moreover, we assume that the difference between taking expectations of the numerator and denominator individually, rather than jointly, is small.

We again check this by the same simulation process as Section 2.3.1 and find that indeed the convexity adjustment is generally very small, with maximum errors of 0.01 for the in-sample Sharpe ratio and 0.005 for the out-of-sample Sharpe ratio,

as can be seen in Figure 2.3.3. Compared to the magnitude of the sample Sharpe ratios (shown in the second row), the errors due to convexity are negligible, at most 2.5%. In these simulations we let $K = N$ and $T = T_{\text{IS}} = T_{\text{OOS}}$.⁵

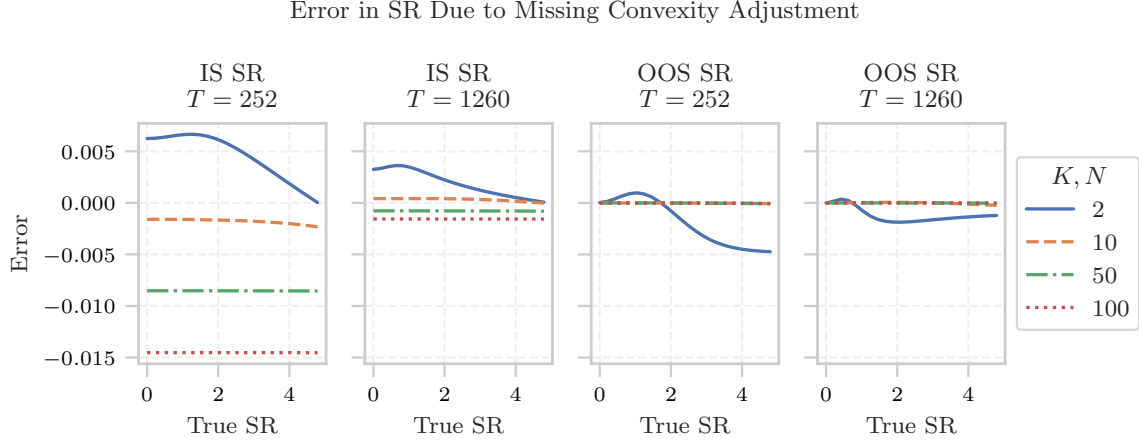


Figure 2.3.3: Difference between the expected Sharpe Ratio $\mathbb{E}[\text{SR}]$ and the ratio of the expected mean and variance $\mathbb{E}[\widehat{\mathbb{E}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_i}]] / \sqrt{\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_i}]]}$, i.e. the error due to the missing convexity adjustment for both in-sample and out-of-sample Sharpe ratios.

Armed with these analytical approximations, we can study the “replication ratio”, i.e., the fraction $\text{SR}_{\text{EOOS}}/\text{SR}_{\text{EIS}}$ of the in-sample Sharpe ratio that is recovered out of sample. A replication ratio of 100% means that the strategy completely replicates out of sample. Conversely, a replication ratio of 0% means that it returns zero, e.g., because the signals are completely uninformative (i.e., the true parameter β is $\mathbf{0}$) and thus the expected out-of-sample Sharpe ratio is zero.

Univariate Case To gain some intuition we first evaluate the replication ratio in the simplest case with one asset ($N = 1$) and one dynamic signal with no intercept ($K = 1$) and where we scale σ_s to be equal to σ_ϵ , which is always possible as one

⁵ Note these estimates are purely calculated via simulation, and we do not use our analytical values. This is as the error due to the missing ε terms (as studied in Section 2.3.1) dominates the small convexity error when $K = 100, T_{\text{IS}} = 252$, yielding no insight into the magnitude of the convexity error.

can always scale the original signal to achieve this. In this case the formulas derived above simplify as

$$\begin{aligned} \text{SR} &= \frac{\beta^2}{\sqrt{2\beta^4 + \beta^2}}, \\ \text{SR}_{\text{EIS}} &= \frac{\beta^2 + \frac{1}{T_{\text{IS}}}}{\sqrt{2\beta^4 + \left(1 + \frac{15}{T_{\text{IS}}-2} - \frac{2}{T_{\text{IS}}}\right) \beta^2 + \frac{4}{T_{\text{IS}}} - \frac{3}{T_{\text{IS}}+2} - \frac{1}{T_{\text{IS}}^2}}}, \\ \text{SR}_{\text{EOOS}} &= \frac{\beta^2}{\sqrt{2\beta^4 + \left(1 + \frac{2}{T_{\text{IS}}-2}\right) \beta^2 + \frac{1}{T_{\text{IS}}-2}}}. \end{aligned}$$

Whence, the replication ratio only depends on the size T_{IS} of the training dataset and the parameter β (or, equivalently, the true Sharpe ratio) in this case:

$$\frac{\text{SR}_{\text{EOOS}}}{\text{SR}_{\text{EIS}}} = \frac{\beta^2 \sqrt{2\beta^4 + \left(1 + \frac{15}{T_{\text{IS}}-2} - \frac{2}{T_{\text{IS}}}\right) \beta^2 + \frac{4}{T_{\text{IS}}} - \frac{3}{T_{\text{IS}}+2} - \frac{1}{T_{\text{IS}}^2}}}{\left(\beta^2 + \frac{1}{T_{\text{IS}}}\right) \sqrt{2\beta^4 + \left(1 + \frac{2}{T_{\text{IS}}-2}\right) \beta^2 + \frac{1}{T_{\text{IS}}-2}}}. \quad (2.3.3)$$

We plot a heat map of this relationship in Figure 2.3.4. As the Sharpe ratio (equivalently β) increases, overfitting becomes less of an issue. Indeed, as the signal-to-noise ratio becomes arbitrarily large for $\beta \rightarrow \infty$, the replication ratio approaches 100% ($\text{SR}_{\text{EOOS}}/\text{SR}_{\text{EIS}} \rightarrow 1$). Similarly, the difference between in- and out-of-sample performance also vanishes as the training data set becomes arbitrarily large ($T_{\text{IS}} \rightarrow \infty$). Both these relationships are nonlinear, with large initial improvements that eventually saturate.

Multivariate Case We now turn to the case of multiple assets and signals ($N, K > 1$). To illustrate some of the basic mechanisms, we look at a representative example. To wit, we consider an in-sample Sharpe ratio of 2 with a ten-year backtest of daily returns ($T_{\text{IS}} = 2520$), and assume that assets and signals are uncorrelated, $\Sigma_s = \mathbf{I}_K$, $\Sigma_\epsilon = \mathbf{I}_N$. For simplicity, we also assume that the signals include no intercept ($\mu_s = \mathbf{0}$) and that investments are directly proportional to the signals without

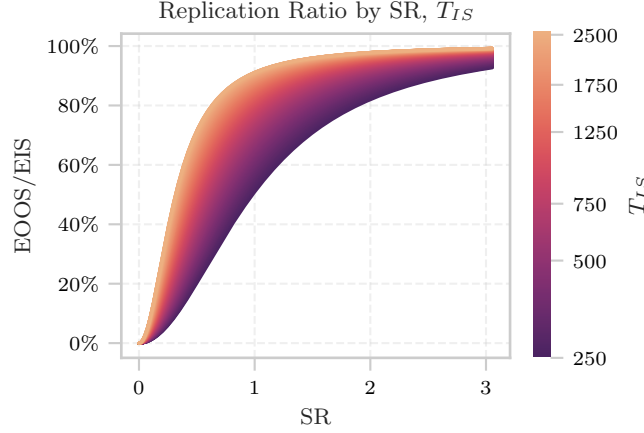


Figure 2.3.4: Replication ratio $\frac{\text{SR}_{\text{EOOS}}}{\text{SR}_{\text{EIS}}}$, SR_{EIS} and SR_{EOOS} in the 1-asset, 1-signal case as in (2.3.3), with varying SR , T_{IS} .

reweighing ($\mathbf{Z} = \mathbf{I}_N$). Finally, to achieve the desired in-sample Sharpe ratio of 2, we let $\boldsymbol{\beta} = b\mathbf{1}_{N,K}$ for a suitable constant b and the matrix $\mathbf{1}_{N,K} \in \mathbb{R}^{N \times K}$ of all ones.

This is a kind of worst-case scenario for overfitting, because independent signals and returns introduce a maximal amount of freedom into the model. In this case the replication ratio becomes

$$\frac{\text{SR}_{\text{EOOS}}}{\text{SR}_{\text{EIS}}} = \frac{b^2 K N}{b^2 K N + \frac{K N}{T_{\text{IS}}}} \frac{\sqrt{2(b^2 K N)^2 + (c_1 + \tilde{c}_1)b^2 K N + c_2 + \tilde{c}_2}}{\sqrt{2(b^2 K N)^2 + c_1 b^2 K N + c_2}}, \quad (2.3.4)$$

for the constants $c_i, \tilde{c}_i$ defined in (2.3.1). Whence, for independent signals and residuals, we obtain a similar expression as in the univariate case (2.3.3).

Figure 2.3.5 shows that, conditional on constant in-sample Sharpe ratio $\text{SR}_{\text{EIS}} = 3$, the replication ratio (2.3.4) declines more quickly in K as the number of assets N increases. Why does this happen? For each additional asset or signal we increase the freedom in the model to overfit to the historical period, therefore reducing the replication ratio.

One way to address this is to reduce the number of model parameters, e.g., by running a pooled regression that assumes that, up to some suitable normalisations, a family of common signals affects all assets in the same manner. For example,

[GP13] normalise their momentum signals for each asset to make them comparable in the cross-section and then only fit one parameter across all stocks for each of these signals, rather than many stock-specific coefficients.⁶ While it is entirely possible that the assets have differing true reactions to stimuli (for example, stocks in different sectors may respond differently to the same economic shock), in many cases this fact is overruled by the benefit that can be gained through reducing the number of parameters and avoiding overfitting. A concrete example of this is provided in Section 2.5 where we observe that a panel regression increases the replication ratio from 71% to 30% for a ten-year backtest.

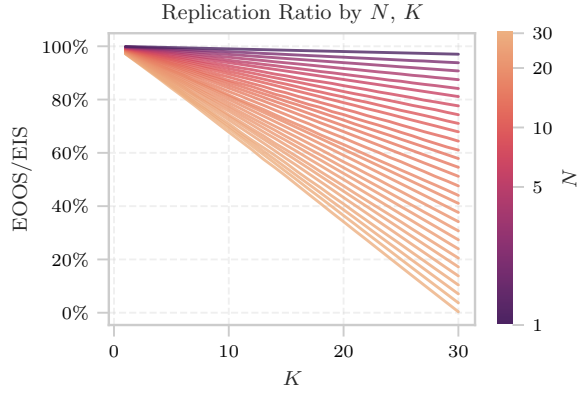


Figure 2.3.5: Replication ratio $\frac{SR_{EOOS}}{SR_{EIS}}$ for increasing assets N and signals K , and fixed in-sample Sharpe ratio $SR_{EIS} = 3$ with $T_{IS} = 252 \cdot 100$. Large T_{IS} is required to keep SR_{EIS} fixed at 3 as p, m increase, with smaller T_{IS} we cannot always find β such that $SR_{EIS} = 3$.

Importantly, the above discussion is linked to the Sharpe ratio being held constant as the number of assets N increases. Generally, we observe that Sharpe ratios increase as we add additional uncorrelated assets to strategies, primarily due to diversification. To test when this additional performance boost outstrips the overfitting effect we consider a simple case with one signal and an increasing number of assets that the signal works equally well on. Figure 2.3.6 shows that if these assets are

⁶ This is reminiscent to the fitting of price impact models where, after normalising for each asset's volatility and trading volume, universal models also often outperform stock-specific models out of sample, as in [MWW24] and references therein.

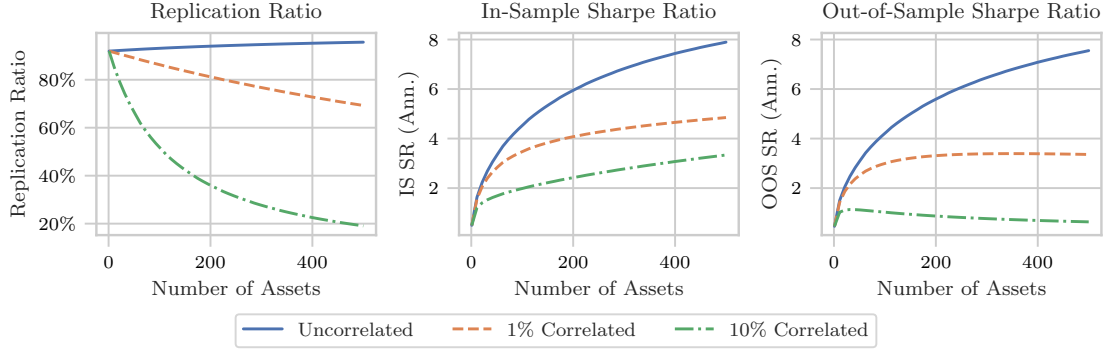


Figure 2.3.6: Replication ratio, IS Sharpe ratio and OOS Sharpe ratio as assets increase under different correlations

completely independent, then adding additional assets results in both an increased replication ratio and out-of-sample Sharpe ratio. However, even with just 1% average correlation between assets we see that the replication ratio already becomes monotonically decreasing. For such a low level of correlation between the assets, the out-of-sample Sharpe ratio does initially increase as more assets are added because diversification more than compensates for overfitting. However, already with 10% correlation this effect is reversed. In summary, adding assets generally is bad for your replication ratio. But if you can find some lowly correlated assets with consistent predictive power, it is certainly a benefit to the out-of-sample performance to include them.

2.3.3 Unknown Covariances

In Proposition 2.3.1 we have assumed that the true covariance matrix Σ_ϵ is known when we construct the portfolio $\hat{\mathbf{w}}_t$. Of course, in practice we must also estimate this parameter from the observed data. This is not an easy task, and much has been written on how one can improve your estimation of the covariance matrix of returns [BBP17]. However, the corresponding estimation errors are usually of secondary importance compared to misestimation of the regression parameter β .

When the covariance matrix is also estimated, it is no longer feasible to obtain

analytical results on the replication ratio. However, it is straightforward to assess this in a simulation study. To this end we perform a Monte Carlo simulation with 30,000 samples where, as in Section 2.3.1, we select random parameters centred around the empirical point estimates from Section 2.5 (cf. Appendix 2.D for more details) for varying numbers of signals K and assets N , along with varying lookback windows T_{IS} . We form portfolios using either the known precision matrix Σ_ϵ^{-1} or the estimated precision matrix from the observed residuals $\widehat{\Sigma}_\epsilon^{-1}$.

In the high-dimensional regime where the dimension N exceeds the effective sample size $T_{\text{eff}} = T_{\text{IS}} - 1$, the sample covariance matrix $\widehat{\Sigma}_\epsilon$ is rank-deficient and singular. To ensure the matrix is invertible we use a simple trace-preserving linear shrinkage estimator. We define the regularised covariance matrix as:

$$\widehat{\Sigma}_{\text{reg}} = 1/q \widehat{\Sigma}_\epsilon + (1 - 1/q) \bar{\lambda} \mathbf{I}_N,$$

where $\mathbf{I}_N$ is the identity matrix, $\bar{\lambda} = \text{tr}(\widehat{\Sigma}_\epsilon)/N$ represents the average eigenvalue, and the shrinkage intensity is determined by the ratio $q = N/T_{\text{eff}}$. This approach preserves the trace (total variance of the residuals) while giving us an invertible covariance matrix $\widehat{\Sigma}_{\text{reg}}$. Note that when $q < 1$ we simply use the estimated covariance matrix directly.

Figure 2.3.7 displays the in-sample Sharpe ratio, the out-of-sample Sharpe ratio and the replication ratio for $N = 20$ and $N = 500$. In the first case, where the asset universe is small, results using the estimated and true covariance matrices are nearly identical. In the high-dimensional setting ($N = 500$), the in-sample Sharpe ratio is inflated further when using the estimated covariance matrix instead of the true one. Regarding the out-of-sample Sharpe ratios, we see that when the number T of time steps observed is less than the number of assets ($N = 500$, i.e., where the covariance matrix is rank-deficient) then the out-of-sample Sharpe ratio is much worse with the estimated covariance matrix (almost zero). However, once this is resolved ($T > N$), the estimated portfolio actually slightly outperforms the one constructed with the

true covariance matrix. Moreover, as long as $T > N$, the replication ratio is very well aligned even when N is large.

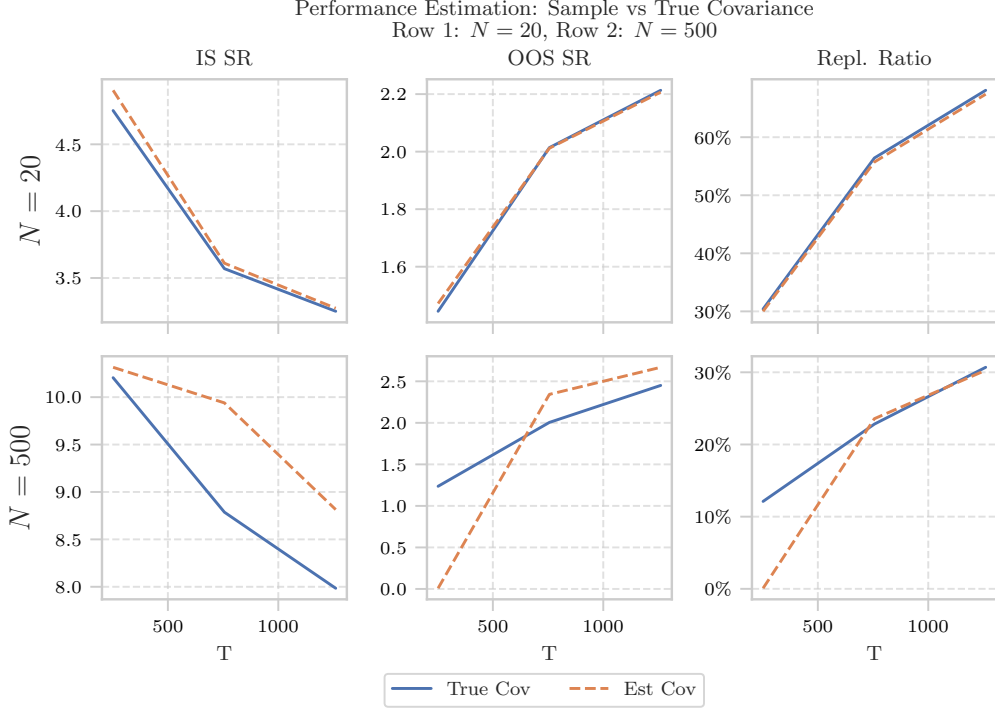


Figure 2.3.7: In- and out-of-sample Sharpe ratio and replication ratio using either the true or estimated covariance matrices Σ_ϵ , $\hat{\Sigma}_\epsilon$ to form $\hat{\mathbf{w}}_t$ in Monte Carlo simulations.

2.3.4 Autoregressive Signals

So far, we assumed for analytical tractability that signals and residuals are iid. A typical assumption (in [GP13] for example) is that instead the signal follows an AR(1) process, so that signals are persistent and investment opportunities in turn autocorrelated across time. Such persistent signals are not analytically tractable in our framework, but we can assess the corresponding replication ratios via simulations. We examine the simplest case with one asset and one AR(1) signal,

$$r_{t+1} = \beta s_t + \epsilon_{t+1}, \quad s_t = \phi s_{t-1} + u_t,$$

where $\epsilon_{t+1} \sim \mathcal{N}(0, 1)$ and u_t is a centred white noise process with unit variance.⁷ In this case, the true Sharpe ratio is

$$\text{SR} = \frac{\mathbb{E}[w_t r_{t+1}]}{\sqrt{\mathbb{V}[w_t r_{t+1}]}} = \frac{\beta^2 \sigma_s^2}{\sqrt{2\beta^4 \sigma_s^4 + \beta^2 \sigma_s^2}} = \frac{\beta}{\sqrt{1 + 2\beta^2 - \phi^2}}, \quad (2.3.5)$$

where we have taken $w_t = \beta s_t$, and assumed $\beta > 0$ without loss of generality. For the second equality in (2.3.5), we have used that the kurtosis of the signal's noise $\gamma_{s,4} = 3$ for u_t normally distributed, and $\sigma_s^2 = \frac{1}{1-\phi^2}$ for the final equality.

To understand the impact of the persistence of AR(1) signal compared to their iid counterparts, we proceed by Monte Carlo simulations. We let $T_{\text{OOS}} = 1260$ and vary $T_{\text{IS}} \in \{252, 1260, 2520\}$, $\phi \in \{0, 0.5, 0.9, 0.99, 0.999\}$ and $\text{SR} \in [0, 2]$. Note that $\phi = 0.9$ corresponds to a signal half-life of 6.5 (daily) time-steps, whereas $\phi = 0.999$ implies a half-life of 690 time-steps. For a given signal persistence ϕ we vary the parameter β to achieve the desired true Sharpe ratio in each case using (2.3.5).

In Figure 2.3.8 we compare the replication ratios (and their difference to their counterparts computed in our model with iid signals) for different signal persistences parametrised by ϕ . Unless ϕ is very close to 1 (> 0.99), the difference compared to the iid case $\phi = 0$ is surprisingly small. Thus, although our assumption of iid signals may initially seem like a strong one, it in fact has little impact on the replication ratio in practice.

⁷ We assume that $|\phi| < 1$ so that (s_t) is weakly stationary thus $\mathbb{E}[s_t] = 0$ and $\mathbb{V}[s_t] = \frac{1}{1-\phi^2}$.

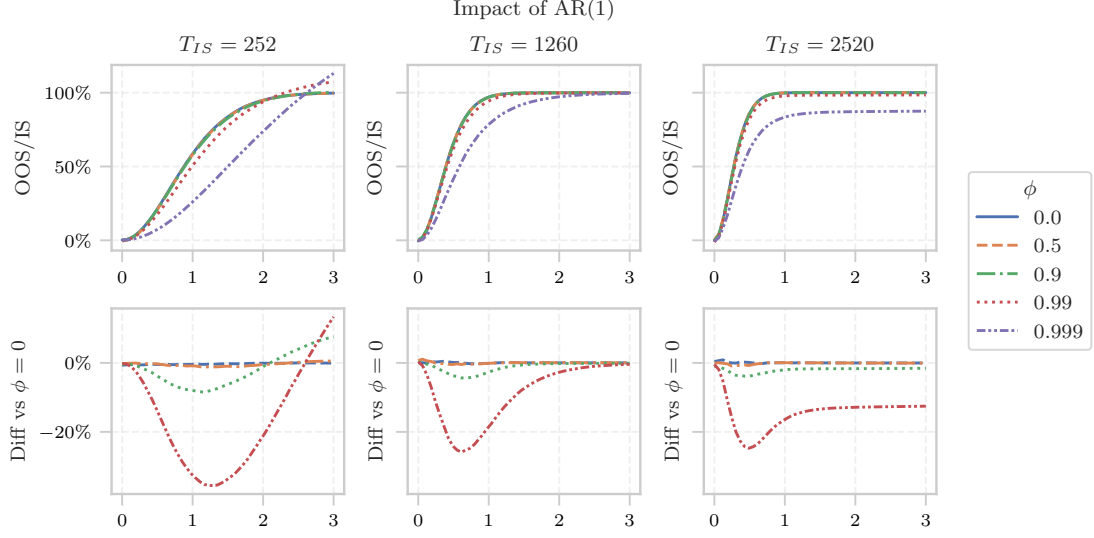


Figure 2.3.8: Impact of s_t following an AR(1) process. The top row shows the replication ratio, and the bottom row shows the simple difference of this ratio vs $\phi = 0$, corresponding to the iid case.

2.4 Comparison with the Literature

Most closely related to the present study is the work of Kan, Wang, and Zheng [KWZ24].

They compare the expected in- and out-of-sample Sharpe ratios of the Markowitz portfolio for N assets whose expected returns are estimated on a training dataset of size T . They find that the expected in-sample Sharpe ratio is

$$\mathbb{E}[\hat{\theta}] = \frac{\Gamma\left(\frac{N+1}{2}\right) \Gamma\left(\frac{T-N-1}{2}\right)}{\Gamma\left(\frac{N}{2}\right) \Gamma\left(\frac{T-N}{2}\right)} {}_1F_1\left(-\frac{1}{2}; \frac{N}{2}; -\frac{N\theta^2}{2}\right),$$

where $\Gamma(\cdot)$ is the Gamma function, ${}_1F_1(\cdot; \cdot; \cdot)$ the confluent hypergeometric function, and $\theta = \sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}$ the true Sharpe ratio of the Markowitz portfolio. Likewise, the expected out-of-sample Sharpe ratio can also be computed in closed form as

$$\mathbb{E}[\tilde{\theta}] = \frac{\theta^2 \sqrt{T} \Gamma\left(\frac{N+1}{2}\right) \Gamma\left(\frac{T-N+2}{2}\right) \Gamma\left(\frac{T}{2}\right)}{\sqrt{2} \Gamma\left(\frac{N+2}{2}\right) \Gamma\left(\frac{T-N+1}{2}\right) \Gamma\left(\frac{T+1}{2}\right)} {}_1F_1\left(\frac{1}{2}; \frac{N+2}{2}; -\frac{T\theta^2}{2}\right).$$

In the framework of [GP13], the setting of [KWZ24] corresponds to using a single,

infinitely persistent trading signal ($\phi \rightarrow 1$) for each asset. In contrast, we consider multiple signals per asset that decay quickly ($\phi \rightarrow 0$). Incorporating multiple signals per asset allows us to cover many common trading strategies, such as momentum strategies based on several different moving averages of past returns. Figure 2.4.1 illustrates how overfitting becomes an increasingly important concern when allowing for additional freedom in the model by considering multiple signals per asset. We fix $T = T_{\text{IS}} = T_{\text{OOS}} = 2520$, $\text{SR} = 1.5$, vary the number of assets N in both models and set either $K = 1$ or $K = 10$ in our model. We observe that for a single signal ($K = 1$) our haircuts are similar to the ones of [KWZ24], but for $K = 10$ signals our haircut is much more severe. Note that this is conditional on a fixed true Sharpe ratio, and therefore the strength of each signal must be lower as K increases; in practice, adding signals involves a trade-off between increasing the true SR and raising the risk of overfitting.

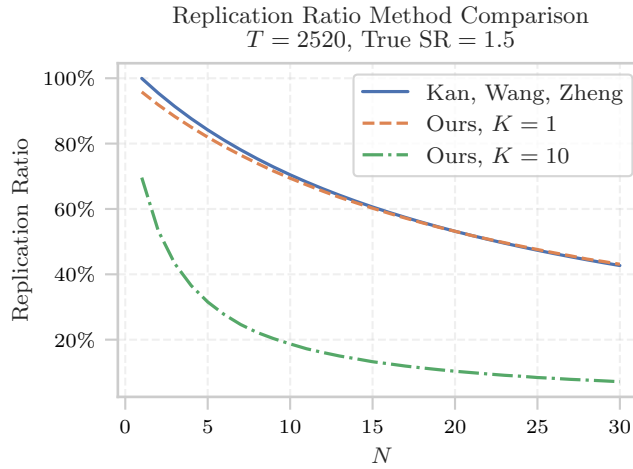


Figure 2.4.1: Comparison of the replication ratio for Kan, Wang, and Zheng and ours with $T = T_{\text{IS}} = T_{\text{OOS}} = 2520$ and $\text{SR} = 1.5$ for varying N . For ours we set $K = 10$ or $K = 1$.

An important common conclusion of both [KWZ24] and the present study is the non-linear dependence of the replication ratio on the true Sharpe ratio. To wit, for small values of the true Sharpe ratio, the expected out-of-sample Sharpe

ratio is much lower than the expected in-sample Sharpe ratio. A takeaway from this is that when developing investment strategies, it is preferable to develop a single high Sharpe strategy than many low Sharpe strategies. This aligns with the accepted wisdom of successful practitioners, for example Peter Muller (Founder, PDT Partners) who quipped [Mul01]:

In my opinion it is far better to refine an individual strategy...than to attempt to put together lots of weaker strategies. Depth is more important than breadth for investment strategies... I would much rather have a single strategy with an expected Sharpe ratio of 2 than a strategy that has an expected Sharpe ratio of 2.5 formed by putting together five supposedly uncorrelated strategies each with an expected Sharpe ratio of 1.

2.5 Predicting Commodity Futures

To test our method we follow the approach of Gârleanu and Pedersen [GP13] in the empirical application section of their seminal paper on dynamic trading. We select the commodity futures in Table 2.5.1 and collect close prices for the closest-to-expiry contracts from May 1998 to December 2023. The daily returns are computed as close-to-close percent changes in prices.

As in [GP13], we compute the 5-day, 1-year and 5-year Sharpe ratios for each asset and use these as signals. In effect these are normalised versions of the moving averages typically used in momentum-style strategies. We additionally transform the signals to be mean zero and unit variance. Normalizing the variances enables us to easily apply L^2 -regularisation when fitting β to calibrate the model we later use as a simulation engine; the centring is required per our conditions on the signals in the theoretical model. This yields a total of 36 dynamic (cross-sectional) signals, which with the addition of the asset-specific intercept yields 37 regressors per asset.

To assess the validity of our results we need to compare our analytical approximations for the expected in-sample and out-of-sample Sharpe ratios, SR_{EIS} and

Commodity	Venue	Ticker
Aluminium	LME	MAL3=LX
Copper	LME	MCU3=LX
Nickel	LME	MNI3=LX
Zinc	LME	MZN3=LX
Lead	LME	MPB3=LX
Gasoil	ICE	LGOc1
WTI	NYMEX	CLc1
Gold	COMEX	GCc1
Silver	COMEX	SIc1
Coffee	NYBOT	KCc1
Cocoa	NYBOT	CCc1
Sugar	NYBOT	SBc1

Table 2.5.1: Commodities selected for testing from the LME (London Metal Exchange), ICE (Intercontinental Exchange), NYMEX (New York Mercantile Exchange), COMEX (Commodity Exchange) and NYBOT (New York Board of Trade).

SR_{EOOS} , to similarly averaged observed equivalents. This requires more than the one realisation of the returns available historically. To resolve this, we simulate many realisations of a fitted model and use these to check our results.

The first step in enabling this simulation is to estimate a model for the returns, which naturally we take as

$$\mathbf{r}_{t+1} = \boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1}.$$

We estimate $\boldsymbol{\beta} \in \mathbb{R}^{37 \times 12}$ by ridge regression with regularisation parameter 0.1. We additionally estimate the covariance of the signals $\boldsymbol{\Sigma}_s$ and residuals $\boldsymbol{\Sigma}_\epsilon$ from the historical data (noting that the means are zero by construction).

To simulate non-iid and non-Gaussian signals and residuals we also fit an AR(1) process to the historically observed signals,

$$\mathbf{s}'_t = \Phi \mathbf{s}'_{t-1} + \mathbf{u}_t,$$

where $\mathbf{s}'_t \in \mathbb{R}^{36}$ denotes the non-intercept signals, $\mathbf{u}_t \in \mathbb{R}^{36}$ denotes the shocks to the AR(1) process, and $\Phi \in \mathbb{R}^{36 \times 36}$ denotes the AR(1) parameter.

Additionally, we consider the impact of ϵ_{t+1} and $\mathbf{u}_t$ not being multivariate Gaussian, by fitting a t -distribution to each dimension of the residuals and shocks and using these to simulate residuals and shocks with fat tails. To induce the original covariance structure of the residuals and shocks we scale the independent samples by the Cholesky transformation of the relevant covariance matrix, Σ_ϵ or Σ_u .

With the fitted models we simulate 10,000 realisations of the signals and residuals under both the Gaussian IID model, and with the AR(1) with fat tails model. All fitted parameter values from this section can be found on the authors' webpages. For context, we observe that the average correlation between signals is 20%, with most signals exhibiting very low correlation, and a few exhibiting very high correlation. The average correlation between residuals is 28% and the average excess kurtosis for the residuals is 9. The diagonal of the Φ matrix has average 0.88, and the off-diagonal has average 0.02.

As mentioned in the introduction, the haircuts we observe for these trials are quite steep, and they could be improved by reducing the amount of freedom in the model, for example by panel regression. Indeed, making the structural assumption that the normalised momentum signals have universal weights across assets reduces the number of model parameters for $N = 12$ assets and $K = 3 \times 12 + 1 = 37$ signals each from 444 to 3. In simulations where the full model is used to generate the data but the panel regression is used for the portfolio construction, we see that the replication ratio is 70%, much higher than the 38% observed when the portfolio is constructed using the beta from the full regression. This is due to reduced overfitting, causing us to have lower in-sample Sharpe ratios (0.9 vs 8.2) but more consistent out-of-sample Sharpe ratios (0.6 vs 3.1). On the real data, when we compare panel regression with the full model we observe that the panel portfolio not only has a better replication ratio than the full model (68% vs 9%), but it also has a higher out-of-sample Sharpe ratio (0.5 vs 0.3), a sharp reversal given the latter's substantially better in-sample performance (panel 0.7, full 3.6). This further cements our advice

to prefer simpler models wherever possible.

In addition to the study below, we conduct an additional study on the impact of using the estimated covariance matrix $\hat{\Sigma}_\epsilon$ rather than the true covariance matrix Σ_ϵ to construct the portfolio, the results of which can be seen in Appendix 2.E. Broadly, we find that the impact is minimal, which affirms our previous study in Section 2.3.3.

2.5.1 Gaussian and iid Signals and Residuals

We first test our approach in the case where the data-generating model satisfies our key assumptions, namely that the signals and residuals are Gaussian and iid.

We simulate 13 years, with 252 trading days per year, of new signals and returns, and use 10 years as in-sample and 3 years as out-of-sample data. For each in-sample simulation we compute the OLS estimator for $\hat{\beta} = \mathbf{R}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}$, where $\mathbf{R} \in \mathbb{R}^{12 \times 252 \cdot 10}$ is the stacked matrix of returns and $\mathbf{S} \in \mathbb{R}^{37 \times 252 \cdot 10}$ is the stacked matrix of signals. We then form the portfolio $\hat{\mathbf{w}}_t = \Sigma_\epsilon^{-1} \hat{\beta} \mathbf{s}_t$ as described in Section 2.2, utilising the true covariance matrix from the historical data. (As noted in Section 2.3.3 estimating the covariance matrix as well only has a minor impact. This is confirmed also for this example, see Table 2.E.1). This portfolio is then tested on the in-sample and out-of-sample periods to compute the observed in-sample and out-of-sample Sharpe ratios for each simulation.

In Figure 2.5.1 we see that the simulated expected in-sample and out-of-sample Sharpe ratios align perfectly with our analytical results, and so does the replication ratio. This shows that with real-world parameter values our analytical approximations are highly accurate. We note here that we have plotted the sample Sharpe ratios directly, and are thus comparing $\mathbb{E}[\widehat{\text{SR}}_{\text{IS}}]$ with SR_{EIS} and similarly for the out-of-sample Sharpe ratios. For completeness, we report the average mean, variance and other metrics in Table 2.E.1. In the next section we additionally check the robustness of our results when our assumptions are violated.

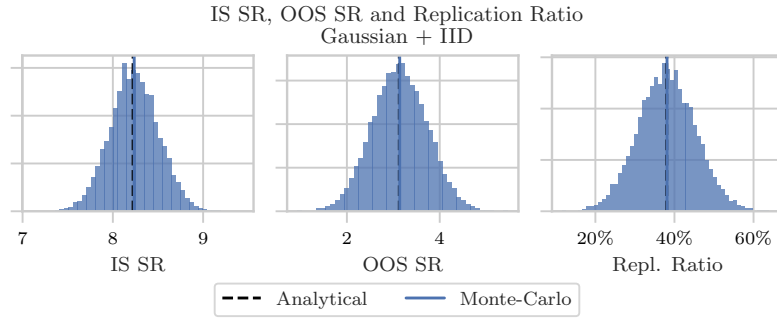


Figure 2.5.1: The expected in-sample Sharpe ratio, expected out-of-sample Sharpe ratio and replication ratio for the experiments in Section 2.5.1.

2.5.2 AR(1) Signals and t -distributed Residuals and Shocks

Next, we examine how our calculations hold up when the assumptions of the model are no longer true by simulating the returns and signals using the AR(1) model with fat-tail residuals and shocks.

In Figure 2.5.2 we see that even when our assumptions are violated, our analytical replication ratio is still remarkably close to its simulation observed counterpart. Relaxing the Gaussian assumption to allow for fat tails causes the in-sample Sharpe ratio to decrease, but also causes the out-of-sample Sharpe ratio to decrease. Additionally, allowing the signals to be AR(1) rather than iid causes the out-of-sample Sharpe ratio to decrease, however this effect is rather small. Finally, as the out-of-sample and in-sample Sharpe ratio are affected by the non-Gaussian returns, and the AR(1) signals have a relatively small effect, the replication ratio with both is almost identical to our analytical value under the Gaussian and iid assumptions. This confirms the practical relevance of our analytical results.

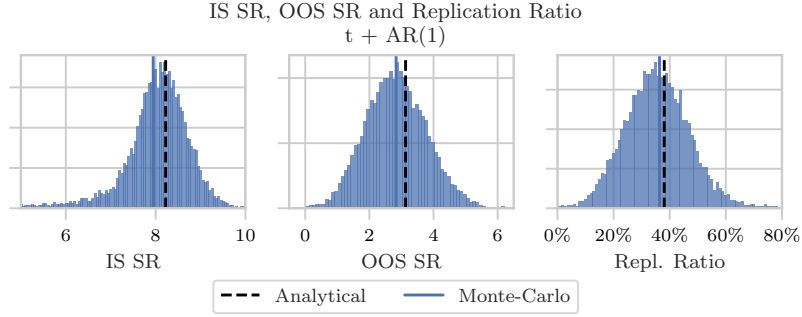


Figure 2.5.2: The expected in-sample Sharpe ratio, expected out-of-sample Sharpe ratio and replication ratio for the experiments in Section 2.5.2.

2.6 Empirical Study: Goyal, Welch, and Zafirov Dataset

Dataset

As another test of our approach, we use the extensive dataset compiled by Goyal, Welch, and Zafirov [GWZ24]. The authors have constructed monthly, quarterly and annual signals which have been proposed in the academic literature to predict the equity premium of the CRSP US Index, a broad US equity index. We utilise the monthly predictors, of which there are 39 with at least some data from 1926 to 2024, almost 1200 months. These predictors are highly diverse, ranging from macroeconomic indicators, measures of investor sentiment, and variance-related signals to stock cross-sectional data, technical indicators, and commodity prices. For further details, an extensive description of each predictor is given in [GWZ24]. To pre-process these signals and the corresponding returns we wish to predict, we follow the approach of [KMZ24], who also leverage this dataset, and normalise the returns by the rolling 12-month volatility and the signals by their respective expanding volatility.

Empirical Methodology

Our analytical results are derived for the *expected* replication ratio, so to check their validity we must perform some sort of averaging to mitigate the random out- or under-performance of any given signal and sample period. To achieve that on this dataset we proceed as follows.

We repeatedly select random groups of signals of sizes between 1 and 39. We then randomly draw total lengths between 10 and 70 years and split them into (chronologically ordered) in- and out-of-sample periods with both legs at least five years long. Specifically, we perform this sampling 5,000,000 times. We constrain the sampled signals and periods to be such that each signal has full data coverage over the in- and out-of-sample period (as not all signals have data coverage for the full 1926 to 2024 period).

For each draw, we centre the selected predictors over the entire $T_{IS}+T_{OOS}$ window and whiten them using a Cholesky decomposition of their sample covariance. This is useful later on when we wish to estimate β for our analytical values. We subsequently add a column of ones to this whitened design matrix to include an intercept in the regression model.

We then estimate the OLS parameter $\hat{\beta}$ using the sampled signals' in-sample data (per Equation (2.2.3)) and test this on the in-sample and out-of-sample periods using the estimated Markowitz portfolio $\hat{\mathbf{w}}_t = \hat{\Sigma}_\epsilon^{-1} \hat{\beta} \mathbf{s}_t$, where $\hat{\Sigma}_\epsilon^{-1}$ is also estimated from the observed in-sample residuals. In this case we cannot know the true covariance matrix Σ_ϵ , but as we are predicting just one asset (the CRSP Index), $\hat{\Sigma}_\epsilon$ is simply the variance of the residuals σ_ϵ^2 , which is relatively easy to estimate (in comparison to a large covariance matrix of many assets).

Parameter Estimation for Analytical Values

This process yields pairs of in-sample and out-of-sample Sharpe ratios $\widehat{SR}_{IS}$ and $\widehat{SR}_{OOS}$. According to our analysis above, the primary drivers of their ratio $\widehat{SR}_{OOS}/\widehat{SR}_{IS}$

should be: the number of signals K , in-sample periods T_{IS} , and true signal strength β for the strategy. Clearly, the number of signals K and backtest length T_{IS} are directly available here, but we will need to estimate β . From our analysis throughout the chapter, we've shown that estimating β is quite difficult.

However, for our needs, the exact value of each entry of β is less important than having the correct magnitude of β and in turn the correct magnitude of the true Sharpe ratio (see Figure 2.3.4). We found in experimentation that using the $\hat{\beta}$ value directly led to the analytical values being poor estimates, and therefore devised the following method to calibrate β as an input to the analytical computations. Specifically, rather than using $\hat{\beta}$ directly, we compute an implied parameter β^* by minimising the difference between the observed out-of-sample Sharpe ratio $\widehat{\text{SR}}_{\text{OOS}}$ and the expected out-of-sample Sharpe ratio SR_{EOOS} :

$$\beta^* = \arg \min_{\beta} \left(\widehat{\text{SR}}_{\text{OOS}} - \text{SR}_{\text{EOOS}}(\mu_s, \Sigma_s, \Sigma_\epsilon, T_{\text{IS}}, \beta) \right)^2.$$

We perform this calibration separately for each resample. We fix the intercept coefficient within β to zero (assuming we predict day-to-day price moves but not the unconditional drift). For the remaining active signals, $\mu_s = \mathbf{0}_p$ and $\Sigma_s = \mathbf{I}_K$ due to our whitening step. Additionally, Σ_ϵ is estimated by $\hat{\Sigma}_\epsilon$ from the particular sample, and T_{IS} is the length of the sample's in-sample period.

For the minimisation, we take a reduced form for β and assume equal predictive power for all the active signals, $\beta = (0, b, b, \dots)$. We do this to further reduce the degrees of freedom. We use a root-finding routine to choose the value of b such that the analytical SR_{EOOS} matches the realised out-of-sample Sharpe ratio for each specific trial. Consequently, we view the resulting β^* as a local calibration of the true Sharpe ratio for that particular set of signals and period.

Once we have this implied β , we can use it to estimate the expected in- and out-of-sample mean, variance and Sharpe ratio, and expected replication ratio for that particular set of signals over that particular sampled time period. We note here

that in the following analysis, all analytical values use this implied β .

Analysis

We recap what we now have in hand. From the process described above we have obtained many sets of observed in- and out-of-sample empirical and analytical metrics. As discussed, the analytical metrics are expectations, while the empirical metrics are simply observed data points for a particular set of signals over a particular period, which will naturally have measurement error and deviate from their (unknown) expected values.

To resolve this, we look at averages for both the empirical and analytical metrics as a function of our identified key variables: number of signals K , length of backtest T_{IS} and true Sharpe ratio SR . By computing the averages as a function of these variables across all the samples, the effect of selecting different signals over different periods, and the finite sample measurement error, should average out, and our average empirical values should approach our average analytical values.

The results of this procedure are displayed in Figure 2.6.1. In each panel, we plot both the expected ratio of the out-of-sample and in-sample Sharpe ratios, and the ratio of their expectations, and the analytical replication ratio. Generally, we see that all the estimates are very close to each other. In the first panel, we observe that indeed the replication ratio decreases considerably as the number of signals K increases.

In the second panel, we plot the replication ratio as a function of T_{IS} . We found that this was not smoothly monotonic (perhaps due to insufficient samples), so we plot both the means per backtest length in our samples, and a smooth fit. We see that in general the three estimates agree, with the replication ratio increasing as the backtest length increases.

The final panel shows the replication ratio as a function of the implied true Sharpe ratio. Again, all three cases show the expected (increasing) behaviour, and

2.7 Conclusion

the empirical value, which corresponds more closely to our analytical estimate (the ratio of the expectations) is indeed closer to our analytical values than the alternative (expectation of the ratios). In addition to the replication ratio displayed, we report the individual metrics (mean, variance and Sharpe ratio) in Appendix 2.E.

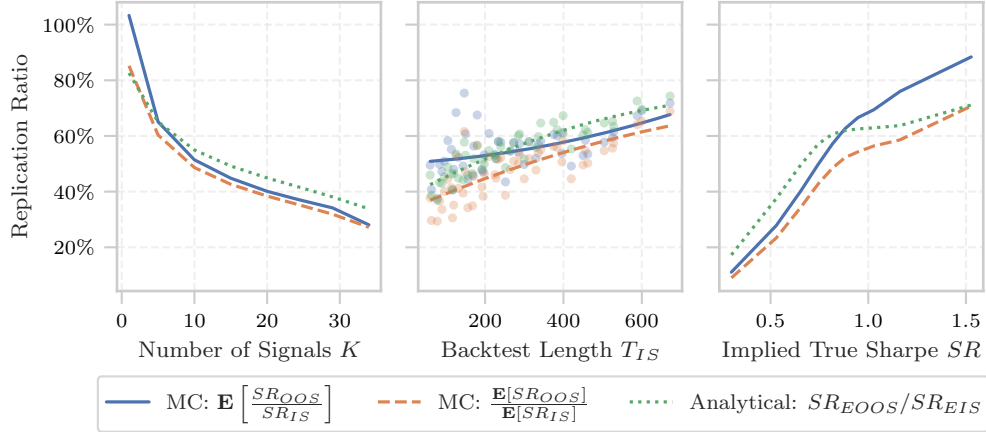


Figure 2.6.1: Empirical and implied analytical replication ratio by number of signals K , backtest length T_{IS} and implied true Sharpe ratio SR (annualised).

2.7 Conclusion

This chapter derives analytical approximations for the in-sample and out-of-sample Sharpe ratios of portfolios constructed using linear prediction models. We show that increasing either the number of signals or assets too much makes this procedure susceptible to overfitting and thereby yields wildly overestimated in-sample Sharpe ratios.

We show that low true Sharpe ratio signals are particularly vulnerable to overfitting. Conversely, by extending the length of the in-sample period one can reduce the overfitting risk, and can produce a higher replication ratio out of sample.

We test our results on commodity futures using momentum-style signals and find that allowing AR(1) signals and non-Normal signals/residuals does not significantly impact the validity of our results. In particular, once we match the theoretical out-

of-sample Sharpe ratio to the observed value, we see that the replication ratio is primarily a function of the out-of-sample Sharpe ratio and the curves for the AR(1) signals closely matches those of iid signals.

From this analysis, it seems that the best way to minimise the potential for overfitting is to minimise the number of signals and assets that are being used for any predictive model used to trade, and utilise the largest amount of data possible. To conclude, we quote Nick Patterson [Pat16] on his experience at Renaissance Technologies, which is very much in line with these results:

It's funny that I think the most important thing to do on data analysis is to do the simple things right. So, here's a kind of non-secret about what we did at Renaissance: in my opinion, our most important statistical tool was simple regression with one target and one independent variable.

They might be on to something.

Chapter Appendix

2.A Derivation of Equation (2.2.2)

To derive (2.2.2) we need to derive the expressions for

$$\mathbb{E}[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top(\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1})] \quad \text{and} \quad \mathbb{V}[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top(\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1})].$$

We start with the former:

$$\begin{aligned} \mathbb{E}[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top(\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1})] &= \mathbb{E}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t + \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] \\ &= \mathbb{E}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t] \\ &= \text{tr}(\mathbf{G}\boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\mu}_s, \end{aligned}$$

where $\mathbf{G} := \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta}$. We have used $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = 0$ in the second equality, and identity 5.24 from [Bro20] in the final equality.

Next we tackle the variance,

$$\begin{aligned} \mathbb{V}[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top(\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1})] &= \mathbb{V}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t + \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] \\ &= \mathbb{V}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t] + \mathbb{V}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] + \text{Cov}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t, \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] \\ &= \mathbb{V}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t] + \mathbb{V}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] \\ &= 2 \text{tr}((\mathbf{G}\boldsymbol{\Sigma}_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\Sigma}_s \mathbf{G}\boldsymbol{\mu}_s + \mathbb{V}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] \\ &= 2 \text{tr}((\mathbf{G}\boldsymbol{\Sigma}_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\Sigma}_s \mathbf{G}\boldsymbol{\mu}_s \\ &\quad + \mathbb{E}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1} \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] - \mathbb{E}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}]^2 \\ &= 2 \text{tr}((\mathbf{G}\boldsymbol{\Sigma}_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\Sigma}_s \mathbf{G}\boldsymbol{\mu}_s + \mathbb{E}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1} \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] \\ &= 2 \text{tr}((\mathbf{G}\boldsymbol{\Sigma}_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\Sigma}_s \mathbf{G}\boldsymbol{\mu}_s + \mathbb{E}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \mathbf{Z}^\top \boldsymbol{\epsilon} \boldsymbol{\epsilon}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t] \\ &= 2 \text{tr}((\mathbf{G}\boldsymbol{\Sigma}_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\Sigma}_s \mathbf{G}\boldsymbol{\mu}_s + \text{tr}(\mathbf{F}\boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{F}\boldsymbol{\mu}_s, \end{aligned}$$

2.B General Case Result of Proposition 2.3.1

where $\mathbf{F} := \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta}$. We have used $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = 0$ to deduce

$$\text{Cov}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t, \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\epsilon}_{t+1}] = 0,$$

identity 379 from [PP12, sec. 8.2.2] to compute $\mathbb{V}[\mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t]$, identity 322 from [PP12, sec. 6.2.2] to compute the penultimate line and identity 5.24 from [Bro20] in the final equality.

Thus,

$$\text{SR} = \frac{\text{tr}(\mathbf{G} \boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{G} \boldsymbol{\mu}_s}{\sqrt{2 \text{tr}((\mathbf{G} \boldsymbol{\Sigma}_s)^2) + 4 \boldsymbol{\mu}_s^\top \mathbf{G} \boldsymbol{\Sigma}_s \mathbf{G} \boldsymbol{\mu}_s + \text{tr}(\mathbf{F} \boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{F} \boldsymbol{\mu}_s}},$$

where

$$\mathbf{F} := \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \quad \text{and} \quad \mathbf{G} := \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta}.$$

2.B General Case Result of Proposition 2.3.1

Proposition 2.B.1. *The expected average in-sample $P\mathcal{E}L$ is*

$$\mathbb{E} \left[\widehat{\mathbb{E}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}] \right] = \text{tr}(\mathbf{G} \boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{G} \boldsymbol{\mu}_s + \frac{K}{T_{\text{IS}}} \text{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon).$$

The expected variance of the in-sample P&L is

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{V}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= 2 \operatorname{tr} \left((\mathbf{G} \boldsymbol{\Sigma}_s)^2 \right) + 4 \boldsymbol{\mu}_s^\top \mathbf{G} \boldsymbol{\Sigma}_s \mathbf{G} \boldsymbol{\mu}_s + \operatorname{tr}(\mathbf{F} \boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{F} \boldsymbol{\mu}_s \\
 &+ \frac{3}{T_{\text{IS}} - K - 1} \operatorname{tr} \left(2 \mathbf{F} \boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{F} \boldsymbol{\mu}_s \cdot \boldsymbol{\Sigma}_{s,+}^{-1} \boldsymbol{\Sigma}_{s,-} + K \mathbf{F} \boldsymbol{\Sigma}_s \right) \\
 &+ \operatorname{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z}) \left(\frac{p(K + T_{\text{IS}} + 4)}{T_{\text{IS}}(T_{\text{IS}} + 2)} - \frac{2 \mathbf{1}_K^\top \boldsymbol{\mu}_s (K - T_{\text{IS}} - 2)(K - T_{\text{IS}})}{T_{\text{IS}}^2 (T_{\text{IS}} + 1)(T_{\text{IS}} + 2)} \right) \\
 &- \operatorname{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z})^2 \frac{2(K - T_{\text{IS}}) \left((K - 2) \mathbf{1}_K^\top \boldsymbol{\mu}_s + T_{\text{IS}}(K - \mathbf{1}_K^\top \boldsymbol{\mu}_s) + K \right)}{T_{\text{IS}}^2 (T_{\text{IS}} + 1)(T_{\text{IS}} + 2)} \\
 &+ \frac{2}{T_{\text{IS}} - K - 1} \operatorname{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon) \operatorname{tr} \left(2 \mathbf{G} \boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{G} \boldsymbol{\mu}_s \cdot \boldsymbol{\Sigma}_{s,+}^{-1} \boldsymbol{\Sigma}_{s,-} + K \mathbf{G} \boldsymbol{\Sigma}_s \right) \\
 &- \frac{2K}{T_{\text{IS}}} \operatorname{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon) \left(\operatorname{tr}(\mathbf{G} \boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{G} \boldsymbol{\mu}_s \right) \\
 &+ \varepsilon_1 + \varepsilon_2 + \varepsilon_3,
 \end{aligned}$$

where $\mathbf{F} = \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta}$, $\mathbf{G} = \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{s,+} = \boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top$, $\boldsymbol{\Sigma}_{s,-} = \boldsymbol{\Sigma}_s - \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top$, and

$$\begin{aligned}
 \varepsilon_1 &= \operatorname{tr} \left(\operatorname{Cov} \left[(\mathbf{S} \mathbf{S}^\top)^{-1}, \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right), \\
 \varepsilon_2 &= \operatorname{tr} \left(\operatorname{Cov} \left[(\mathbf{S} \mathbf{S}^\top)^{-1}, \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right), \\
 \varepsilon_3 &= \operatorname{tr} \left(\operatorname{Cov} \left[(\mathbf{S} \mathbf{S}^\top)^{-1}, \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right),
 \end{aligned}$$

for any $t \in \mathcal{T}_{\text{IS}}$ by the iid property of $(\mathbf{s}_t)_{t \in \mathcal{T}_{\text{IS}}}$.

The expected average out-of-sample P&L is

$$\mathbb{E} \left[\widehat{\mathbb{E}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] = \operatorname{tr}(\mathbf{G} \boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{G} \boldsymbol{\mu}_s,$$

The expected variance of the out-of-sample $P\mathcal{E}L$ is

$$\begin{aligned}
& \mathbb{E} \left[\widehat{\mathbb{V}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] \\
&= 2 \operatorname{tr} ((\mathbf{G}\boldsymbol{\Sigma}_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\Sigma}_s \mathbf{G}\boldsymbol{\mu}_s + \operatorname{tr}(\mathbf{F}\boldsymbol{\Sigma}_s) + \boldsymbol{\mu}_s^\top \mathbf{F}\boldsymbol{\mu}_s \\
&\quad + \boldsymbol{\mu}_s^\top \left(2\mathbf{D}\boldsymbol{\Sigma}_s^\top \mathbf{F} + \operatorname{tr}(\mathbf{F}\boldsymbol{\Sigma}_s^\top) \mathbf{D} + \operatorname{tr}(\boldsymbol{\Sigma}_s^\top \mathbf{D}) \mathbf{F} \right) \boldsymbol{\mu}_s \\
&\quad + \operatorname{tr}(\mathbf{D}\boldsymbol{\Sigma}_s^\top \mathbf{F}\boldsymbol{\Sigma}_s) + \operatorname{tr}(\mathbf{D}\boldsymbol{\Sigma}_s) \operatorname{tr}(\mathbf{F}\boldsymbol{\Sigma}_s^\top) \\
&\quad + \operatorname{tr}((\mathbf{Z}\boldsymbol{\Sigma}_\epsilon)^2) \left(\operatorname{tr}(\boldsymbol{\Sigma}_s^\top \mathbf{D}) + \boldsymbol{\mu}_s^\top \mathbf{D}\boldsymbol{\mu}_s \right),
\end{aligned}$$

where $\mathbf{D} := \frac{(\boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top)^{-1}}{T_{\text{IS}} - K - 1}$.

2.C Proof of Proposition 2.B.1

Recall that the classical Ordinary Least Square method yields the estimator

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1},$$

where $\mathbf{E} \in \mathbb{R}^{N \times T_{\text{IS}}}$ is the matrix of stacked residuals. This is a more convenient expression to work with than (2.2.3), and we will use it to evaluate the expectations of the sample means, variances and covariances. Since $(\mathbf{s}_t)_t$ and $(\boldsymbol{\epsilon}_{t+1})_t$ are iid sequences, the expectation of the sample expectation, variance and covariance are simply the true expectation, variance and covariance respectively.

2.C.1 In-sample expected average

From the definition (2.2.4), the in-sample expected P&L $\mathbb{E}[\widehat{\mathbb{E}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]$ can be evaluated as

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left((\mathbf{Z} \widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top (\boldsymbol{\beta}_{\mathbf{s}_t} + \boldsymbol{\epsilon}_{t+1}) \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\{\mathbf{Z} \widehat{\boldsymbol{\beta}}_{\mathbf{s}_t}\}^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] + \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\{\mathbf{Z} \widehat{\boldsymbol{\beta}}_{\mathbf{s}_t}\}^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\left\{ \mathbf{Z} \left(\boldsymbol{\beta} + \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \right) \mathbf{s}_t \right\}^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &+ \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\left\{ \mathbf{Z} \left(\boldsymbol{\beta} + \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \right) \mathbf{s}_t \right\}^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= \mathbb{E} \left[(\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right] + \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\left\{ (\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right\} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &+ \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\left\{ \mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t \right\}^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right],
 \end{aligned}$$

since the sample mean is an unbiased estimator, and since $(\mathbf{s}_t)_t$ and $(\boldsymbol{\epsilon}_{t+1})_t$ are respectively iid, and $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$.

Since we can “extract” the t -th column from a matrix $\mathbf{X}$ by multiplying it by a vector $\mathbf{e}_t$ which has 1 in the t -th position and zero everywhere else, then

$$\begin{aligned}
 \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{E}^\top \mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t} \right] \\
 &= 0, \quad \text{for all } t \in \mathcal{T}_{\text{IS}},
 \end{aligned}$$

since $\mathbf{E}$ has a Matrix Normal distribution $\mathcal{MN}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon, \mathbf{I}_{T_{\text{IS}}})$ by assumption and since $\mathbf{Z}$ is assumed symmetric. As previously mentioned, the first equality is true for all $t \in \mathcal{T}_{\text{IS}}$ by the iid property of $(\mathbf{s}_t)_t$, and the fact the sample mean is an unbiased

estimator.

The next term reads

$$\begin{aligned}
 \mathbb{E} \left[\widehat{\mathbb{E}} \left[\left(\left(\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t \right)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{E}^\top \mathbf{Z} \mathbf{E} \mathbf{e}_t \right], \quad \forall t \in \mathcal{T}_{\text{IS}}, \\
 &= \mathbf{e}_t^\top \mathbb{E}[\mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S}] \mathbb{E}[\mathbf{E}^\top \mathbf{Z} \mathbf{E}] \mathbf{e}_t, \\
 &= \text{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon) \mathbf{e}_t^\top \mathbb{E}[\mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S}] \mathbf{e}_t \\
 &= \frac{K}{T_{\text{IS}}} \text{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon),
 \end{aligned}$$

where we have used the Matrix Normal identity

$$\mathbb{E}[\mathbf{X} \mathbf{A} \mathbf{X}^\top] = \mathbf{U} \text{tr}(\mathbf{A}^\top \mathbf{V}) + \mathbf{M} \mathbf{A} \mathbf{M}^\top,$$

with $\mathbf{X} \sim \mathcal{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V})$ [Gup00, Theorem 2.3.5]. We also used that

$$\text{tr}(\mathbb{E}[\mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S}]) = K$$

since the projection matrix $\mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S}$ is idempotent and thus its trace equals its rank, and since $(\mathbf{s}_t)_t$ are iid, then the expectation of each element of the diagonal of $\mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S}$ are equal, thus $\mathbf{e}_t^\top \mathbb{E}[\mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{S}] \mathbf{e}_t = \frac{K}{T_{\text{IS}}}$. Therefore, the in-sample expected P&L is given by

$$\mathbb{E} \left[\widehat{\mathbb{E}} \left[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] = \text{tr} \left(\boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z} \right) + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\mu}_s + \frac{K}{T_{\text{IS}}} \text{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon).$$

2.C.2 In-sample expected variance

We turn next to the expected variance $\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]$:

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top (\boldsymbol{\beta}_{\mathbf{s}_t} + \boldsymbol{\epsilon}_{t+1}) \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left(\left(\mathbf{Z} \left(\boldsymbol{\beta} + \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \right) \mathbf{s}_t \right)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right. \right. \right. \\
 &\quad \left. \left. \left. + \left(\mathbf{Z} \left(\boldsymbol{\beta} + \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \right) \mathbf{s}_t \right)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right], \\
 &= \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} + (\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right. \right. \right. \\
 &\quad \left. \left. \left. + (\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} + (\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right], \\
 &= \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] + \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[2 \widehat{\text{Cov}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[2 \widehat{\text{Cov}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[2 \widehat{\text{Cov}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[2 \widehat{\text{Cov}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[2 \widehat{\text{Cov}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &\quad + \mathbb{E} \left[2 \widehat{\text{Cov}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right].
 \end{aligned}$$

Since the signals and residuals are iid, and the sample variance is an unbiased estimator for the population variance, we can replace all the expected sample variance terms with the true variance. We are thus left with computing these true variances. The first two terms are

$$\begin{aligned}
 \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= \mathbb{V} \left[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right] \\
 &= 2 \operatorname{tr} \left(\mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \right) + 4\boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\mu}_s, \\
 \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= \mathbb{V} \left[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right] \\
 &= \operatorname{tr} \left(\mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \right) + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\mu}_s,
 \end{aligned}$$

where the first equality in both expressions is for all $t \in \mathcal{T}_{\text{IS}}$, and the first equality comes from the iid assumption and the fact the sample variance is an unbiased estimator.

The next term is

$$\begin{aligned}
 \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= \mathbb{V} \left[(\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right], \quad \text{for all } t \in \mathcal{T}_{\text{IS}} \\
 &= \mathbb{E} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)^2 \right] - \mathbb{E} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right) \right]^2 \\
 &= \mathbb{E} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)^2 \right],
 \end{aligned}$$

since $\mathbb{E}[\mathbf{E}] = 0$ and $\mathbf{E}$ and $\mathbf{S}$ are independent. To evaluate this further we again use $\mathbf{s}_t = \mathbf{S}\mathbf{e}_t$ and let $\mathbf{H} := \mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{S}$ (the so called ‘hat’ matrix in linear regression),

so that

$$\begin{aligned}
 & \mathbb{E} \left[\left((\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \beta \mathbf{s}_t \right)^2 \right] \\
 &= \mathbb{E} \left[\left(\mathbf{e}_t^\top \mathbf{S}^\top (\mathbf{SS}^\top)^{-1} \mathbf{SE}^\top \mathbf{Z} \beta \mathbf{Se}_t \right)^2 \right] \\
 &= \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{HE}^\top \mathbf{Z} \beta \mathbf{Se}_t \mathbf{e}_t^\top \mathbf{HE}^\top \mathbf{Z} \beta \mathbf{Se}_t \right] \\
 &= \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{H} (\mathbf{I}_{T_{\text{IS}}} \mathbf{H}^\top \mathbf{e}_t \mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \Sigma_\epsilon) \mathbf{Z} \beta \mathbf{Se}_t \right] \\
 &= \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{He}_t \mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \Sigma_\epsilon \mathbf{Z} \beta \mathbf{Se}_t \right], \\
 &= \text{tr} \left(\mathbb{E} \left[\mathbf{e}_t^\top \mathbf{He}_t \mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \Sigma_\epsilon \mathbf{Z} \beta \mathbf{Se}_t \right] \right) \\
 &= \text{tr} \left(\mathbb{E} \left[(\mathbf{SS}^\top)^{-1} \mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \Sigma_\epsilon \mathbf{Z} \beta \mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \right] \right) \\
 &= \text{tr} \left(\mathbb{E} \left[(\mathbf{SS}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \Sigma_\epsilon \mathbf{Z} \beta \mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \right] + \mathcal{E}_1 \right),
 \end{aligned}$$

where

$$\begin{aligned}
 \mathcal{E}_1 &:= \mathbb{E} \left[(\mathbf{SS}^\top)^{-1} \mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \Sigma_\epsilon \mathbf{Z} \beta \mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \right] \\
 &\quad - \mathbb{E} \left[(\mathbf{SS}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \Sigma_\epsilon \mathbf{Z} \beta \mathbf{Se}_t \mathbf{e}_t^\top \mathbf{S}^\top \right].
 \end{aligned}$$

The $\mathcal{E}_1$ term results from the fact that $(\mathbf{SS}^\top)^{-1}$ is of course not independent of the other terms, however experimentally we find that $\text{tr}(\mathcal{E}_1)$ is negligible compared to the trace of the expectation product and decreases rapidly in the ratio K/T_{IS} . Provided we are in the regime where $T_{\text{IS}} \gg p$, then this approximation is reasonable as can be seen in Figure 2.3.2.

With this, we can finish the evaluation. To simplify the expression it is convenient

to swap $\mathbf{S}\mathbf{e}_t$ with $\mathbf{s}_t$ again, and define the matrix $\mathbf{F} := \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta}$, so that

$$\begin{aligned}
 & \mathbb{E} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta} \mathbf{s}_t \right)^2 \right] \\
 &= \text{tr} \left(\mathbb{E} \left[(\mathbf{S} \mathbf{S}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] + \mathcal{E}_1 \right) \\
 &= \text{tr} \left[\frac{\boldsymbol{\Sigma}_{s,+}^{-1}}{T_{\text{IS}} - K - 1} \left(2\boldsymbol{\Sigma}_{s,+} \mathbf{F} \boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{F} \boldsymbol{\mu}_s \boldsymbol{\Sigma}_{s,-} + \text{tr}(\mathbf{F} \boldsymbol{\Sigma}_s) \boldsymbol{\Sigma}_{s,+} \right) \right] + \varepsilon_1 \\
 &= \frac{1}{T_{\text{IS}} - K - 1} \text{tr} \left(2\mathbf{F} \boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{F} \boldsymbol{\mu}_s \cdot \boldsymbol{\Sigma}_{s,+}^{-1} \boldsymbol{\Sigma}_{s,-} + K \mathbf{F} \boldsymbol{\Sigma}_s \right) + \varepsilon_1,
 \end{aligned}$$

with $\varepsilon_1 := \text{tr}(\mathcal{E}_1)$ and $\boldsymbol{\Sigma}_{s,+} = \boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top$, $\boldsymbol{\Sigma}_{s,-} = \boldsymbol{\Sigma}_s - \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top$. Thus, the expected variance term is given by

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta} \mathbf{s}_t \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= \frac{1}{T_{\text{IS}} - K - 1} \text{tr} \left(2\mathbf{F} \boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{F} \boldsymbol{\mu}_s \cdot \boldsymbol{\Sigma}_{s,+}^{-1} \boldsymbol{\Sigma}_{s,-} + K \mathbf{F} \boldsymbol{\Sigma}_s \right) + \varepsilon_1.
 \end{aligned}$$

Moving on, we tackle the next term in the expected variance similarly:

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] = \mathbb{V} \left[(\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right], \quad \text{for all } t \in \mathcal{T}_{\text{IS}}, \\
 &= \mathbb{E} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)^2 \right] - \mathbb{E} \left[(\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right]^2 \\
 &= \mathbb{E} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)^2 \right] - \left(\text{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon) \frac{K}{T_{\text{IS}}} \right)^2.
 \end{aligned}$$

We can evaluate the first term as

$$\begin{aligned}
 & \mathbb{E} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)^2 \right] \\
 &= \mathbb{E} \left[\mathbf{e}_t \mathbf{H} \mathbf{E}^\top \mathbf{Z} \mathbf{E} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{E}^\top \mathbf{Z} \mathbf{E} \mathbf{H}^\top \mathbf{e}_t \right] \\
 &= \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z} \mathbf{Z}^\top \boldsymbol{\Sigma}_\epsilon \mathbf{Z}) \left(\mathbb{E} \left[\mathbf{e}_t^\top \mathbf{H} \mathbf{e}_t \right] + \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{H} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{H}^\top \mathbf{e}_t \right] \right) + \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z})^2 \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{H} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{H} \mathbf{e}_t \right] \\
 &= \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z} \mathbf{Z}^\top \boldsymbol{\Sigma}_\epsilon \mathbf{Z}) \left(\frac{K}{T_{\text{IS}}} + \left(\frac{K(K+2)}{T_{\text{IS}}(T_{\text{IS}}+2)} - \mathbf{1}_K^\top \boldsymbol{\mu}_s \frac{2(T_{\text{IS}}-K)(T_{\text{IS}}-K+2)}{T^2(T+1)(T+2)} \right) \right) \\
 &\quad + \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z})^2 \left(\frac{K(K+2)}{T_{\text{IS}}(T_{\text{IS}}+2)} - \mathbf{1}_K^\top \boldsymbol{\mu}_s \frac{2(T_{\text{IS}}-K)(T_{\text{IS}}-K+2)}{T^2(T+1)(T+2)} \right),
 \end{aligned}$$

where we have used the fact that when there is no intercept term the diagonal entries of the hat matrix $\mathbf{H}$ are distributed $\beta \left(\frac{K}{2}, \frac{(T-K)}{2} \right)$ [BKW04, Appendix 2A], and for the case where there is an intercept term, Chatterjee and Hadi [CH88, Section 2.3.7] derive an approximate distribution for the diagonal elements of the hat matrix,

$$\frac{T_{\text{IS}} - K}{K - 1} \frac{h_{tt} - \frac{1}{T_{\text{IS}}}}{1 - h_{tt}} \sim F(K - 1, T_{\text{IS}} - K),$$

where $F(K - 1, T_{\text{IS}} - K)$ denotes the F distribution with $K - 1$ and $T_{\text{IS}} - K$ degrees of freedom. However, according to Rao [Rao09, p. 542] this is an exact distribution, rather than an approximate one when $\boldsymbol{\mu}_s = \mathbf{0}$, which holds in our assumptions. With a little work this yields

$$\mathbb{E}[h_{tt}^2] = \frac{K(K+2)}{T_{\text{IS}}(T_{\text{IS}}+2)} - \frac{2(T_{\text{IS}}-K)(T_{\text{IS}}-K+2)}{T^2(T+1)(T+2)},$$

and noting that $\mathbf{1}_K^\top \boldsymbol{\mu}_s$ is zero with no intercept, and one with an intercept, this yields the final line of the derivation above. Putting it all together we obtain the

expected variance term

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z}) \left(\frac{K}{T_{\text{IS}}} + \left(\frac{K(K+2)}{T_{\text{IS}}(T_{\text{IS}}+2)} - \mathbf{1}_K^\top \boldsymbol{\mu}_s \frac{2(T_{\text{IS}}-K)(T_{\text{IS}}-K+2)}{T^2(T+1)(T+2)} \right) \right) \\
 &+ \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z})^2 \left(\frac{K(K+2)}{T_{\text{IS}}(T_{\text{IS}}+2)} - \mathbf{1}_K^\top \boldsymbol{\mu}_s \frac{2(T_{\text{IS}}-K)(T_{\text{IS}}-K+2)}{T^2(T+1)(T+2)} \right) - \left(\text{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\epsilon) \frac{K}{T_{\text{IS}}} \right)^2 \\
 &= \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z}) \frac{KT_{\text{IS}}(T_{\text{IS}}+1)(K+T_{\text{IS}}+4) - 2(K-T_{\text{IS}}-2)(K-T_{\text{IS}})\mathbf{1}_K^\top \boldsymbol{\mu}_s}{T_{\text{IS}}^2(T_{\text{IS}}+1)(T_{\text{IS}}+2)} \\
 &- \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z})^2 \frac{2(K-T_{\text{IS}}) \left((K-2)\mathbf{1}_K^\top \boldsymbol{\mu}_s + T_{\text{IS}}(K - \mathbf{1}_K^\top \boldsymbol{\mu}_s) + K \right)}{T_{\text{IS}}^2(T_{\text{IS}}+1)(T_{\text{IS}}+2)}.
 \end{aligned}$$

Next we begin to tackle the covariance terms, again we use the fact that the sample covariance is an unbiased estimator and the iid property. The first two covariance terms are zero:

$$\begin{aligned}
 \mathbb{E} \left[\widehat{2\text{Cov}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] &= 2\text{Cov}[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t}, (\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1}] \\
 &= 0,
 \end{aligned}$$

since $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$, and

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{2\text{Cov}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= 2\text{Cov} \left[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t}, (\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right]
 \end{aligned}$$

is null for all $t \in \mathcal{T}_{\text{IS}}$ as well since $\mathbb{E}[\epsilon_{t+1}] = \mathbf{0}$. Regarding the next covariance term,

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{2\text{Cov}} \left[\left((\mathbf{Z}\beta_{\mathbf{s}_t})^\top \beta_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \epsilon_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= 2\text{Cov} \left[(\mathbf{Z}\beta_{\mathbf{s}_t})^\top \beta_{\mathbf{s}_t}, (\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \epsilon_{t+1} \right], \quad \text{for all } t \in \mathcal{T}_{\text{IS}}, \\
 &= 2\mathbb{E} \left[(\mathbf{Z}\beta_{\mathbf{s}_t})^\top \beta_{\mathbf{s}_t} (\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \epsilon_{t+1} \right] \\
 &\quad - 2\mathbb{E} \left[(\mathbf{Z}\beta_{\mathbf{s}_t})^\top \beta_{\mathbf{s}_t} \right] \mathbb{E} \left[(\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \epsilon_{t+1} \right].
 \end{aligned}$$

Starting with the first expectation,

$$\begin{aligned}
 & \mathbb{E} \left[(\mathbf{Z}\beta_{\mathbf{s}_t})^\top \beta_{\mathbf{s}_t} (\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \epsilon_{t+1} \right] \\
 &= \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \beta \mathbf{S} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{H} \mathbf{E}^\top \mathbf{Z} \mathbf{E} \mathbf{e}_t \right] \\
 &= \text{tr}(\mathbf{Z}\Sigma_\epsilon) \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{S}^\top \beta^\top \mathbf{Z} \beta \mathbf{S} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{H} \mathbf{e}_t \right] \\
 &= \text{tr}(\mathbf{Z}\Sigma_\epsilon) \text{tr} \left(\mathbb{E} \left[(\mathbf{SS}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{s}_t \mathbf{s}_t^\top \beta^\top \mathbf{Z} \beta \mathbf{s}_t \mathbf{s}_t^\top \right] + \mathcal{E}_2 \right) \\
 &= \frac{\text{tr}(\mathbf{Z}\Sigma_\epsilon) \text{tr} \left(2\beta^\top \mathbf{Z} \beta \Sigma_{s,+} + \boldsymbol{\mu}_s^\top \beta^\top \mathbf{Z} \beta \boldsymbol{\mu}_s \Sigma_{s,+}^{-1} \Sigma_{s,-} + K \beta^\top \mathbf{Z} \beta \Sigma_s \right)}{T_{\text{IS}} - K - 1} + \varepsilon_2,
 \end{aligned}$$

where we have proceeded as before with a mixed expectation of $(\mathbf{SS}^\top)^{-1}$ and $\mathbf{S}$, and

$$\mathcal{E}_2 = \mathbb{E} \left[(\mathbf{SS}^\top)^{-1} \mathbf{s}_t \mathbf{s}_t^\top \beta^\top \mathbf{Z} \beta \mathbf{s}_t \mathbf{s}_t^\top \right] - \mathbb{E} \left[(\mathbf{SS}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{s}_t \mathbf{s}_t^\top \beta^\top \mathbf{Z} \beta \mathbf{s}_t \mathbf{s}_t^\top \right],$$

and $\varepsilon_2 := \text{tr}(\mathcal{E}_2)$. Next, the second expectation reads

$$\mathbb{E} \left[(\mathbf{Z}\beta_{\mathbf{s}_t})^\top \beta_{\mathbf{s}_t} \right] \mathbb{E} \left[(\mathbf{ZES}^\top (\mathbf{SS}^\top)^{-1} \mathbf{s}_t)^\top \epsilon_{t+1} \right] = \left(\text{tr}(\beta \Sigma_s \beta^\top \mathbf{Z}) + \boldsymbol{\mu}_s^\top \beta^\top \mathbf{Z} \beta \boldsymbol{\mu}_s \right) \frac{K}{T_{\text{IS}}} \text{tr}(\mathbf{Z}\Sigma_\epsilon).$$

Putting this together yields

$$\begin{aligned}
 & \mathbb{E} \left[2\widehat{\text{Cov}} \left[\left((\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= \frac{2}{T_{\text{IS}} - K - 1} \text{tr}(\mathbf{Z}\boldsymbol{\Sigma}_\epsilon) \text{tr} \left(2\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\mu}_s \cdot \boldsymbol{\Sigma}_{s,+}^{-1} \boldsymbol{\Sigma}_{s,-} + K\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s \right) \\
 & - \frac{2K}{T_{\text{IS}}} \text{tr}(\mathbf{Z}\boldsymbol{\Sigma}_\epsilon) \left(\text{tr}(\boldsymbol{\beta}\boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z}) + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\mu}_s \right) + \varepsilon_2.
 \end{aligned}$$

The next covariance term is similar, namely

$$\begin{aligned}
 & \mathbb{E} \left[2\widehat{\text{Cov}} \left[\left((\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\
 &= 2\text{Cov} \left[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1}, (\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right], \quad \text{for all } t \in \mathcal{T}_{\text{IS}}, \\
 &= 2\mathbb{E} \left[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} (\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right] \\
 & \quad - 2\mathbb{E} \left[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right] \mathbb{E} \left[(\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right] \\
 &= 2\mathbb{E} \left[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} (\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right],
 \end{aligned}$$

since $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$. The first expectation can be evaluated as

$$\begin{aligned}
 & \mathbb{E} \left[(\mathbf{Z}\boldsymbol{\beta}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} (\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right] \\
 &= \mathbb{E} \left[\mathbf{e}_t^\top \mathbf{S}^\top \boldsymbol{\beta}^\top \mathbf{Z} \text{tr}(\mathbf{e}_t^\top \mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{e}_t) \boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\mathbf{S} \mathbf{e}_t \right] \\
 &= \text{tr} \left(\mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] + \mathcal{E}_3 \right) \\
 &= \frac{1}{T_{\text{IS}} - K - 1} \text{tr} \left(2\mathbf{F}\boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{F}\boldsymbol{\mu}_s \cdot \boldsymbol{\Sigma}_{s,+}^{-1} \boldsymbol{\Sigma}_{s,-} + K\mathbf{F}\boldsymbol{\Sigma}_s \right) + \varepsilon_3,
 \end{aligned}$$

where again we handle the $(\mathbf{S}\mathbf{S}^\top)^{-1}$ terms as before, and

$$\mathcal{E}_3 = \mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] - \mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z} \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right],$$

and $\varepsilon_3 := \text{tr}(\mathcal{E}_3)$. Thus,

$$\begin{aligned} & \mathbb{E} \left[2\widehat{\text{Cov}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\ &= \frac{2}{T_{\text{IS}} - K - 1} \text{tr} \left(2\mathbf{F}\boldsymbol{\Sigma}_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{F}\boldsymbol{\mu}_s \cdot \boldsymbol{\Sigma}_{s,+}^{-1} \boldsymbol{\Sigma}_{s,-} + K\mathbf{F}\boldsymbol{\Sigma}_s \right) + \varepsilon_3, \end{aligned}$$

where again $\mathbf{F} = \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}$. We turn now to the penultimate term:

$$\begin{aligned} & \mathbb{E} \left[2\widehat{\text{Cov}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\ &= 2\text{Cov} \left[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1}, (\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right], \quad \text{for all } t \in \mathcal{T}_{\text{IS}}, \\ &= 2\mathbb{E} \left[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} (\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right] \\ &\quad - 2\mathbb{E} \left[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right] \mathbb{E} \left[(\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right] \\ &= 0, \end{aligned}$$

as the derivation yields a cubic in $\mathbf{E}$ and $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$.

For the final term, we can write

$$\begin{aligned} & \mathbb{E} \left[2\widehat{\text{Cov}} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}}, \left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] \\ &= 2\mathbb{E} \left[\widehat{\mathbb{E}} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} (\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right. \\ &\quad \left. - \widehat{\mathbb{E}} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \widehat{\mathbb{E}} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{IS}}} \right] \right] = 0, \end{aligned}$$

which again yields terms with only $\mathbf{E}$ or cubics in $\mathbf{E}$, and from the identity for cubic expectations we see that every term in the expectation includes $\boldsymbol{\mu}_\epsilon$ which is zero, and thus the full expectation is zero. We can now collate all the terms for the expected

variance $\mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]]$:

$$\begin{aligned}
 & \mathbb{E}[\widehat{\mathbb{V}}[(\widehat{\text{PnL}}_t)_{t \in \mathcal{T}_{\text{IS}}}]] \\
 &= 2 \operatorname{tr}((\mathbf{G}\Sigma_s)^2) + 4\boldsymbol{\mu}_s^\top \mathbf{G}\Sigma_s \mathbf{G}\boldsymbol{\mu}_s + \operatorname{tr}(\mathbf{F}\Sigma_s) + \boldsymbol{\mu}_s^\top \mathbf{F}\boldsymbol{\mu}_s \\
 &+ \frac{3}{T_{\text{IS}} - K - 1} \operatorname{tr}\left(2\mathbf{F}\Sigma_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{F}\boldsymbol{\mu}_s \cdot \Sigma_{s,+}^{-1} \Sigma_{s,-} + K\mathbf{F}\Sigma_s\right) \\
 &+ \operatorname{tr}(\Sigma_\epsilon \mathbf{Z}\Sigma_\epsilon \mathbf{Z}) \frac{KT_{\text{IS}}(T_{\text{IS}} + 1)(K + T_{\text{IS}} + 4) - 2(K - T_{\text{IS}} - 2)(K - T_{\text{IS}}) \mathbf{1}_K^\top \boldsymbol{\mu}_s}{T_{\text{IS}}^2 (T_{\text{IS}} + 1)(T_{\text{IS}} + 2)} \\
 &- \operatorname{tr}(\Sigma_\epsilon \mathbf{Z})^2 \frac{2(K - T_{\text{IS}}) \left((K - 2) \mathbf{1}_K^\top \boldsymbol{\mu}_s + T_{\text{IS}}(K - \mathbf{1}_K^\top \boldsymbol{\mu}_s) + K \right)}{T_{\text{IS}}^2 (T_{\text{IS}} + 1)(T_{\text{IS}} + 2)} \\
 &+ \frac{2}{T_{\text{IS}} - K - 1} \operatorname{tr}(\mathbf{Z}\Sigma_\epsilon) \operatorname{tr}\left(2\mathbf{G}\Sigma_{s,+} + \boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\mu}_s \cdot \Sigma_{s,+}^{-1} \Sigma_{s,-} + K\mathbf{G}\Sigma_s\right) \\
 &- \frac{2K}{T_{\text{IS}}} \operatorname{tr}(\mathbf{Z}\Sigma_\epsilon) \left(\operatorname{tr}(\mathbf{G}\Sigma_s) + \boldsymbol{\mu}_s^\top \mathbf{G}\boldsymbol{\mu}_s \right) \\
 &+ \varepsilon_1 + \varepsilon_2 + \varepsilon_3,
 \end{aligned}$$

where

$$\begin{aligned}
 \mathbf{F} &:= \boldsymbol{\beta}^\top \mathbf{Z}\Sigma_\epsilon \mathbf{Z}\boldsymbol{\beta} \quad \text{and} \quad \mathbf{G} := \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}, \\
 \Sigma_{s,+} &:= \Sigma_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top, \quad \text{and} \quad \Sigma_{s,-} := \Sigma_s - \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top
 \end{aligned}$$

and

$$\begin{aligned}
 \varepsilon_1 &= \operatorname{tr} \left(\mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{S}^\top \boldsymbol{\beta}^\top \mathbf{Z}\Sigma_\epsilon \mathbf{Z}\boldsymbol{\beta} \mathbf{S} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{S}^\top \right] \right. \\
 &\quad \left. - \mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{S} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{S}^\top \boldsymbol{\beta}^\top \mathbf{Z}\Sigma_\epsilon \mathbf{Z}\boldsymbol{\beta} \mathbf{S} \mathbf{e}_t \mathbf{e}_t^\top \mathbf{S}^\top \right] \right), \\
 \varepsilon_2 &= \operatorname{tr} \left(\mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] - \mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right), \\
 \varepsilon_3 &= \operatorname{tr} \left(\mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z}\mathbf{Z}\Sigma_\epsilon \mathbf{Z}\mathbf{T}\boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] - \mathbb{E} \left[(\mathbf{S}\mathbf{S}^\top)^{-1} \right] \mathbb{E} \left[\mathbf{s}_t \mathbf{s}_t^\top \boldsymbol{\beta}^\top \mathbf{Z}\mathbf{Z}\Sigma_\epsilon \mathbf{Z}\mathbf{T}\boldsymbol{\beta} \mathbf{s}_t \mathbf{s}_t^\top \right] \right).
 \end{aligned}$$

2.C.3 Out-of-sample expected average

Next we derive the expected out-of-sample expected return and variance. We now have that $\widehat{\boldsymbol{\beta}}$ is independent of the realisations of the signal and the residuals, thus

the derivation is simplified significantly. Firstly, the expected out-of-sample P&L is given by

$$\mathbb{E} \left[\widehat{\mathbb{E}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top (\boldsymbol{\beta}_{\mathbf{s}_t} + \boldsymbol{\epsilon}_{t+1}) \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] = \mathbb{E}[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t}] = \text{tr}(\boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z}) + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\mu}_s,$$

where we have no $\mathbb{E}[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1}]$ term since $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$ and by independence.

2.C.4 Out-of-sample expected variance

Next the variance is given by

$$\begin{aligned} & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top (\boldsymbol{\beta}_{\mathbf{s}_t} + \boldsymbol{\epsilon}_{t+1}) \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] \\ &= \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] + \widehat{\mathbb{V}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right. \\ & \quad \left. + 2\widehat{\text{Cov}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}}, \left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right]. \end{aligned}$$

The first term can be evaluated as

$$\begin{aligned} & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] \\ &= \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] + \widehat{\mathbb{V}} \left[\left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right. \\ & \quad \left. + 2\widehat{\text{Cov}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}}, \left((\mathbf{Z}\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right], \end{aligned}$$

where we must be careful to note that $\mathbf{E}\mathbf{S}^\top (\mathbf{S}\mathbf{S}^\top)^{-1}$, i.e. the error in $\widehat{\boldsymbol{\beta}}$, are the errors and signals from the first period, not the second. The first term of this is simply given by

$$\begin{aligned} \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] &= \mathbb{V}[(\mathbf{Z}\boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t}] \\ &= 2 \text{tr}(\mathbf{Z}\boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top) + 4 \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\mu}_s, \end{aligned}$$

where the first equality holds for all $t \in \mathcal{T}_{\text{OOS}}$ since $\mathbf{s}_t$ are iid. The covariance term is zero since $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$, so we are left with $\mathbb{E} \left[\widehat{\mathbb{V}} \left[((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right]$.

With $\boldsymbol{\epsilon}_\beta := \mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}$,

$$\begin{aligned} & \mathbb{E} \left[\widehat{\mathbb{V}} \left[((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] = \mathbb{V} \left[(\mathbf{Z}\boldsymbol{\epsilon}_\beta\mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t \right], \quad \text{for all } t \in \mathcal{T}_{\text{OOS}}, \\ & = \mathbb{E} \left[\text{tr} \left(\boldsymbol{\epsilon}_\beta^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s \left(\boldsymbol{\epsilon}_\beta^\top \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\epsilon}_\beta \right) \boldsymbol{\Sigma}_s \right) \right. \\ & \quad \left. + \boldsymbol{\mu}_s^\top \left(\boldsymbol{\epsilon}_\beta^\top \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\epsilon}_\beta \right) \boldsymbol{\Sigma}_s \left(\boldsymbol{\epsilon}_\beta^\top \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\epsilon}_\beta \right) \boldsymbol{\mu}_s \right]. \end{aligned}$$

To evaluate this we require some intermediate expectations,

$$\begin{aligned} \mathbb{E} \left[\boldsymbol{\epsilon}_\beta^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s \boldsymbol{\epsilon}_\beta^\top \mathbf{Z} \right] &= \mathbf{D}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}, \\ \mathbb{E} \left[\boldsymbol{\epsilon}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\epsilon} \right] &= \mathbf{D} \text{tr} \left(\boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta} \right), \\ \mathbb{E} \left[\boldsymbol{\epsilon}_\beta \boldsymbol{\Sigma}_s \boldsymbol{\epsilon}_\beta^\top \right] &= \boldsymbol{\Sigma}_\epsilon \text{tr} \left(\boldsymbol{\Sigma}_s^\top \mathbf{D} \right), \end{aligned}$$

where

$$\mathbf{D} := \frac{(\boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^\top)^{-1}}{T_{\text{IS}} - K - 1}.$$

Putting this all together yields

$$\begin{aligned} & \mathbb{E} \left[\widehat{\mathbb{V}} \left[((\mathbf{Z}\mathbf{E}\mathbf{S}^\top(\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{s}_t)^\top \boldsymbol{\beta}\mathbf{s}_t)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] \\ & = \text{tr} \left(\mathbf{D}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s \right) + \text{tr} \left(\mathbf{D}\boldsymbol{\Sigma}_s \right) \text{tr} \left(\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \right) \\ & \quad + \boldsymbol{\mu}_s^\top \left(2\mathbf{D}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta} + \text{tr} \left(\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \right) \mathbf{D} + \text{tr} \left(\boldsymbol{\Sigma}_s^\top \mathbf{D} \right) \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta} \right) \boldsymbol{\mu}_s. \end{aligned}$$

Thus, the first term for the expected variance out-of-sample is

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\beta}_{\mathbf{s}_t} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] \\
 &= 2 \operatorname{tr}(\mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top) + 4 \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\mu}_s \\
 &+ \operatorname{tr} \left(\mathbf{D} \boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s \right) + \operatorname{tr}(\mathbf{D} \boldsymbol{\Sigma}_s) \operatorname{tr} \left(\boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \right) \\
 &+ \boldsymbol{\mu}_s^\top \left(2 \mathbf{D} \boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} + \operatorname{tr} \left(\boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \right) \mathbf{D} + \operatorname{tr} \left(\boldsymbol{\Sigma}_s^\top \mathbf{D} \right) \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \right) \boldsymbol{\mu}_s.
 \end{aligned}$$

We can proceed similarly for the second term,

$$\begin{aligned}
 & \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] \\
 &= \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] + \widehat{\mathbb{V}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \\
 &\quad + 2 \widehat{\operatorname{Cov}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}}, \left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right].
 \end{aligned}$$

Again the first term is simply

$$\mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] = \mathbb{V} \left[(\mathbf{Z} \boldsymbol{\beta}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right] = \operatorname{tr}(\mathbf{Z} \boldsymbol{\beta} \boldsymbol{\Sigma}_s \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon) + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\beta} \boldsymbol{\mu}_s,$$

and the final term is zero as $\mathbf{E}$ is from the in-sample period, and therefore independent of $\boldsymbol{\epsilon}_{t+1}$ from the out-of-sample period, and $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$. Thus, we need only to find the second term,

$$\begin{aligned}
 \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z} \mathbf{E} \mathbf{S}^\top (\mathbf{S} \mathbf{S}^\top)^{-1} \mathbf{s}_t)^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] &= \mathbb{E} \left[\operatorname{tr}(\mathbf{Z} \boldsymbol{\epsilon}_\beta \boldsymbol{\Sigma}_s \boldsymbol{\epsilon}_\beta^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon) + \boldsymbol{\mu}_s^\top \boldsymbol{\epsilon}_\beta^\top \mathbf{Z} \boldsymbol{\Sigma}_\epsilon \mathbf{Z} \boldsymbol{\epsilon}_\beta \boldsymbol{\mu}_s \right] \\
 &= \operatorname{tr}((\mathbf{Z} \boldsymbol{\Sigma}_\epsilon)^2) \left(\operatorname{tr}(\boldsymbol{\Sigma}_s^\top \mathbf{D}) + \boldsymbol{\mu}_s^\top \mathbf{D} \boldsymbol{\mu}_s \right).
 \end{aligned}$$

And therefore, the full second term is given by

$$\begin{aligned} \mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top \boldsymbol{\epsilon}_{t+1} \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] &= \text{tr}(\mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon) + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\mu}_s \\ &\quad + \text{tr}((\mathbf{Z}\boldsymbol{\Sigma}_\epsilon)^2) \left(\text{tr}(\boldsymbol{\Sigma}_s^\top \mathbf{D}) + \boldsymbol{\mu}_s^\top \mathbf{D}\boldsymbol{\mu}_s \right). \end{aligned}$$

The third term, the covariance term, is necessarily zero by $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = \mathbf{0}$ and independence, thus the expected out-of-sample variance is given by

$$\begin{aligned} &\mathbb{E} \left[\widehat{\mathbb{V}} \left[\left((\mathbf{Z}\widehat{\boldsymbol{\beta}}_{\mathbf{s}_t})^\top (\boldsymbol{\beta}\mathbf{s}_t + \boldsymbol{\epsilon}_{t+1}) \right)_{t \in \mathcal{T}_{\text{OOS}}} \right] \right] \\ &= 2 \text{tr}(\mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top) + 4\boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\mu}_s \\ &\quad + \text{tr}(\mathbf{D}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s) + \text{tr}(\mathbf{D}\boldsymbol{\Sigma}_s) \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}) \\ &\quad + \boldsymbol{\mu}_s^\top \left(2\mathbf{D}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta} + \text{tr}(\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}) \mathbf{D} + \text{tr}(\boldsymbol{\Sigma}_s^\top \mathbf{D}) \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta} \right) \boldsymbol{\mu}_s \\ &\quad + \text{tr}(\mathbf{Z}\boldsymbol{\beta}\boldsymbol{\Sigma}_s\boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon) + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}^\top \mathbf{Z}\boldsymbol{\Sigma}_\epsilon \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\mu}_s + \text{tr}((\mathbf{Z}\boldsymbol{\Sigma}_\epsilon)^2) \left(\text{tr}(\boldsymbol{\Sigma}_s^\top \mathbf{D}) + \boldsymbol{\mu}_s^\top \mathbf{D}\boldsymbol{\mu}_s \right). \end{aligned}$$

2.D Monte Carlo Details

As previously mentioned, we use Monte Carlo simulations both to check that the approximations (2.3.2) are reasonable ($\varepsilon_i = 0$ and ignoring the convexity adjustment) and to support and provide additional analysis. To do this, we observe the point estimates of the parameters from the experiments using commodity futures in Section 2.5 and select distributions to draw new realisations within a cloud around these point estimates. Specifically,

- For the signals we let the standard deviation of the signals be 1, which is unrestrictive as we can always standardise any given signal.
- For the residuals we sample the variances from a χ^2 distribution, which is well known to be the sample distribution for variance.
- For both signals and residuals we draw correlation matrices from the LJK

distribution [LKJ09], commonly used as a prior for correlation matrices in Bayesian modelling, and which can be considered the correlation equivalent of the Wishart distribution used for covariance matrices.

- We can then construct the covariance matrix $\Sigma = \text{diag}(\sigma)\mathbf{P}\text{diag}(\sigma)$, where σ is the vector of variances per signal/residual and $\mathbf{P}$ is the sampled correlation matrix for the signal/residuals.
- For β we sample entries from a Laplace distribution, which provided the best fit when looking at the entries of real β matrices from Section 2.5.

With β and $\Sigma_\epsilon, \Sigma_s$ coming from real data we would expect certain structures such as clusters with the covariance matrix (e.g. similarity between stocks in the same sector), or similar parameters for certain stocks; this random sampling unfortunately misses out on this. However, from studying the expressions for the replication ratio it does not appear that these characteristics would impact our results. This is further confirmed by noting that when using real assets and signals in Section 2.5, the results remain valid. To achieve a desired true Sharpe ratio we simply scale the matrix β appropriately. Finally, we let $\mu_s = \mathbf{0}$ and do not include a drift term.

We note here that to simulate even 100 assets requires a significant amount of compute, particularly with large T , which realistically will almost always be the case for a strategy of this style. Thus, the analytical approximations are useful to enable quick comparative adjustments for strategies, without requiring lengthy simulations.

2.E Additional Values for Section 2.5 and Section 2.6

Section 2.5 - Estimated Covariance

In Section 2.5 we have used the true covariance matrix when constructing our portfolios. To further confirm that the values are robust to the estimation of the covariance matrix (as previously shown in Section 2.3.3), we run the same simulation again,

but this time estimating the covariance matrix for each sample from the observed in-sample data, and using that to construct the portfolio $\hat{\mathbf{w}} = \hat{\Sigma}_\epsilon^{-1} \hat{\beta} \mathbf{s}_t$.

We collect all the metrics for the analytical, Gaussian + IID simulated estimates, and $t + \text{AR}(1)$ simulated estimates into Table 2.E.1. Overall, we see that replacing the true residual covariance matrix with its in-sample estimate has only a negligible impact on both in-sample and out-of-sample performance, in line with the sensitivity analysis in Section 2.3.3. Additionally, we see that the average realised Sharpe ratio is extremely close to the estimate which does not account for convexity (yielded by dividing the average mean by the square root of the average variance), which is in turn close to our analytical value. Thus, to conclude this section, we see that even when the data generating process departs from our idealised assumption, the analytical estimates are still a good estimator for all the metrics of interest.

		Gaussian + IID		$t + \text{AR}(1)$	
	Analytical	True Cov	Est. Cov	True Cov	Est. Cov
<i>In-Sample</i>					
Mean μ	0.28	0.28	0.29	0.31	0.32
Variance σ^2	0.30	0.30	0.30	0.41	0.38
Sharpe Ratio $\widehat{\text{SR}}_{\text{IS}}$	N/A	8.24	8.26	7.95	8.12
μ/σ (SR_{EIS})	8.22	8.24	8.26	7.72	8.10
<i>Out-of-Sample</i>					
Mean μ	0.11	0.11	0.11	0.14	0.14
Variance σ^2	0.29	0.29	0.30	0.59	0.62
Sharpe Ratio $\widehat{\text{SR}}_{\text{OOS}}$	N/A	3.14	3.13	2.83	2.83
μ/σ (SR_{EOOS})	3.13	3.14	3.13	2.79	2.79
<i>Replication Ratio</i>					
$\widehat{\text{SR}}_{\text{OOS}}/\widehat{\text{SR}}_{\text{IS}}$	N/A	38%	38%	36%	35%
$\text{SR}_{\text{EOOS}}/\text{SR}_{\text{EIS}}$	38%	38%	38%	36%	34%

Table 2.E.1: Simulation Results comparing Analytical and Estimated values under Gaussian IID and $\text{AR}(1)+t$ assumptions for Section 2.5. Sharpe ratios are annualised assuming 252 trading days.

Section 2.6 - Full Metrics

We report here the expected metrics in the same style as the previous section. Note that the results reported are averages over all 5,000,000 samples. For example, the interpretation of the empirical expected in-sample mean is: “The average observed in-sample mean for all combinations of signals over all sub-periods”, and the interpretation of the analytical expected in-sample mean is: “The average expected in-sample mean for all combinations of signals over all sub-periods”.

We observe that while the raw mean and variance estimates differ significantly between the analytical and empirical results, this discrepancy is primarily attributable to a scale factor. Most notably, the large empirical out-of-sample variance of 44.58 (compared to the analytical 0.27) is driven by significant kurtosis in the empirical values; for comparison, the modal values are 0.05 and 0.09 and the median values are 0.49 and 0.12 respectively. In particular, some resamples generate very large realised variances for portfolios whose corresponding out-of-sample Sharpe ratios are essentially zero. These extreme observations have little impact on the average Sharpe ratio but inflate the average variance. As the portfolio weights can be multiplied by a constant to scale both the mean and variance without altering the risk-adjusted return, the scale-invariant Sharpe ratio provides a more robust metric for comparison.

In this regard, the analytical and empirical estimates align remarkably well. The in-sample expected analytical Sharpe ratio $\mathbb{E}[\text{SR}_{\text{EIS}}] = 1.16$ is comparable to the derived empirical ratio of 1.21. Similarly, the out-of-sample estimates are similar for the analytical and empirical values: the analytical model predicts $\mathbb{E}[\text{SR}_{\text{EOOS}}] = 0.18$, which matches the average observed $\mathbb{E}[\widehat{\text{SR}}_{\text{OOS}}] = 0.18$. This alignment results in consistent replication ratios across the different aggregation methods, ranging between 46% and 54%.

	Analytical	Empirical
<i>In-Sample</i>		
Mean $\mathbb{E}[\mu]$	0.19	0.33
Variance $\mathbb{E}[\sigma^2]$	0.32	0.87
Sharpe Ratio $\mathbb{E}[\widehat{\text{SR}}_{\text{IS}}]$	N/A	1.33
μ/σ ($\mathbb{E}[\text{SR}_{\text{EIS}}]$)	1.16	1.21
<i>Out-of-Sample</i>		
Mean $\mathbb{E}[\mu]$	0.09	0.41
Variance $\mathbb{E}[\sigma^2]$	0.27	44.58
Sharpe Ratio $\mathbb{E}[\widehat{\text{SR}}_{\text{OOS}}]$	N/A	0.18
μ/σ ($\mathbb{E}[\text{SR}_{\text{EOOS}}]$)	0.18	0.21
<i>Replication Ratio</i>		
$\mathbb{E}[\widehat{\text{SR}}_{\text{OOS}}/\widehat{\text{SR}}_{\text{IS}}]$	N/A	52%
$\mathbb{E}[\widehat{\text{SR}}_{\text{OOS}}]/\mathbb{E}[\widehat{\text{SR}}_{\text{IS}}]$	N/A	46%
$\mathbb{E}[\text{SR}_{\text{EOOS}}/\text{SR}_{\text{EIS}}]$	54%	N/A
$\mathbb{E}[\text{SR}_{\text{EOOS}}]/\mathbb{E}[\text{SR}_{\text{EIS}}]$	52%	N/A

Table 2.E.2: Statistics from resampling study, comparing analytical (implied) and empirical values for Section 2.6. Sharpe ratios are annualised from monthly values.

3 | Correcting In-Sample Sharpe Ratios for Selection Bias and Alpha Decay

3.1	Introduction	84
3.2	Bayesian Model: Theory & Implementation	89
3.3	Empirical Case Study: Hedge Fund Returns	99
3.4	Conclusion	114
	Appendices	115
3.A	Extensions, Other Cases & Additional Theory	115
3.B	Implementation & Robustness	123
3.C	Supplementary Material for Section 3.3	137
3.D	Illustrative Examples	162
3.E	Comparison to Chen and Zimmermann	169
3.F	Proofs	179

Chapter Abstract

We propose a Bayesian framework for estimating the true Sharpe ratio of investments. This allows one to explicitly account for two key drivers of the disparity between historical and future performance: selection bias and alpha decay. By modelling the joint distribution of in-sample and out-of-sample Sharpe ratios, we are able to disentangle these effects and correct for the inflation present in historical track records. In a simulation study, we find that fitting the model to in-sample data alone yields poor parameter estimates, whereas a joint fitting approach successfully recovers the underlying parameters. A case study using a comprehensive dataset

of hedge fund returns illustrates the substantial quantitative importance of these effects for real data. We find that our method yields accurate estimates of post-listing performance that significantly help to improve fund selection. In particular, we show that a bias-aware approach is essential for making robust allocation decisions, specifically in the post-2008 regime where selection effects and alpha decay have become more pronounced. This chapter is based on the pre-print paper “Bayes Is All You Need: Better Estimates of True Sharpe Ratios” [Mul+26].

3.1 Introduction

One of the fundamental questions in finance is how to allocate capital. Investors seek to select investments that are likely to perform well in the future, and often make this choice based on attractive historical performance. The issue is that this historical performance is often not representative of future returns: the wedge between in-sample and out-of-sample performance is often dramatic, cf. Figure 3.1.1. As the disclaimer on every fund prospectus advises: “Past performance does not guarantee future results” [Sec02]. But why is this the case? And what can we do to correct for this? In the present chapter, we propose a Bayesian method which provides a principled solution to this problem, capturing two of the main effects that cause this disparity between historical and future performance: “selection bias” and “alpha decay”. In particular, we show that a method that unifies both effects is necessary to make accurate inferences about future returns.

“Selection bias” is a general phenomenon that takes a number of forms. One prominent instance is the multiple testing problem: researchers test many hypotheses before selecting the significant ones, cf., e.g., [Bai+14; BL14; HL14; HL15; HLZ16; HLS20; HL20; HL21a; LB21; GLX21]. Another well-known example is the backfill bias present in hedge fund databases: funds opportunistically list their historical returns in commercial databases when they have performed well, causing their pre-listing returns to be inflated relative to post-listing performance, cf., e.g.,

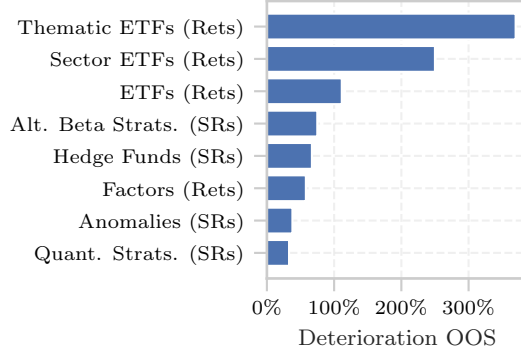


Figure 3.1.1: Performance deterioration between in-sample and out-of-sample performance in different domains. Sources are: Thematic ETFs & Sector ETFs [Ben+23] (estimated), ETFs [BL15], Alternative Beta Strategies [SLP17], Hedge Funds Appendix 3.C.2, Factors [MP16], Anomalies Appendix 3.E and Quantitative Strategies [CV23b].

[FH00; Par95; BGP99; Pv03; FH09; BGR14; JS19].

The second effect, “alpha decay”, is the tendency for investments to underperform out-of-sample beyond the reduction attributable to selection bias alone. [MP16] study this by comparing returns generated between the end of the in-sample period and the publication date against post-publication returns. This decay is often significant. For example, using the open-source asset pricing dataset [CZ22], we find pre-publication Sharpe ratios are 11% lower than in-sample figures, while post-publication Sharpe ratios are 38% lower (Appendix 3.E). Similar dynamics appear in the fund literature where increased scale (assets under management) erodes alpha [BGS22]. In practitioner research, crowding in popular trades has also been linked to performance degradation [MAN19].

In the present chapter, we provide a unified framework to disentangle and quantify both of these effects. This is non-trivial, as a decline in out-of-sample performance can a priori be attributed to either. To overcome this, we use an empirical Bayesian approach that explicitly accounts for selection bias and alpha decay during the fitting step. Specifically, we view the observed Sharpe ratios as the outcome of a selection mechanism: each investment has a true Sharpe ratio SR , with an in-sample

estimate $\widehat{\text{SR}}_{\text{IS}}$, and is included in the dataset with some probability that increases in $\widehat{\text{SR}}_{\text{IS}}$. Conditional on inclusion, we then observe out-of-sample performance from a possibly degraded (alpha-decayed) true Sharpe ratio. Modelling this process allows us to back out the unobserved distribution of the investments that failed the filter, and thereby correct for the inflation in historical track records. Once the model has been fit, it can then be used to perform inference on either the dataset used to fit it, or new candidate investments. The resulting posterior means are then corrected for in-sample measurement error not only through standard Bayesian shrinkage to the prior, but also by explicitly controlling for selection bias.

This methodology has broad applicability across the investment industry. It can aid fund-of-funds and pension funds in selecting managers for capital allocation, assist multi-strategy firms in deciding which managers to launch, retain, or terminate, and support quantitative researchers in evaluating which strategies deserve continued inclusion in their portfolios.

While conceptually sound, this approach presents thorny numerical challenges that make it impractical to implement using naive approaches. To wit, the lack of closed-form expressions necessitates computing complex nested integrals across potentially thousands of sample Sharpe ratios. We overcome this bottleneck using a custom JIT-compiled implementation, rendering the required computations feasible for both fitting and inference across large datasets.

To demonstrate the practical applicability of our approach, we test it on a large dataset of hedge fund returns [Joe+21], covering six major commercial hedge fund databases. We find that the method yields significantly more accurate estimates of individual hedge funds' post-listing performance compared to a number of benchmarks (e.g., a 40% reduction in MSE relative to taking the pre-listing performance as an estimate for future performance) and far better calibration of predictive distributions (the probability integral transforms are statistically indistinguishable from $\text{Uniform}(0, 1)$ under Kolmogorov–Smirnov tests).

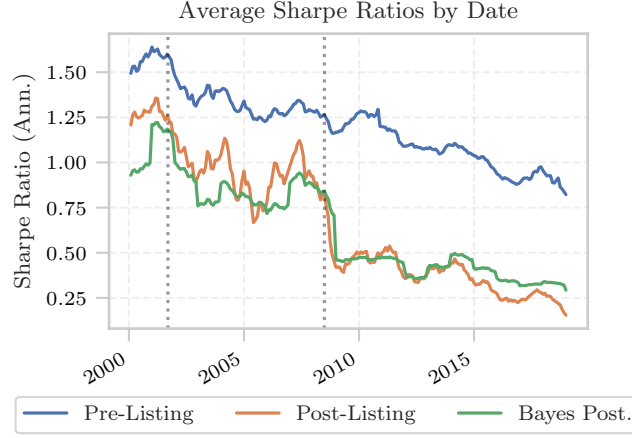


Figure 3.1.2: Average pre- and post-listing Sharpe ratios over time, with a 15-year rolling lookback, and the average Bayesian posterior mean for each date.

We further demonstrate the model’s utility in fund selection, showing that in periods where selection bias and alpha decay are prevalent, such as post-2008, a bias-aware approach is essential for making well-calibrated investment decisions. In Figure 3.1.2 we see how significant the issue has become since the 2008 financial crisis, and how well the Bayesian estimated average posterior means captures this dynamic, closely tracking the average post-listing performance.

Bayesian approaches have a rich history in the finance literature [FS86; Jor86; BMW01; JKP23]. In the fund literature, they have been applied by [PS02] to estimate mutual fund alphas by incorporating information from non-benchmark passive assets, a method [KNT07] extend to hedge funds to overcome short return histories. Empirical Bayesian approaches in particular have previously been used by [HL18] in the estimation of fund alphas. More recently, they have also been applied in the estimation of publication bias in the asset pricing literature [CZ20; HL21b].¹

In our model, we find in a simulation study that fitting to in-sample data alone

¹ A primary advantage of Bayesian frameworks is that, under specific conditions, they are ‘immune to selection bias’ [Daw94; Yek12]. However, this immunity relies on the prior’s parameters being estimated accurately. If the prior is misspecified, for example, through an incorrect location or variance, any resulting inference will still be compromised. For further discussion on this, see Section 3.2.1 and Appendix 3.A.

yields poor parameter estimates, leading to incorrect assessments of selection bias. We address this by fitting in- and out-of-sample performance jointly, producing substantially more accurate estimates. However, the performance decline out-of-sample is not solely attributable to selection bias. If the model does not account for alpha decay, it is forced to ascribe the entire performance drop to selection bias, resulting in poor fits and unrealistic parameters. Incorporating alpha decay not only resolves this issue, but also yields insights into the sources of performance degradation and their relative importance. For example, in the hedge fund data studied in Section 3.3, the level of alpha decay has increased from a relatively moderate level (87% of in-sample true performance maintained out-of-sample) to quite severe (55%, in the post 2008 period).

Although we apply our method to commercial hedge fund data here, the approach has natural applications to many areas of finance, such as quantitative strategy development. The classical challenge in this domain is overfitting. Although studied from various perspectives [BL14; HLS20; CJS25; MJM25], existing frameworks often require additional inputs, such as how the strategy was constructed or an effective number of trials/tests, which are frequently unavailable in practice.

Our Bayesian approach avoids relying on these inputs, which are often unavailable, in the spirit of [LL19; LLZ25]. Instead, we parameterise a reduced-form selection mechanism and fit the model to a dataset of reasonably comparable previously launched strategies with known in- and out-of-sample performance. In this way, the model can encode the estimated degree of overfitting in the parameters of the fitted prior. For example, one could fit the model to a dataset of already launched mid-frequency equity strategies and use it to assess new, similar candidate strategies.

The remainder of the chapter is structured as follows: We provide a brief description of the preliminaries, followed by the model specification and a guide to application, before presenting the results of applying the model to our dataset of hedge fund returns. The appendix presents significant extensions, robustness checks, and

details on the supplementary and exploratory steps required for the empirical study, alongside illustrative examples of how parameter variation affects inference and an additional application of the method to the open-source asset pricing dataset [CZ22].

3.2 Bayesian Model: Theory & Implementation

Our goal is to accurately estimate true Sharpe ratios, corrected for selection bias and alpha decay. In this section we set out the necessary prerequisites and describe both the model we propose to achieve this goal, and how to apply it. This methodology is applicable to a wide variety of capital allocation tasks, ranging from sovereign wealth and pension funds allocating to hedge funds, multi-strategy firms allocating to portfolio managers, or quantitative traders choosing among algorithmic trading strategies. To illustrate our framework, we use a commercially available large dataset of hedge fund returns. For clarity and parsimony, the main text develops a single, illustrative version of our Bayesian model tailored to this particular application. However, the framework is highly flexible and easily applied to a wide variety of investments; we discuss several extensions and alternative specifications in Appendix 3.A, which we draw particular attention to for quantitative researchers wishing to model the effect of parameter tuning from a Bayesian selection perspective.

3.2.1 Preliminaries

Types of Sharpe Ratios

As discussed above, in-sample (historical) performance is often inflated relative to later out-of-sample (future) performance. But, in both the in- and out-of-sample periods, the performance is a noisy measurement of the investment's true, underlying performance. We therefore distinguish between three types of Sharpe ratio:

1. **The True Sharpe Ratio SR:** The true, unobserved, performance of an

investment,

2. **The In-Sample Sharpe Ratio $\widehat{\text{SR}}_{\text{IS}}$:** A noisy *and* biased measurement of the true performance of an investment,
3. **The Out-of-Sample Sharpe Ratio $\widehat{\text{SR}}_{\text{OOS}}$:** A noisy but unbiased measurement of the true performance of an investment.

In theory, as the sample sizes grow, which we will denote by T_{IS} and T_{OOS} respectively, the observed Sharpe ratios $\widehat{\text{SR}}_{\text{IS}}$ and $\widehat{\text{SR}}_{\text{OOS}}$ should converge to the true Sharpe ratio SR . In practice however, the number of useful observations is typically constrained. The amount of in-sample returns for an investment is typically fixed, and out-of-sample returns are only observed gradually over time. If our goal is to estimate the true performance of an investment we must therefore have a robust method to combine the biased in-sample and unbiased out-of-sample performance measurements in a consistent manner. This is exactly what our method provides.

Types of Selection Bias

To understand how selection bias affects inference of statistics it is best to consider a clear example, such as the hedge fund dataset we use in Section 3.3. Selection bias arises in this dataset through funds entering commercial databases according to decisions that are correlated with in-sample performance. Think of a selection filter operating on a fund's in-sample track record: the fund has an unobserved true Sharpe ratio SR , with a corresponding in-sample estimate $\widehat{\text{SR}}_{\text{IS}}$, and is observed in the dataset only if it passes a performance-dependent selection rule.

This is defined as *random* selection bias by Yekutieli [Yek12]. In the random case, we analyse a cross-section of samples that have already passed the selection filter, so both the distribution of the latent parameter SR and the observed statistic $\widehat{\text{SR}}_{\text{IS}}$ are affected by selection. Intuitively, low- SR samples are under-represented because they are less likely to produce high $\widehat{\text{SR}}_{\text{IS}}$ and therefore less likely to be

observed.²

3.2.2 Model Specification

We have two goals: to fit our model to a dataset of reasonably homogeneous investments, and then to perform inference for new individual investments. The first requires the joint marginal likelihood of the in-sample and out-of-sample Sharpe ratios; the second, the posterior for the true Sharpe ratio. The ingredients for both are the same: a prior, a likelihood function, and an acceptance function.

Prior

The prior is the distribution we assume all true Sharpe ratios are drawn from, both for the observed samples in our dataset, and the unobserved samples which were not selected. We take a location-scale t prior,

$$\pi(\text{SR}) \sim \text{lst} \left(\text{SR}_0, \varsigma_0^2 \frac{\nu - 2}{\nu}, \nu \right),$$

which we have parameterised such that $\mathbb{E}[\text{SR}] = \text{SR}_0$ and $\mathbb{V}[\text{SR}] = \varsigma_0^2$, with $\nu > 2$ controlling the kurtosis of the prior. We choose this prior as it strikes a good balance between flexibility – allowing us to fit to the observed data well – and relatively few parameters to fit. In particular, allowing for excess kurtosis is vital, as we observe that a few funds exhibit significantly higher Sharpe ratios than the rest. In contrast, the empirical distribution of out-of-sample Sharpe ratios is rather symmetric, and therefore we do not use an even more flexible distribution for parsimony.

² Two related cases are useful for intuition. In *fixed* selection bias, a single fund’s SR is treated as fixed and only the observed statistic $\widehat{\text{SR}}_{\text{IS}}$ is conditioned on selection, which generally requires larger posterior corrections. In *hierarchical* selection bias, selection occurs after choosing among correlated variants (e.g. parameter-tuned strategies within a family), which is particularly relevant for quantitative strategy development. We treat both cases and provide examples to aid intuition in Appendix 3.A.2.

Likelihood

The likelihood of the sample Sharpe ratios is determined by the assumption on the distribution of the underlying returns. When the returns are iid Gaussian, then the Sharpe ratio for a sample of T_{IS} observations follows a scaled non-central t -distribution with non-centrality parameter $T_{\text{IS}}\text{SR}$ and $T_{\text{IS}} - 1$ degrees of freedom, $T_{\text{IS}}\widehat{\text{SR}}_{\text{IS}} \sim t_{T_{\text{IS}}-1}(T_{\text{IS}}\text{SR})$ [Pav21].

Computing non-central t -densities is slow in implementations, though. To speed up computations, we use a Normal approximation for the sampling distribution $f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$. The sampling distribution of the Sharpe ratio was first derived by Lo [Lo02], and later extended to cover cases where the returns have skewness and kurtosis [Mer02], or autocorrelation [Pav21], or both simultaneously [LLZ25], which is the form used here:

$$\sqrt{T_{\text{IS}}} \left(\widehat{\text{SR}}_{\text{IS}} - \text{SR} \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{1+\phi}{1-\phi} - \frac{1+\phi+\phi^2}{1-\phi^2} \gamma_{3,r} \text{SR} + \frac{1+\phi^2}{1-\phi^2} \frac{\gamma_{4,r}-1}{4} \text{SR}^2 \right).$$

Here, $\xrightarrow{d}$ denotes convergence in distribution as the sample size grows ($T_{\text{IS}} \rightarrow \infty$), ϕ is an AR(1) autocorrelation parameter, $\gamma_{3,r} = \mathbb{E} \left[\left(\frac{r_t - \mu}{\sigma} \right)^3 \right]$ is the skewness of the underlying returns, and $\gamma_{4,r} = \mathbb{E} \left[\left(\frac{r_t - \mu}{\sigma} \right)^4 \right]$ is the raw kurtosis of the underlying returns (not the excess kurtosis, $\gamma_{4,r} - 3$). This likelihood is suitably flexible to allow us to consider common departures from iid Gaussian returns, but a simpler likelihood can be recovered by setting $\phi = 0, \gamma_{3,r} = 0, \gamma_{4,r} = 3$. Further discussion is given in Appendix 3.A.1, where we also retain a proof of the result. Note that in our notation for the likelihood $f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$, we suppress the dependence on T_{IS} (or T_{OOS}) for clarity.

Acceptance Probability

To model selection, we introduce an *acceptance function* $p_{\text{acc}}(\widehat{\text{SR}}_{\text{IS}})$: the probability that an investment appears in our observed dataset given its in-sample Sharpe

ratio. In the hedge fund setting, this captures the reduced-form effect of managers choosing whether (and when) to report returns conditional on their in-sample performance. In other applications, p_{acc} can represent publication decisions, internal capital allocation, or any performance-based gating mechanism.

Following Chen and Zimmermann [CZ20], we use a logistic specification,

$$p_{\text{acc}}(\widehat{\text{SR}}_{\text{IS}}) = \frac{1}{1 + \exp\left(-\ell \left(\widehat{\text{SR}}_{\text{IS}} - \kappa\right)\right)}. \quad (3.2.1)$$

The midpoint κ is the in-sample Sharpe ratio at which inclusion probability is 50%. The slope ℓ controls how sharp the filter is: small ℓ corresponds to a soft threshold where non-performance considerations matter, while large ℓ approximates a hard cutoff driven primarily by $\widehat{\text{SR}}_{\text{IS}}$.

Alpha Decay

Selection bias alone does not fully explain why out-of-sample performance is often materially worse than in-sample performance. We therefore include *alpha decay*, defined here as systematic performance deterioration out-of-sample beyond what would be implied by selection effects and estimation noise alone. We model decay by scaling the true Sharpe ratio in the out-of-sample period:

$$\text{SR}_{\delta} := \delta \text{SR}, \quad \delta < 1,$$

so that out-of-sample returns are generated as if the investment's true Sharpe ratio has been shrunk toward zero.³

Including δ is essential for identification. The empirical gap between $\widehat{\text{SR}}_{\text{IS}}$ and $\widehat{\text{SR}}_{\text{OOS}}$ can be explained by either (i) a stronger selection filter (larger inflation in $\widehat{\text{SR}}_{\text{IS}}$) or (ii) genuine decay (lower SR_{δ}). Without an explicit decay component, the model is forced to attribute the full gap to selection bias, which can lead to

³ While we considered an additive form, $\text{SR} - \delta$, the multiplicative form consistently provides a superior fit to observed performance patterns.

implausible selection parameters and poor fits. With δ included, the joint shape of in-sample and out-of-sample performance allows the model to disentangle these two sources of degradation.

3.2.3 Model Application

Bias Adjustment

Without selection bias, one can simply use the prior and likelihood defined in the previous section for model fitting and inference. But, as previously discussed, we must incorporate these two effects to correctly perform these tasks. Specifically, we must adjust the prior and likelihood to correct for the effect of the selection filter. These are computed by adapting the results of Yekutieli [Yek12] from hard to soft selection thresholds. The results somewhat differ depending on the kind of selection bias considered, and we collect the relevant results for all cases in Appendix 3.A.2. For random selection bias, the bias adjusted prior $\pi_B(\text{SR})$ and likelihood $f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$ are given by

$$\pi_B(\text{SR}) := \frac{\mathbb{P}(\text{acc} \mid \text{SR})\pi(\text{SR})}{\mathbb{P}(\text{acc})}, \quad f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR}) := \frac{p_{\text{acc}}(\widehat{\text{SR}}_{\text{IS}}) f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})}{\mathbb{P}(\text{acc} \mid \text{SR})},$$

where $p_{\text{acc}}(\widehat{\text{SR}}_{\text{IS}})$ is the pointwise acceptance probability evaluated at an observed in-sample Sharpe ratio $\widehat{\text{SR}}_{\text{IS}}$, while $\mathbb{P}(\text{acc} \mid \text{SR})$ and $\mathbb{P}(\text{acc})$ are (normalising) selection probabilities obtained by integrating this acceptance function over the relevant distributions:

$$\mathbb{P}(\text{acc} \mid \text{SR}) := \int_{\widehat{\text{SR}}_{\text{IS}}} p_{\text{acc}}(\widehat{\text{SR}}_{\text{IS}}') f(\widehat{\text{SR}}_{\text{IS}}' \mid \text{SR}) d\widehat{\text{SR}}_{\text{IS}}', \quad (3.2.2)$$

$$\mathbb{P}(\text{acc}) := \int_{\text{SR}'} \pi(\text{SR}') \mathbb{P}(\text{acc} \mid \text{SR}') d\text{SR}'. \quad (3.2.3)$$

The bias adjusted prior $\pi_B(\text{SR})$ is the unbiased prior $\pi(\text{SR})$, adjusted by the probability that a fund with true Sharpe ratio SR is selected, $\mathbb{P}(\text{acc} \mid \text{SR})$, and

renormalised by the total marginal probability of selection $\mathbb{P}(\text{acc})$. Here, $\mathbb{P}(\text{acc} \mid \text{SR})$ is the expected acceptance probability for a fixed true Sharpe ratio, computed by integrating the acceptance function $p_{\text{acc}}(\cdot)$ over the sampling distribution of $\widehat{\text{SR}}_{\text{IS}}$ given SR. The marginal probability $\mathbb{P}(\text{acc})$ averages $\mathbb{P}(\text{acc} \mid \text{SR})$ over the unbiased prior $\pi(\text{SR})$, ensuring π_B integrates to one. In general these normalising constants are not available in closed form under flexible priors and likelihoods, so they must be computed numerically; we discuss efficient and stable approximations in Appendix 3.B.2.

Model Fitting

We fit the Bayesian model to the observed cross-section of pre- and post-listing Sharpe ratios using maximum likelihood estimation (i.e. empirical Bayes [Rob56]). This requires fitting the prior parameters, $\text{SR}_0, \varsigma_0, \nu$, the acceptance function parameters, κ, ℓ , and the alpha decay parameter δ . These collectively form the parameter set $\Theta = \{\text{SR}_0, \varsigma_0, \nu, \kappa, \ell, \delta\}$.

To this end, we first compute the joint marginal likelihood of the in- and out-of-sample Sharpe ratios, $f(\widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}})$. This is obtained by integrating the product of the bias-adjusted in-sample likelihood, the bias-adjusted prior, and the out-of-sample likelihood over the support of the true Sharpe ratio SR' . Note that the out-of-sample likelihood is conditional on the alpha-decayed parameter SR'_δ :

$$f(\widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}}) = \int_{\text{SR}'} f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR}') f(\widehat{\text{SR}}_{\text{OOS}} \mid \text{SR}'_\delta) \pi_B(\text{SR}') d\text{SR}'. \quad (3.2.4)$$

Note that we must use the bias adjusted in-sample likelihood, $f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$, and prior, $\pi_B(\text{SR})$. The out-of-sample Sharpe ratio however is unbiased, and we therefore use the standard likelihood but with the alpha-decayed true Sharpe ratio SR_δ rather than the standard true Sharpe ratio SR. While one could try to fit the model using only in-sample data, we find that this method often fails to recover the true parameters accurately when tested on samples simulated from a known

data generating process. In particular, the alpha decay parameter δ is impossible to identify without the joint fitting approach. In Appendix 3.B.1 we provide simulation results with known parameters demonstrating the need to fit with both in- and out-of-sample data, and how it significantly improves parameter estimates.

Next, we maximise the joint likelihood eq. (3.2.4) across all pairs of in- and out-of-sample Sharpe ratios. Assuming the pairs are independent across investments, i.e. $(\widehat{\text{SR}}_{\text{IS},i}, \widehat{\text{SR}}_{\text{OOS},i})$ and $(\widehat{\text{SR}}_{\text{IS},j}, \widehat{\text{SR}}_{\text{OOS},j})$ are independent for $i \neq j$, we can equivalently maximise the sum of the log-likelihoods,

$$\max_{\Theta} \sum_i \log f(\widehat{\text{SR}}_{\text{IS},i}, \widehat{\text{SR}}_{\text{OOS},i} \mid \Theta).$$

The independence assumption is strong and unlikely to hold for raw returns, as they typically consist of both systematic exposure (β) and idiosyncratic performance (α). However, when comparing investments such as hedge funds, we typically wish to identify the hedge funds which deliver α rather than β , as there are much cheaper ways to buy β exposure. We therefore focus on the idiosyncratic component of the returns, which we find is sufficiently independent for empirical Bayes to yield useful parameter estimates. We study the effect of violating this assumption in Appendix 3.B.4.

In practice, this fitting procedure is challenging due to a complex loss surface and the numerical cost of calculating the marginal likelihood eq. (3.2.4). We detail how we overcome these challenges and make it feasible to apply the method in Appendix 3.B.2. The key lies in the usage of an optimised, compiled, and GPU-enabled log likelihood calculation (using JAX [FJL19]), and the robust fitting methods provided by CERN’s Minuit2 library [JR75; Dem+25].

The MLE approach is extremely efficient, and the best choice when speed is required (for example, when fitting many models for walkforward analysis). But, when the selection effect is relatively weak (for example $\ell < 5$), the loss surface becomes flat and a pure MLE approach tends to select extreme “non-physical” κ

values. To resolve this, we suggest a hybrid approach, utilising Generalised Method of Moments to fit a subset of parameters (typically κ and δ , which are well identified by the sample moments), with the remainder being fit by MLE. This method acts as a form of regularisation, constraining the parameters to intuitive regions. Full details of this two-step approach, which we term MLE-GMM, are given in Appendix 3.B.3. Note that a full GMM approach is inappropriate for the task at hand as some of the parameters, for example ℓ , are not well identified by the sample moments.

Inference

Once the model has been fit to the observed distribution of in- and out-of-sample Sharpe ratios, it can then be used to compute the posterior for each investment's true Sharpe ratio in a standard manner. To do so, we simply apply Bayes rule and use the bias adjusted prior and likelihood,

$$\pi_B(\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}) = \frac{\pi_B(\text{SR})f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})}{\int_{\text{SR}'} \pi_B(\text{SR}')f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR}') d\text{SR}'}. \quad (3.2.5)$$

In the case of random selection bias, we observe that the bias adjustments to the prior and likelihood “cancel out”, yielding the same posterior as if there was no selection bias,

$$\begin{aligned} \pi_B(\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}) &= \frac{\pi_B(\text{SR})f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})}{\int_{\text{SR}'} \pi_B(\text{SR}')f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR}') d\text{SR}'} \\ &= \frac{\pi(\text{SR})f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})}{\int_{\text{SR}'} \pi(\text{SR}')f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR}') d\text{SR}'} = \pi(\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}). \end{aligned}$$

Intuitively, this says that if you know the correct prior $\pi(\text{SR})$, then you can perform Bayesian inference on a biased in-sample Sharpe ratio $\widehat{\text{SR}}_{\text{IS}}$ as if there was no selection bias (i.e. using the unadjusted likelihood $f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$). But this is only true if you have the correct prior $\pi(\text{SR})$. By using the bias adjusted prior $\pi_B(\text{SR})$

and likelihood $f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$ when fitting the model via MLE, we have been able to estimate the parameters of the unbiased prior $\pi(\text{SR})$.

If we had naively estimated the parameters of the prior, the Bayesian approach would fail to correct for the selection bias as we would have misestimated the parameters of the prior $\pi(\text{SR})$, likely leading to estimating an overly optimistic prior mean SR_0 .⁴

To compute the posterior for the alpha-decayed Sharpe ratio, i.e. the true Sharpe ratio in the out-of-sample period, we apply a simple change-of-variables transformation,

$$\pi(\text{SR}_\delta \mid \widehat{\text{SR}}_{\text{IS}}) = \pi(\text{SR}_\delta / \delta \mid \widehat{\text{SR}}_{\text{IS}}) / \delta.$$

We can also use the out-of-sample data to further update the posterior via Bayesian continuous updating,

$$\pi_B(\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}}) \propto f(\widehat{\text{SR}}_{\text{OOS}} \mid \text{SR}_\delta) \pi_B(\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}).$$

This allows us to use all the data available, both the biased in-sample returns and the unbiased out-of-sample returns, to estimate the true performance of a given investment. This is very useful in practice. For example, we can compute the initial posterior once for a fund when it enters the dataset using its backfilled returns, and then continuously update it as new returns arrive, refining our estimate of the true Sharpe ratio for that fund. Updating in this fashion forms a key application we explore later, using evolving estimates of the true Sharpe ratio to decide whether a fund should continue to be included in our portfolio as we observe out-of-sample returns.

In addition to updating the posterior with out-of-sample data, we may also want to check if the out-of-sample data is in line with our expectations given the posterior

⁴ In the other (fixed and hierarchical) cases, the posterior requires further adjustment, as in Appendix 3.A.2.

computed on the in-sample data. This is done by comparing the observed out-of-sample Sharpe ratio to the posterior predictive distribution computed using the posterior from the in-sample Sharpe ratio alone, specifically,

$$f(\widehat{\text{SR}}_{\text{OOS}} \mid \widehat{\text{SR}}_{\text{IS}}) = \int_{\text{SR}'} f(\widehat{\text{SR}}_{\text{OOS}} \mid \text{SR}'_{\delta}) \pi(\text{SR}'_{\delta} \mid \widehat{\text{SR}}_{\text{IS}}) d\text{SR}'_{\delta}. \quad (3.2.6)$$

Here it is important to note that we must use the alpha decayed posterior to account for the possible alpha decay in the out-of-sample period. This is another application we explore later, where we can detect if a fund is performing even worse (or better than) in the out-of-sample period than we would expect given our bias adjusted estimate for their true Sharpe ratio using their in-sample returns.⁵

3.3 Empirical Case Study: Hedge Fund Returns

With the model fully described, we turn now to an empirical application of our approach. To do so, we utilise a dataset of hedge fund returns collected by Joenväärä et al. [Joe+21] by merging six major commercial hedge fund databases (BarclayHedge, EurekaHedge, Hedge Fund Management, Hedge Fund Research, Lipper TASS and Morningstar) into a single consistent dataset from January 1990 to December 2023. The dataset is only survivorship bias-free from 1994 onwards, so we restrict ourselves to data after this date.

The dataset consists of monthly returns net of fees, monthly assets under management and a collection of characteristics for each fund, such as style, domicile, use of leverage, etc. Importantly, we have a “listing date” for each fund, either directly given by the database or implied using the method from Jorion and Schwarz [JS19]. Using this listing date, we can split the returns into pre- and post-listing returns, which correspond to in- and out-of-sample returns in our terminology.

⁵ This is also useful for example for answering questions such as: What is the probability that this fund will do worse out-of-sample than in-sample? We can compute this simply with the conditional CDF of the out-of-sample Sharpe ratio, evaluated at the observed in-sample Sharpe ratio: $F_{\widehat{\text{SR}}_{\text{OOS}} \mid \widehat{\text{SR}}_{\text{IS}}}(\widehat{\text{SR}}_{\text{IS}})$.

As the hedge fund literature documents (e.g., Bhardwaj, Gorton, and Rouwenhorst [BGR14]; Fung and Hsieh [FH09]), pre-listing returns are often inflated relative to post-listing performance. One driver is the incentive for successful funds to list and attract capital, see Appendix 3.C.7 for figures showing AUM change relative to listing date; consequently, backfilled returns tend to exceed non-backfilled ones. Furthermore, funds often undergo an incubation period, deploying the manager’s capital to build a track record before listing. Because funds with poor records rarely list, this induces an incubation bias. This phenomenon serves as an ideal case study for our approach.

Setup

To use the hedge fund data for our approach, some preprocessing must be done. This includes filtering the full set of funds into a clean subset, and separating the returns into systematic and non-systematic components. The details of these two steps can be found in Appendix 3.C. The procedure yields a clean dataset of pre- and post-listing returns which have limited systematic risk exposure (as required by our approach).

We fit the model using the MLE-GMM approach, as described in Appendix 3.B.3, with κ, δ as the profile set. A full discussion of why we select these two parameters for profiling is available in Appendix 3.C.3. For the walkforward analysis, where we are less concerned with parameter interpretability but need to minimise fitting time, we utilise the direct MLE approach, as described in Section 3.2.3.

In addition to studying the usage of our model across the entire cross-section of funds, we investigate how the model parameters vary across different fund types, and how this affects inference, in Appendix 3.C.4. To conduct this analysis we cluster funds according to their style, domicile and other quantitative and qualitative factors, and find that different types of funds exhibit significantly different amounts of selection bias and alpha decay.

We divide the remainder of this section into two parts: a statistical analysis, covering all aspects independent of portfolio construction, and an application of the model to fund selection.

Statistical Analysis

Pre- and Post-listing Returns Over Time

It is well documented that hedge fund performance has broadly reduced over time, see Appendix 3.C.2 and Bollen, Joenväärä, and Kauppila [BJK21]. But is the pre- and post-listing problem also getting worse? We find that yes, it is.

To this end, we split the full time series of returns into three periods according to the regimes identified in Appendix 3.C.2: pre-September 2001, September 2001 to July 2008, and post-July 2008. We restrict the sample for each period to funds launched within each specific window. Without this restriction, funds would lack the pre-listing returns necessary for our method.

In fig. 3.3.1 we plot cumulative average returns as a function of months since listing. Before averaging across funds, we rescale returns so that its volatility equals 1, and then take cross-fund averages. This ensures that high-volatility funds do not dominate the aggregate pattern. We find that, in the post-2008 period, average returns are substantially higher before listing than after listing, as reflected by the noticeably steeper slope of the cumulative return curve in the pre-listing window. In contrast, this pre- versus post-listing difference is much weaker in the two earlier periods.

Bayesian Model Parameters

How does this materialise in the pre- and post-listing Sharpe ratios? In fig. 3.3.2 we plot the empirical distributions of the pre- and post-listing Sharpe ratios, along with the fitted marginal densities from our Bayesian model. To fit the model here

3.3 Empirical Case Study: Hedge Fund Returns

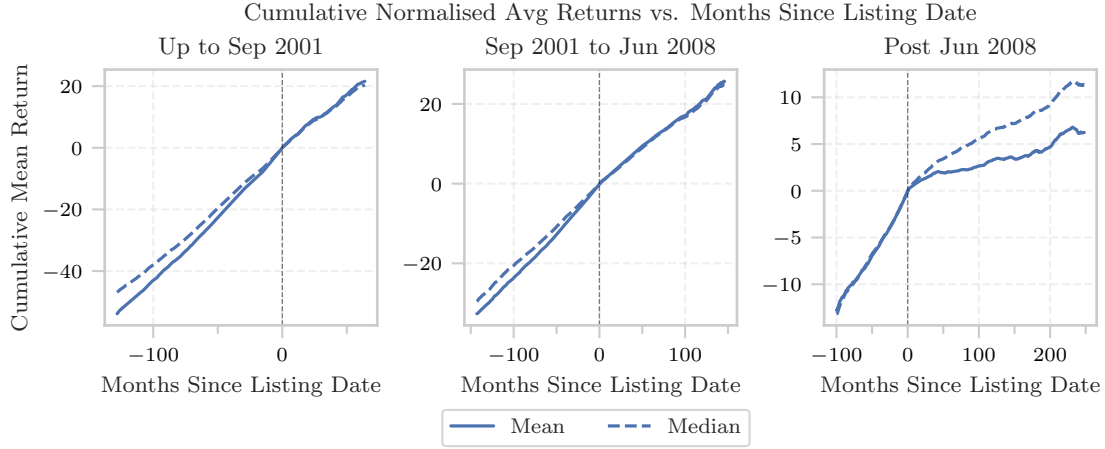


Figure 3.3.1: Cumulative (normalised) average returns vs months pre- and post-listing split into the three identified performance periods.

we have simply taken the observed pre- and post-listing Sharpe ratios for each period (and pre- and post-listing period lengths, see fig. 3.C.17 for the distribution of period lengths in each period), and run the MLE-GMM fitting procedure. The fitted parameters are reported in table 3.3.1.

We see that there is a reduction in risk-adjusted performance post-listing in all three periods, with the effect being most pronounced in the post 2008 period, where we observe that the mean annualised Sharpe ratio post-listing is just 0.1 in our sample of funds. We remind the reader that these are the non-systematic returns only, i.e. the performance hedged against a large array of risk factors (see Appendix 3.C.2 for further details).

From the fitted parameters, we see that the location of the prior SR_0 is similar in the first two periods, but there is little to no selection bias in the second period ($\kappa \approx 0$), while in the first period the selection effect inflates the observed pre-listing Sharpe ratios. In the most recent period, we see that the prior for the true Sharpe ratios is essentially zero, with the positive average performance being driven strongly by selection bias (characterised by a larger ℓ value than the first two periods). Additionally, we see that the alpha decay has become stronger over

3.3 Empirical Case Study: Hedge Fund Returns

time, with δ decreasing from 0.87 to 0.68 and finally 0.55. Finally, we see that the uncertainty in the prior is much larger in the 2001 to 2008 period (large ς_0 and small ν). From this analysis we see that the bias correction has become more and more important over time, and it seems unlikely that this trend won't continue.

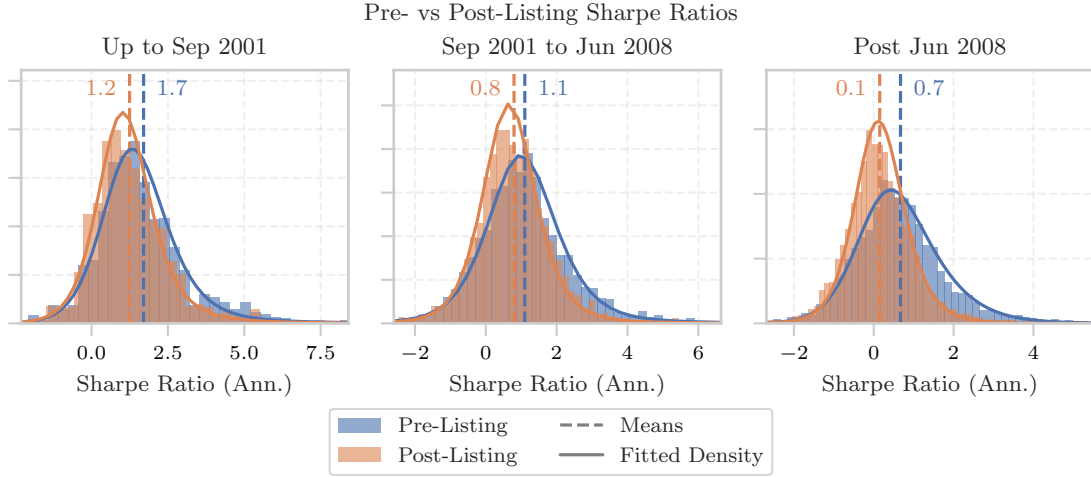


Figure 3.3.2: Empirical distribution of pre- and post-listing Sharpe ratios by period. Empirical means are marked with dashed lines and annotated with their value, and the fitted marginal likelihood from the Bayesian model is shown with a solid line.

	Up to Sep 2001	Sep 2001 to July 2008	Post July 2008
SR_0	0.30 (1.04)	0.27 (0.94)	-0.01 (-0.03)
ς_0	0.30 (1.04)	1.79 (6.20)	0.21 (0.73)
ν	3.26	2.01	5.21
κ	2.33 (8.07)	-0.07 (-0.24)	1.19 (4.12)
ℓ	1.24	0.86	2.50
δ	0.87	0.68	0.55

Table 3.3.1: Fitted parameters for the Bayesian model by period. Annualised equivalents are added in (brackets) for SR_0 , ς_0 and κ .

Decomposition of Selection and Decay Effects

By examining the distribution of the prior $\pi(SR)$, bias adjusted prior $\pi_B(SR)$, and bias & decay adjusted prior $\pi_B(SR_\delta)$ (calculated using the expressions in Sec-

3.3 Empirical Case Study: Hedge Fund Returns

tion 3.2.3), we can see how the selection and decay effects influence the distribution of true Sharpe ratios, which helps us understand why the post-listing Sharpe ratios are lower than their pre-listing counterparts.

In fig. 3.3.3 we plot these densities and in table 3.3.2 we provide the mean and mode of the densities. In line with the discussion above, the middle period has little difference between the unadjusted and bias adjusted prior, while the first and last period have large adjustments. Additionally, by noting the difference between the mean and modal values for the first and last period, we observe how selection bias introduces skewness into the conditional distribution of true Sharpe ratios. While the underlying unadjusted prior $\pi(\text{SR})$ is symmetric, the population of funds that pass the selection filter is not. As discussed in section 3.2.1, the selection filter applies jointly to the in-sample and true parameters, which causes this skewness in the selection adjusted true Sharpe ratios. Funds with low true Sharpe ratios are less likely to have high in-sample Sharpe ratios and are therefore less represented in the observed sample. Finally, we see that the multiplicative alpha decay has a dual effect: by shrinking values towards zero, it affects not only the location of the prior but also its variance. In essence, it enforces a stronger reversion to the mean.

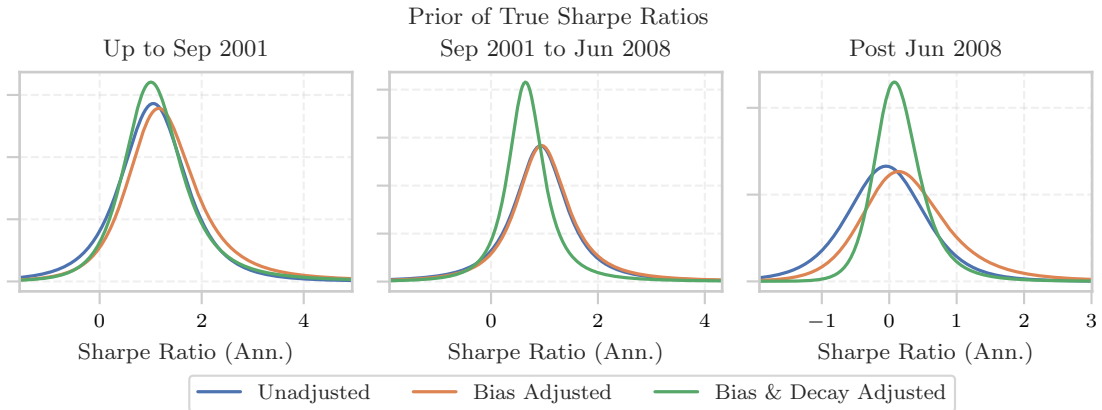


Figure 3.3.3: Unadjusted, bias adjusted, and bias & decay adjusted prior densities for the three periods.

3.3 Empirical Case Study: Hedge Fund Returns

	Up to Sep 2001		Sep 2001 to July 2008		Post July 2008	
	Mean	Mode	Mean	Mode	Mean	Mode
Unadjusted	1.06	1.07	0.94	0.94	-0.05	-0.05
Bias Adjusted	1.36	1.16	1.03	0.94	0.32	0.12
Bias & Decay Adj.	1.18	1.00	0.70	0.64	0.18	0.07

Table 3.3.2: Unadjusted, bias adjusted and bias & decay adjusted expected true Sharpe ratios $\mathbb{E}[\text{SR}]$ for the three periods.

Analysis of Posterior Estimates and Performance Haircuts

Finally, we can look at the primary quantity of interest: the posteriors of the true Sharpe ratio. To compute the posterior we take the parameters for each period from table 3.3.1 and compute the expected posterior Sharpe ratio $\mathbb{E}[\text{SR}_i \mid \widehat{\text{SR}}_{\text{IS},i}]$, and decayed equivalent $\delta \mathbb{E}[\text{SR}_i \mid \widehat{\text{SR}}_{\text{IS},i}]$ for each fund which launched in that period. For comparison, we also compute the posterior using both the pre- and post-listing Sharpe ratios, $\mathbb{E}[\text{SR}_i \mid \widehat{\text{SR}}_{\text{IS},i}, \widehat{\text{SR}}_{\text{OOS},i}]$. We are interested in how much shrinkage the Bayesian method applies to the pre-listing Sharpe ratio, and how this is split between Bayesian shrinkage/selection effects and alpha decay.

To visualise this, we can look at the “haircut” estimated by the Bayesian model, i.e. the percentage reduction from the pre-listing performance to our best guess for the true Sharpe ratio. To do so, we compute the ratio of the posterior estimate (undecayed or decayed) to the in-sample Sharpe ratio $\widehat{\text{SR}}_{\text{IS},i}$, and define the “haircut” as one minus this quantity, e.g. $1 - \mathbb{E}[\text{SR}_i \mid \widehat{\text{SR}}_{\text{IS},i}]/\widehat{\text{SR}}_{\text{IS},i}$.

In Figure 3.3.4 we plot the median of these haircuts across funds, decomposed into the quantity attributable to Bayesian shrinkage and selection bias⁶ and attributable to alpha decay. We observe that in the post 2008 period, the shrinkage has grown significantly in comparison to the earlier two periods. We also note that

⁶ Note that it is difficult to distinguish between Bayesian shrinkage and selection bias purely from the posterior, as the bias adjusted posterior is equal to the unbiased posterior in the random case, see the discussion in Section 3.2.3. This is why inspecting the prior (which does change with selection bias) can be useful for insights.

when comparing the first two periods, alpha decay was the primary cause for the reduction in post-listing performance, indicating that the selection effect was relatively consistent in the two periods, but post-listing alpha decay was more severe in the middle period.

We note that when we compute the posterior using just the in-sample data, $\pi(\text{SR}|\widehat{\text{SR}}_{\text{IS}})$, the median haircut is larger than when we compute it using the in- and out-of-sample data, $\pi(\text{SR}|\widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}})$. As with all Bayesian approaches, the arrival of new evidence (the out-of-sample returns) diminishes the influence of the prior, leading to a posterior that is slightly closer to the observed data and less dominated by the initial shrinkage.

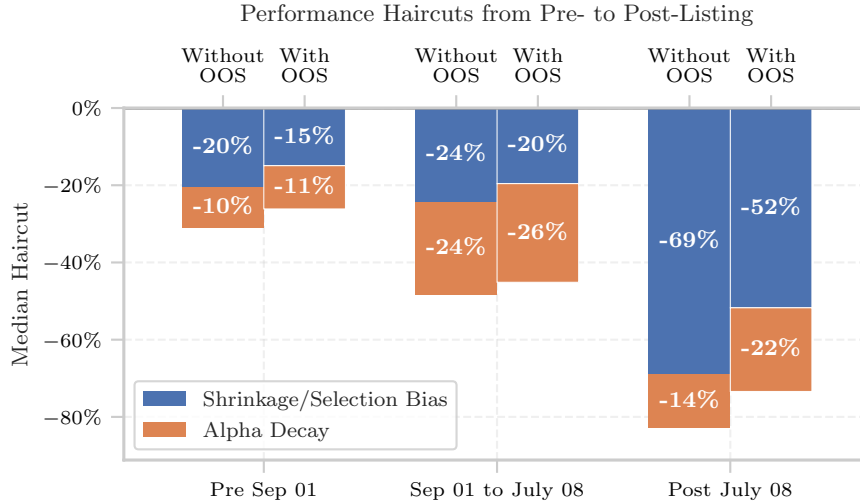


Figure 3.3.4: Bayesian estimated haircuts by period, with the posterior estimated either with pre-listing data alone, or with pre- and post-listing data.

Model Evaluation and Comparative Performance

Having established the changing nature of bias and decay over time, we now turn to a formal evaluation of our approach against baseline methods: a simple James–Stein shrinkage estimator [JS60; EM75]; our Bayesian model fit without accounting for selection bias (both with and without alpha decay); and the bias-aware model fit to

pre-listing Sharpe ratios alone, or fit to pre- and post-listing Sharpe ratios split by the clusters described in Appendix 3.C.4. We provide a more extensive description of these comparators in Appendix 3.C.5, and we report the parameters for each model and each period in table 3.C.4. The following analysis serves to validate that our complex Bayesian model provides a strictly superior fit to the data compared to simpler alternatives, a necessary prerequisite before assessing its economic utility in portfolio construction.

We evaluate both point prediction accuracy and model calibration following the recommendations of [Gel+13, Ch. 6 & 7]. Specifically, we compute the mean squared error (MSE) and log pointwise predictive density (LPPD) as measures of predictive accuracy, and probability integral transforms (PITs) together with Kolmogorov–Smirnov (KS) statistics as diagnostics of calibration and model fit. We evaluate these diagnostics relative to the naive baseline (percentage reduction in MSE, increase in MLPPD, and reduction in KS distance), so that moving to the right corresponds to better performance for all metrics. We provide a full description of the calculation of these statistics in Appendix 3.C.5.⁷

Figure 3.3.5 shows the results, where we have ordered the models (broadly) by their efficacy from worst to best. In the two periods with relatively low selection effects (as shown in the analysis above), the unbiased model with alpha decay performs almost as well as the full bias-aware approach. And when looking purely at the point predictive metrics, it continues to perform well in the post 2008 period, with the bias adjusted method providing a modest improvement in performance. However, the Kolmogorov–Smirnov statistic (which measures how closely the model’s predicted probabilities line up with realised outcomes) clearly shows that accounting for selection bias is essential for accurately matching the true distribution of out-of-sample Sharpe ratios. This is important if using these probabilities to make

⁷ There is some bias here, in that both the pre- and post-listing Sharpe ratios are used to calibrate the model parameters Θ initially; however our goal here is a comparative analysis between models, most of which benefit from this bias. In Appendix 3.C.6 we conduct a walkforward analysis that avoids this issue, but in turn has its own limitations.

3.3 Empirical Case Study: Hedge Fund Returns

investment decisions, as we do in the economic section below.

Additionally, we observe that the model fit to the in-sample data alone does not perform as well as the model fit to the in- and out-of-sample data. This is due to two reasons: 1. As discussed elsewhere (see Appendix 3.B.1), fitting to the in- and out-of-sample data yields improved parameter estimates; 2. Capturing alpha decay (which we cannot do when fitting to in-sample data alone) is vital (as demonstrated by the generally good performance of the unbiased model which does have alpha decay).



Figure 3.3.5: Improvement in statistical metrics relative to a naive baseline for the models described in Appendix 3.C.5.

This analysis confirms that:

1. Fitting to in- and out-of-sample data is key for good model fits,
2. Alpha decay is a vital model component, and,
3. Inclusion of selection bias is required for well calibrated predictive distributions.

Dynamic Fund Selection

We conclude by demonstrating the economic utility of our approach in guiding fund selection. A primary advantage of the Bayesian framework is the generation of full densities, enabling probabilistic decision-making.⁸ For instance, an investor might choose to allocate capital only to funds with a 95% probability of a positive Sharpe ratio.

To implement this rule, we compute the posterior density for each fund, and select those satisfying the condition

$$\mathbb{P}(\text{SR}_{\delta,i} \geq 0 \mid \widehat{\text{SR}}_{\text{IS},i}, \widehat{\text{SR}}_{\text{OOS},i}) \geq 0.95.$$

Another useful question we often like to answer is: Is this fund performing as well as expected, given its pre-listing performance? We address this by computing the predictive distribution defined in Equation (3.2.6) using the posterior computed using only the pre-listing performance,

$$f(\widehat{\text{SR}}_{\text{OOS}} \mid \widehat{\text{SR}}_{\text{IS}}) = \int_{\text{SR}'_{\delta}} f(\widehat{\text{SR}}_{\text{OOS}} \mid \text{SR}'_{\delta}) \pi(\text{SR}'_{\delta} \mid \widehat{\text{SR}}_{\text{IS}}) d\text{SR}'_{\delta},$$

and then compare the observed out-of-sample Sharpe ratio $\widehat{\text{SR}}_{\text{OOS}}$ to this distribution. We then select funds where the quantile of the post-listing performance exceeds some threshold, for example,

$$F(\widehat{\text{SR}}_{\text{OOS}} \mid \widehat{\text{SR}}_{\text{IS}}) \geq a,$$

where $F(\cdot)$ denotes the CDF of the predictive distribution.

We examine the profitability of these trading rules using the naive, James–Stein and “Bias Adjusted” (hereafter Bayesian) estimators, excluding the other variants for clarity. As before, we use the walkforward fitted parameters, refit yearly, using only returns observed up to time $t - 1$ to compute posteriors at time t . For example,

⁸ Earlier probability-threshold rules based on Sharpe-ratio uncertainty appear in [BL12; BL14; LLZ25], for example.

to compute the posterior for January 2002, we use the data up to and including December 2001.

We use the estimated non-systematic returns for both posterior calculation and performance computation. As detailed in Appendix 3.C.2, we estimate the non-systematic returns using the full history for each fund. Similarly, we assume the strategy can invest in any fund, at any size and at any time. While these idealised conditions yield elevated Sharpe ratios, they allow us to isolate our approach’s comparative performance against naive benchmarks which also benefit from these conditions. The results demonstrate that the Bayesian model can be successfully fitted, and is applicable to real data exhibiting autocorrelation, heavy tails and other statistical effects. Consequently, one should view the reported Sharpe ratios as metrics of relative predictive power rather than realistic live trading returns.

Having selected the funds for each month, we allocate risk equally so that high volatility funds do not dominate the portfolio; the weight for each fund is

$$\mathbf{w}_{i,t} = \frac{\mathbb{1}_{i \in \text{sel}_t} \sigma_{i,t}^{-1}}{\sum_{i=1}^N \mathbb{1}_{i \in \text{sel}_t} \sigma_{i,t}^{-1}},$$

where sel_t denotes the selected funds for month t and $\sigma_{i,t}$ is the volatility estimate for fund i using an expanding window of returns up to (but not including) month t . We restrict the selection universe to funds live in month t , allocating zero weight to delisted funds.

Consider first the posterior-based selection rule. We parameterise the rule by the Sharpe ratio cutoff SR_C and the probability threshold p_C , selecting funds where

$$\mathbb{P}(\text{SR}_{\delta,i} \geq \text{SR}_C \mid \widehat{\text{SR}}_{\text{IS},i}, \widehat{\text{SR}}_{\text{OOS},i}) \geq p_C.$$

We run this trading strategy for the Naive, James–Stein⁹ and Bayesian approaches, analysing performance for three time variants:

⁹ For the Naive and James–Stein estimators we use the fund’s entire pre- and post-listing history (truncated at the corresponding start date) to estimate the Sharpe ratio, and use the standard sampling distribution to calculate the probability of exceeding the threshold.

1. From 1996 to 2008, using parameters estimated since 1996,
2. From 2009 to 2020, using parameters estimated since 1996,
3. From 2009 to 2020, using parameters estimated since 2009.

This enables us to isolate the effect of the post-2008 step change from the parameter estimates (see fig. 3.C.18 for the impact of start dates on fitted parameters).

Figure 3.3.6 shows the annualised Sharpe ratios of the portfolios constructed by the three models, with the probability threshold fixed at 50%. In the 1996–2008 period, the naive model outperforms for most cutoffs, with the Bayesian model being best only when the cutoff is raised to 1, a quite restrictive level. This is expected: pre-2008, hedge fund performance was generally strong, and selection effects were modest. However, in the later period, the Bayesian model dominates. The naive and James–Stein approaches are overoptimistic regarding fund Sharpe ratios and fail to exclude under-performers. In contrast, the Bayesian model identifies and cuts these funds aggressively. Notably, the Bayesian model performs similarly for both start dates, whereas the naive and James–Stein estimators deteriorate with the 1996 start date, demonstrating a slower reaction to the regime change.

Figure 3.3.7 presents a similar analysis, fixing the Sharpe ratio cutoff at 0.5 and varying the probability threshold. The results echo the previous plot. The Naive model performs marginally better in the first period, where selection bias is lower, but the Bayesian model outperforms in the latter period, where selection bias is prevalent. At very high thresholds (e.g., 95%), the models converge, selecting the same funds where confidence in the Sharpe ratio is high. The Bayesian model’s outperformance stems primarily from marginal cases where the signal is less clear.

To visualise both parameters simultaneously, fig. 3.3.8 plots the improvement in Sharpe ratio against the Naive and James–Stein models, taking the maximum of the two as the benchmark. In the pre-2008 period, the Bayesian model slightly underperforms the Naive approach; in the post-2008 period, it strongly outperforms, particularly as the cutoff increases. This outperformance arises because the Bayesian

3.3 Empirical Case Study: Hedge Fund Returns

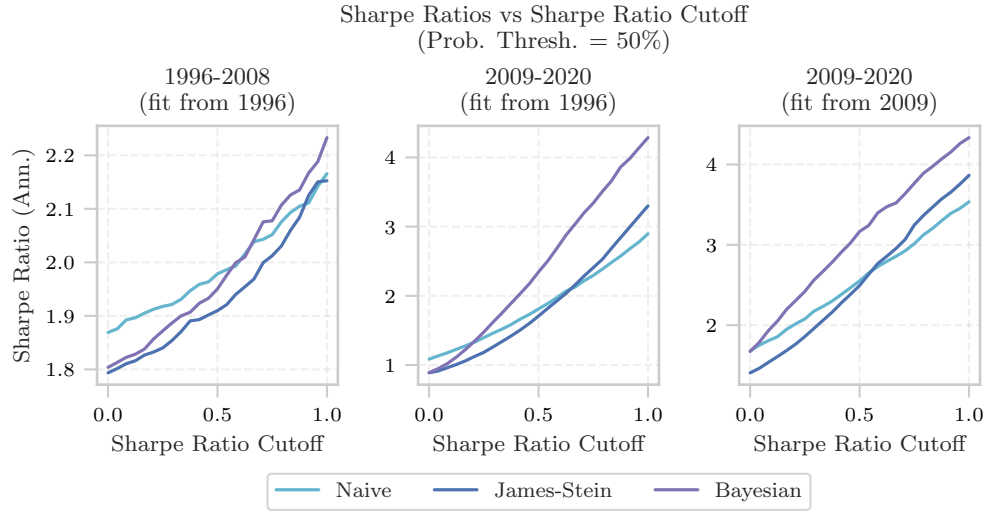


Figure 3.3.6: Sharpe ratios for the posterior based trading strategy for the Naive, James–Stein and Bayesian models by Sharpe ratio cutoff, with the probability threshold set at 50%.

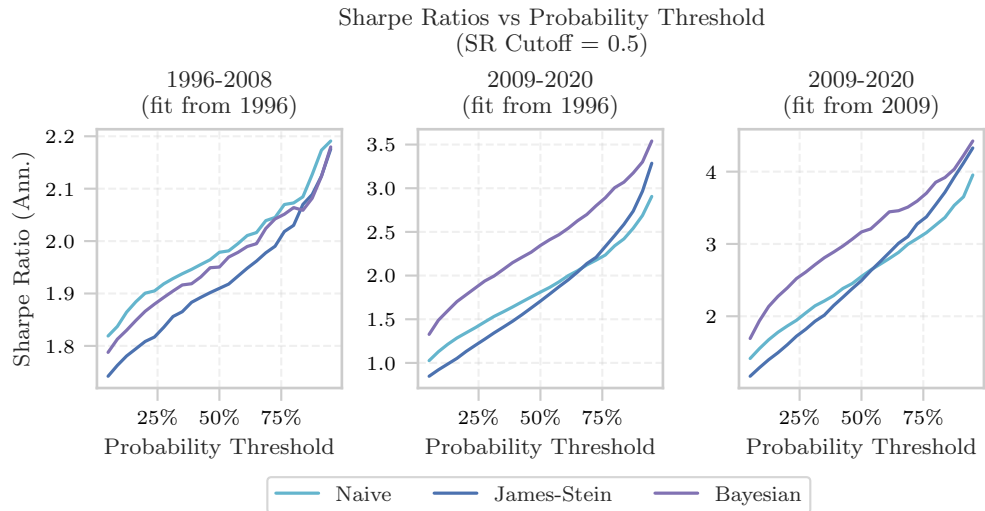


Figure 3.3.7: Sharpe ratios for the posterior based trading strategy for the Naive, James–Stein and Bayesian models by probability threshold, with the Sharpe ratio cutoff set at 0.5.

model estimates future fund performance more accurately (consistent with the prior statistical analysis), effectively pruning underperforming funds.

Finally, in Figure 3.3.9, we examine the predictive-based strategy as a function

3.3 Empirical Case Study: Hedge Fund Returns

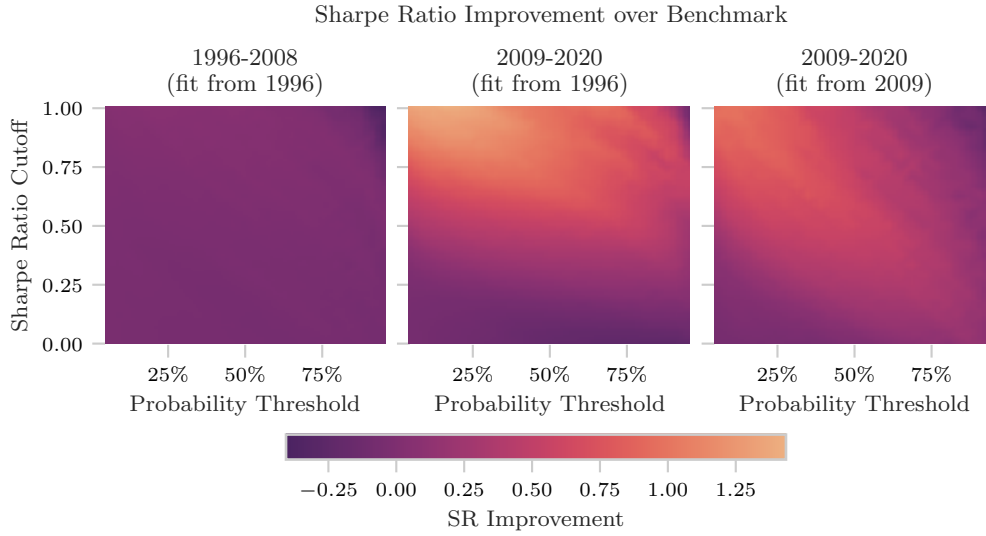


Figure 3.3.8: Improvement over the benchmark for the Bayesian strategy by Sharpe ratio cutoff and probability threshold.

of the quantile threshold. We observe that the James–Stein estimator performs best in the 1996 to 2008 period, where the Bayesian approach exhibits instability, potentially due to the smaller sample size of funds. Post-2008 however, the results are more interesting. For the model fit using data since 1996, the Bayesian model outperforms the James–Stein approach from $a \approx 0.6$; but for the model fit using data from 2009, this transition occurs at exactly $a = 0.5$, the point where funds switch from underperforming to outperforming expectations. This implies that funds outperforming post-listing relative to pre-listing expectations tend to persist in their outperformance, and the Bayesian model is best at identifying this.

This empirical study validates our proposed framework as beneficial for risk management and investment selection. In particular, in regimes where selection effects and alpha decay are strong (such as hedge fund returns post-2008), naive approaches fail to accurately forecast the performance of candidate investments, and a methodology which takes these effects into account is vital.

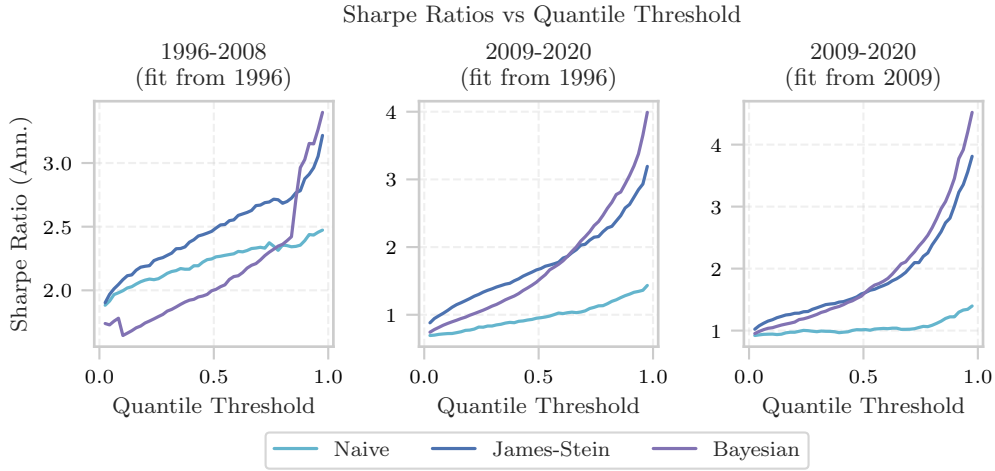


Figure 3.3.9: Sharpe ratios for the predictive based trading strategy for the Naive, James–Stein and Bayesian models by quantile threshold.

3.4 Conclusion

This chapter develops a Bayesian framework to accurately estimate true Sharpe ratios from historical data polluted by selection bias and alpha decay. We show that naive reliance on in-sample performance leads to significant overestimation of future performance. Furthermore, our theoretical and simulation results demonstrate that attempting to correct for these biases using in-sample data alone is insufficient. To correctly identify the model parameters a joint fitting approach utilising the cross-section of both in- and out-of-sample performance is essential.

By providing a principled assessment of true performance, this framework offers a necessary guide to temper overoptimistic estimates, leading to better investment decisions.

Chapter Appendix

3.A Extensions, Other Cases & Additional Theory

3.A.1 Asymptotic Distributions of the Sample Sharpe Ratio

As discussed in Section 3.2.2, we give here the asymptotic distribution for the Sharpe ratio under AR(1) non-Gaussian returns.

Proposition 3.A.1. *Let (r_t) be a time series of asset returns that follows an AR(1) process with $|\phi| < 1$ (hence weakly stationary), namely*

$$r_t - \mu = \phi(r_{t-1} - \mu) + \epsilon_t, \quad \text{for } t \geq 1,$$

starting from $r_0 \in \mathbb{R}$, where $(\epsilon_t)_{t \geq 0}$ are iid shocks with $\mathbb{E}[\epsilon_t] = 0$, $\mathbb{E}[\epsilon_t^2] = \sigma_\epsilon^2$, $\mathbb{E}[\epsilon_t^3] = \gamma_{3,\epsilon}\sigma_\epsilon^3$, and $\mathbb{E}[\epsilon_t^4] = \gamma_{4,\epsilon}\sigma_\epsilon^4$, with all first four moments being finite. Let $\text{SR} := \mu/\sigma$ be the true Sharpe ratio and $\widehat{\text{SR}}_{\text{IS}} := \widehat{\mu}/\widehat{\sigma}$ its sample estimator from a sample length of T_{IS} . As T_{IS} tends to infinity, the following convergence in distribution holds:

$$\sqrt{T_{\text{IS}}} \left(\widehat{\text{SR}}_{\text{IS}} - \text{SR} \right) \xrightarrow{d} \mathcal{N} \left(0, V(\text{SR}, \phi, \gamma_{3,\epsilon}, \gamma_{4,\epsilon}) \right),$$

$$V(\text{SR}, \phi, \gamma_{3,\epsilon}, \gamma_{4,\epsilon}) = \frac{1 + \phi}{1 - \phi} - \gamma_{3,\epsilon} \text{SR} \sqrt{\frac{1 + \phi}{1 - \phi}} + \frac{\text{SR}^2}{4(1 - \phi^2)} \left((\gamma_{4,\epsilon} - 1) + (5 - \gamma_{4,\epsilon})\phi^2 \right).$$

Alternatively, the result can be stated in terms of the skewness and kurtosis of the returns rather than the shocks,

$$\sqrt{T_{\text{IS}}} \left(\widehat{\text{SR}}_{\text{IS}} - \text{SR} \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{1 + \phi}{1 - \phi} - \frac{1 + \phi + \phi^2}{1 - \phi^2} \gamma_{3,r} \text{SR} + \frac{1 + \phi^2}{1 - \phi^2} \frac{\gamma_{4,r} - 1}{4} \text{SR}^2 \right).$$

This result, together with a closely related derivation, has now been published

in [LLZ25]. We retain a proof in this chapter because their argument leaves implicit the justification for the asymptotic normality of the estimator; in Appendix 3.F.1 we make that step explicit by appealing to [PS92], while otherwise following a similar computation of the long-run covariance terms.

Special Cases

This general result subsumes several special cases from the literature, which are useful to consider for context and attribution.

For iid Gaussian returns, as analysed by Lo [Lo02], we set $\phi = 0$, $\gamma_{3,r} = 0$, and $\gamma_{4,r} = 3$. The variance in Proposition 3.A.1 simplifies to $1 + \frac{1}{2}\text{SR}^2$, recovering the classical result

$$\sqrt{T_{\text{IS}}} \left(\widehat{\text{SR}}_{\text{IS}} - \text{SR} \right) \xrightarrow{d} \mathcal{N} \left(0, 1 + \frac{1}{2}\text{SR}^2 \right).$$

In this case the exact distribution is also known and the sample Sharpe ratio is a scaled non-central t -distribution. However, in practice the Normal approximation is often preferable. The density of the conditional likelihood must be evaluated repeatedly when fitting the model via MLE, or performing inference for the posteriors of the true Sharpe ratios, but the non-central t -density is orders of magnitude slower to compute than the normal density, and is prone to numerical inaccuracy. Thus in almost all numerical calculations, we highly recommend using one of the Gaussian approximations to the distribution of the sample Sharpe ratio.

In a response to Lo, Mertens [Mer02] provides the correction to the variance when the returns have skewness and kurtosis. This corresponds to setting $\phi = 0$ in Proposition 3.A.1, which yields their result:

$$\sqrt{T_{\text{IS}}} \left(\widehat{\text{SR}}_{\text{IS}} - \text{SR} \right) \xrightarrow{d} \mathcal{N} \left(0, 1 - \gamma_{3,r}\text{SR} + \frac{\gamma_{4,r} - 1}{4}\text{SR}^2 \right).$$

More recently, Pav [Pav21, Section 3.1.4] provides the variance correction for

autocorrelated Gaussian returns. Setting $\gamma_{3,r} = 0$ and $\gamma_{4,r} = 3$ yields

$$\sqrt{T_{\text{IS}}} \left(\widehat{\text{SR}}_{\text{IS}} - \text{SR} \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{1 + \phi}{1 - \phi} + \frac{1 + \phi^2}{2(1 - \phi^2)} \text{SR}^2 \right).$$

A Study on Likelihood Choice

As noted, it is generally accepted that financial returns are not Gaussian, or iid [Con01].

We can study this on our own dataset of hedge fund returns. Figure 3.A.1 shows the distribution using violin plots of the autocorrelation ϕ , skewness $\gamma_{3,r}$ and excess kurtosis $\gamma_{4,r} - 3$ for the hedge fund returns used in Section 3.3. We observe that on average, fund returns exhibit little autocorrelation (note the returns are monthly frequency), and generally low skewness and kurtosis, with a few outliers taking extreme values.

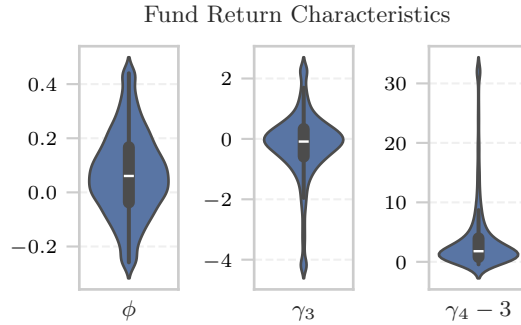


Figure 3.A.1: Distributions of autocorrelation ϕ , skewness $\gamma_{3,r}$ and excess kurtosis $\gamma_{4,r} - 3$ for the hedge fund returns used in Section 3.3.

The question is, how do these characteristics affect the distribution of sample Sharpe ratios? If the shape of the resulting distribution is significantly different from the iid Gaussian case, then failing to account for this will lead to poor parameter estimates. Indeed, in our own studies we found that in particular the model tended to converge towards extreme values for κ to manipulate the shape of the marginal likelihoods to match the observed data.

In Figure 3.A.2 we plot the distributions of the in- and out-of-sample Sharpe

ratios as a function of the autocorrelation ϕ , skewness $\gamma_{3,r}$ and kurtosis $\gamma_{4,r}$ for some example parameter values per Table 3.A.1. Starting with the most impactful, we see that the autocorrelation has the most effect on the distribution of in- and out-of-sample Sharpe ratios, with negative autocorrelation causing the distribution of out-of-sample Sharpe ratios to become more positively skewed, while positive autocorrelation causes the opposite effect. Next, we see that increasing skewness mildly affects the distribution of out-of-sample Sharpe ratios, but significantly less than the autocorrelation, and the kurtosis has little effect at all, even for the extreme value of $\gamma_{4,r} = 20$.

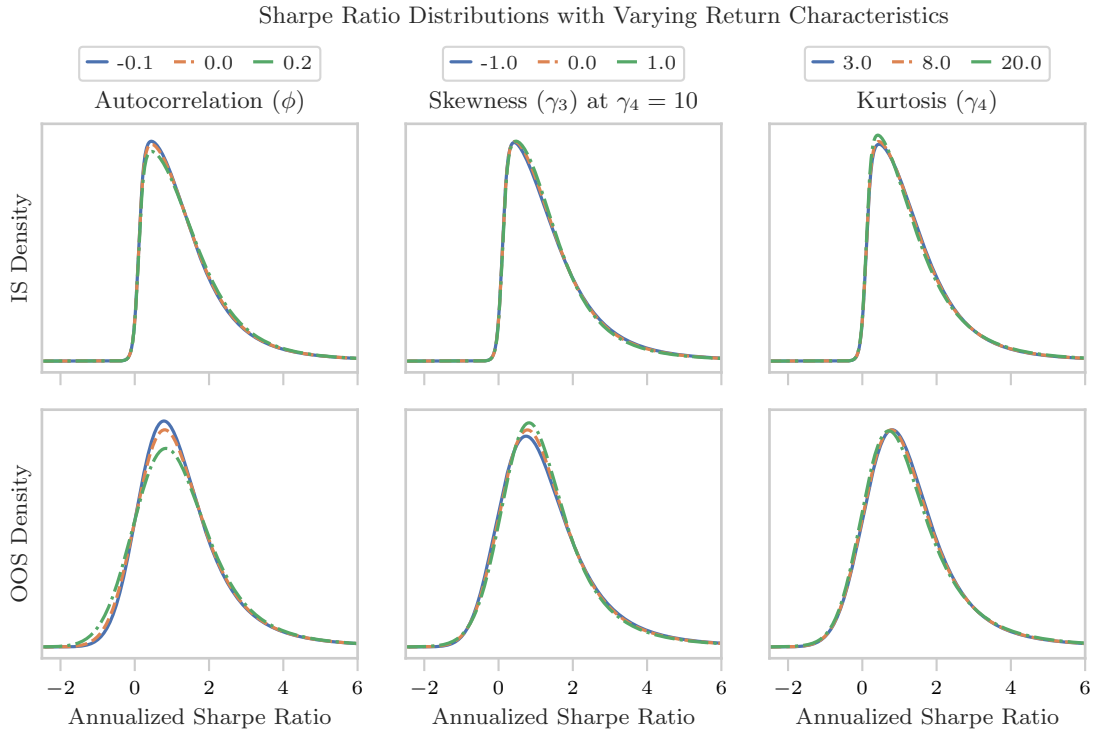


Figure 3.A.2: In- and Out-of-Sample Sharpe ratio distributions as functions of the autocorrelation of returns ϕ , skewness of returns $\gamma_{3,r}$ and kurtosis of returns $\gamma_{4,r}$.

Although the overall effect may seem small, we have found that in some cases the fitting procedure can be sensitive to the shape of the observed distribution of in- and out-of-sample Sharpe ratios. Therefore, if the fitting procedure with a Gaussian iid likelihood results in extreme parameter values, we recommend trying a likelihood

Prior	Selection	Sample	Decay
$SR_0 = 0.35/\sqrt{12}$	$\kappa = 0.1/\sqrt{12}$	$T_{IS} = 12 \cdot 5$	$\delta = 1$
$\varsigma_0^2 = 3.5/12$	$\ell = 50$	$T_{OOS} = 12 \cdot 5$	
$\nu = 3$			

Table 3.A.1: Prior, selection, sample and decay parameters for the experiments of Appendix 3.A.1.

function that captures the average autocorrelation of the returns. To this end, a single, representative autocorrelation value should be selected for the entire sample, rather than estimating a unique value for each fund. This approach is motivated by a crucial computational reason (see Appendix 3.B.2). The calculation’s complexity is driven not by the total number of funds, but by the number of unique likelihood parameterisations. Allowing the autocorrelation to vary by fund would make each parameterisation unique, slowing the computation significantly. Using a common parameter is therefore essential for tractability, as it ensures most funds share an identical likelihood function.

3.A.2 Other Types of Selection Bias

The main text focused exclusively on correcting for random selection bias. This is the kind studied by Dawid [Daw94] where he declared Bayesian statistics immune to selection bias when computing the posterior (but importantly not when fitting the prior), and the kind used by Chen and Zimmermann [CZ20] in their study of publication bias in the asset pricing literature. Here, we complete the picture by developing the equivalent Bayesian corrections for the other two forms: fixed and hierarchical bias. We begin by describing fixed and hierarchical selection bias, and then provide the necessary adjustments to the prior, likelihood and posterior for both cases.

Fixed Selection Bias

As with random selection bias, fixed selection bias is best understood through an example. In the random case, we act as a filter on a population: a fund only enters our dataset if its in-sample Sharpe ratio exceeds a certain threshold. While each fund's true skill level is fixed, the collective distribution of funds that survive this filter is skewed upward. This is intuitive: assuming some correlation between past and future performance, picking funds with high in-sample returns successfully weeds out the unskilled managers. In contrast, fixed selection bias distorts the view of a single manager rather than a group. Consider a manager who privately observes their own track record and decides to raise capital only if their performance looks attractive. Their true Sharpe ratio has not changed; they are the same manager they were before the backtest. However, because the data is only revealed conditional on success, the sample is no longer representative of their true performance. We are effectively observing a selectively reported sample rather than a random one, causing our estimate of their skill to be significantly biased.

Hierarchical Selection Bias

Hierarchical selection bias arises from a two-stage process, which is best explained by a quantitative researcher developing a trading strategy. First, the researcher picks a strategy type, such as momentum. This choice corresponds to drawing a family parameter ϑ from a hyperprior $\pi(\vartheta)$, where ϑ represents the mean of the true Sharpe ratio for all momentum strategies. Second, the researcher tests variations within that style, for example 3-, 6-, or 12-month momentum, and selects the best performer. Each variant has its own true Sharpe ratio SR drawn from a conditional prior $\pi(SR \mid \vartheta)$. Because they all depend on the shared family parameter ϑ , the strategies true Sharpe ratios are correlated. This two-level selection exemplifies hierarchical selection bias, and any form of parameter tuning is an example of this process.

To further clarify, let us make this mathematically precise. We take a location-scale t -distribution for both the hyperprior and conditional prior,

$$\begin{aligned}\vartheta &\sim \text{lst} \left(\text{SR}_0, \omega \varsigma_0^2 \frac{\nu_\vartheta - 2}{\nu_\vartheta}, \nu_\vartheta \right), \\ \text{SR} \mid \vartheta &\sim \text{lst} \left(\vartheta, (1 - \omega) \varsigma_0^2 \frac{\nu - 2}{\nu}, \nu \right).\end{aligned}$$

The parameter $\omega \in [0, 1]$ allocates the total variance ς_0^2 between the two levels. If $\omega = 1$, all variance is at the family level, the true Sharpe ratio is fixed at $\text{SR} = \vartheta$ for all variants, and we recover the fixed case. If $\omega = 0$, all variance is at the variant level, recovering the random case. The choice of ω therefore models our belief about the efficacy of parameter tuning. A high ω implies that tuning has little effect on the true Sharpe ratio, making the variants highly similar and thus increasing the risk of overfitting. Conversely, a low ω implies that tuning can produce genuinely distinct strategies, which lessens this risk.

More Bias Adjustments

As in the random case, we require the bias adjusted versions of the prior and likelihood for the fixed and hierarchical cases to perform model fitting and inference. The results below adapt those of Yekutieli [Yek12], replacing the original hard threshold with a soft threshold p_{acc} .

In the fixed case, the bias adjusted prior $\pi_B(\text{SR})$ is just $\pi(\text{SR})$, it does not require adjustment. By the definition of fixed selection bias, the act of selection does not affect the true Sharpe ratio, so no adjustment is needed to the prior. This is not the case for the conditional likelihood which does require a bias adjustment, and is the same as the bias adjusted conditional likelihood from the random case, $f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$. Therefore, unlike in the random case where the adjustments to the prior and the likelihood cancel when we compute the posterior, this does not occur in the fixed case, and the posterior is a more cautious, bias adjusted posterior.

The hierarchical case interpolates the two, and we denote the bias adjusted un-

conditional prior by $\pi_H(\text{SR})$. The bias adjusted likelihood is again $f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$. We collect the definitions of all the bias adjusted priors and likelihoods in Table 3.A.2.

	Prior	Likelihood
Unbiased	$\pi(\text{SR})$	$f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$
Random	$\pi_B(\text{SR})$	$f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$
Fixed	$\pi(\text{SR})$	$f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$
Hierarchical	$\pi_H(\text{SR})$	$f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})$

$$\pi_B(\text{SR}) = \frac{\mathbb{P}(\text{acc} \mid \text{SR})\pi(\text{SR})}{\mathbb{P}(\text{acc})},$$

$$f_B(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR}) = \frac{p_{\text{acc}}(\widehat{\text{SR}}_{\text{IS}})f(\widehat{\text{SR}}_{\text{IS}} \mid \text{SR})}{\mathbb{P}(\text{acc} \mid \text{SR})},$$

$$\pi_H(\text{SR}) = \mathbb{P}(\text{acc} \mid \text{SR}) \int_{\vartheta'} \frac{\pi(\vartheta')\pi(\text{SR} \mid \vartheta')}{\mathbb{P}(\text{acc} \mid \vartheta')} d\vartheta',$$

$$\mathbb{P}(\text{acc} \mid \text{SR}) = \int_{\widehat{\text{SR}}'_{\text{IS}}} p_{\text{acc}}(\widehat{\text{SR}}'_{\text{IS}})f(\widehat{\text{SR}}'_{\text{IS}} \mid \text{SR}) d\widehat{\text{SR}}'_{\text{IS}},$$

$$\mathbb{P}(\text{acc}) = \int_{\text{SR}'} \pi(\text{SR}')\mathbb{P}(\text{acc} \mid \text{SR}') d\text{SR}',$$

$$\mathbb{P}(\text{acc} \mid \vartheta) = \int_{\text{SR}'} \pi(\text{SR}' \mid \vartheta)\mathbb{P}(\text{acc} \mid \text{SR}') d\text{SR}'.$$

Table 3.A.2: Bias adjusted priors and likelihoods.

As with the random case, to perform model fitting and inference we simply substitute the correct bias adjusted prior and likelihood into the expressions to compute the marginal likelihood, Equation (3.2.4), and posterior, Equation (3.2.5).

3.B Implementation & Robustness

3.B.1 The Necessity of Joint Fitting, Alpha Decay & Parameter Recovery

As described in Section 3.2.2, we select a symmetric prior as the empirically observed distributions of out-of-sample Sharpe ratios are typically symmetric. This section expands on that by showing how fitting to in- and out-of-sample Sharpe ratios is key to identify selection bias, and additionally shows why including alpha decay is key for model fitting.

A skewed distribution of in-sample Sharpe ratios, viewed in isolation, is ambiguous. It could arise from a genuinely skewed prior (e.g. a Gamma distribution) or from selection bias applied to a symmetric one (e.g. a t distribution). As fig. 3.B.1 demonstrates, these two scenarios can produce nearly indistinguishable marginal likelihoods for the observed data. Without the out-of-sample evidence, we cannot tell them apart. But in the empirical study, Section 3.3, we observe a symmetric distribution in the out-of-sample Sharpe ratios, validating our choice.

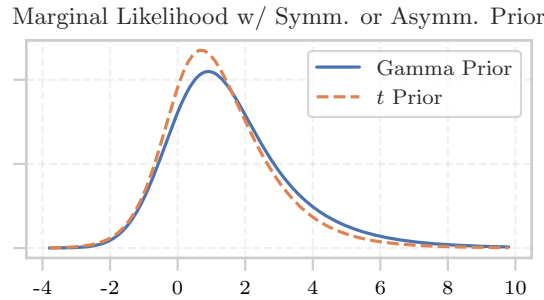


Figure 3.B.1: Indistinguishable marginal likelihoods: an unbiased gamma prior, $\Gamma(1, 10)$ versus biased selection with a t prior, $lst(-0.07, 0.02, 7.58)$ with selection parameters $\kappa = 0, \ell = 28$ (see (3.2.1)).

While enough in-sample data allows for the accurate recovery of parameters related to selection bias, we often lack sufficient samples. Further, we want the best

possible estimate of out-of-sample performance, which requires accounting for both selection bias and alpha decay (as discussed in Section 3.2.3). Fitting our model jointly to in- and out-of-sample Sharpe ratios addresses both problems.

To illustrate, we simulate 10,000 in- and out-of-sample Sharpe ratio pairs from a known distribution (Table 3.B.1) with hierarchical selection bias and perform three model fits: one using only in-sample data; a joint fit using both data sets; and a joint fit with the decay parameter fixed at $\delta = 1$. For all three we use a Gaussian iid likelihood. The results are in Figure 3.B.2 and Table 3.B.2. Fitting to in-sample data alone recovers most parameters well but cannot identify δ . (As an aside, the hierarchical model is the most difficult to fit; fixing ω helps significantly where possible.¹⁰) In contrast, the joint fit with a flexible decay parameter recovers all parameters well, successfully distinguishing selection bias from alpha decay by using the shape of the in-sample distribution and the location of the out-of-sample distribution. If we perform the joint fit but fix $\delta = 1$, the model converges to a poor fit with extreme parameter values, as the model is unable to explain the difference between the in- and out-of-sample performance by selection bias alone.

This test shows why the alpha decay parameter is essential: it is necessary for producing better out-of-sample estimates and for fitting the model to both data sets without generating nonsensical results.

While the above test with 10,000 samples recovered the model's parameters almost perfectly, practical applications rarely afford such a large dataset. To investigate the effect of sample size, we therefore generated a new set of 10,000 samples with random selection bias using the parameters from Table 3.B.3. From this population, we drew 1,000 subsamples of 100, 500, 1,000 and 5,000 in- and out-of-sample Sharpe ratio pairs. We then fitted the model to these subsamples using two methods: first, with in-sample Sharpe ratios only, and second, with both in- and out-of-sample

¹⁰ For example, if using the hierarchical model to model overfitting in algorithmic trading strategies, you may have a prior belief on the amount of influence that tuning parameters can have on the true Sharpe ratio of the resulting strategy. By selecting a ω value which gives the correct variance to the SR prior, you can model this.

3.B Implementation & Robustness

Prior	Selection	Sample	Decay
$SR_0 = 0.5/\sqrt{252}$	$\kappa = 1/\sqrt{252}$	$T_{IS} = 252 \cdot 3$	$\delta = 0.25$
$\varsigma_0^2 = 0.5/252$	$\ell = 100$	$T_{OOS} = 252 \cdot 3$	
$\nu = 10$			
$\nu_\vartheta = 10$			
$\omega = 0.5$			

Table 3.B.1: Prior, selection, sample and decay parameters for the experiments of Appendix 3.B.1.

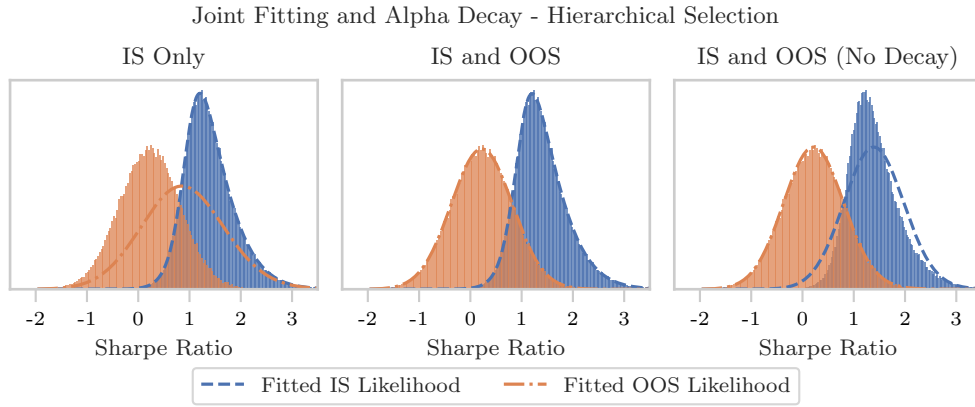


Figure 3.B.2: Comparing fitting with in-sample data only, in- and out-of-sample data, and in- and out-of-sample data without allowing for alpha decay.

	IS Only	IS + OOS	IS + OOS No Decay
SR_0	0.59	0.60	0.19
ς_0^2	0.48	0.47	0.01
ν	7.03	7.20	2.50
ν_ϑ	14.75	13.38	194.96
ω	0.62	0.59	0.01
κ	1.00	1.00	9.98
ℓ	100.73	100.32	55.96
δ	1.0000	0.2516	1.0000

Table 3.B.2: Fitted parameters from fitting with in-sample data only, in- and out-of-sample data, and in- and out-of-sample data without allowing for alpha decay.

Sharpe ratios. Again, we use a Gaussian iid likelihood for both.

Figure 3.B.3 plots the Kullback–Leibler (KL) divergence between the fitted and true models, with lower values indicating a better fit. We clearly see that fitting with both in- and out-of-sample data provides a significantly better fit than fitting with in-sample data alone. The in-sample only method requires 5,000 samples to match the quality that the joint method achieves with just 100.

As KL divergence is not particularly intuitive, Figure 3.B.4 shows the relative errors of the fitted parameters themselves. Using both in- and out-of-sample data not only reduces the variance of the estimates, but also removes the significant bias observed in the in-sample only method. The prior’s kurtosis parameter ν proves particularly difficult to recover from small samples under both methods, but improves quickly for with more samples when using the joint method.

Prior	Selection	Sample	Decay
$SR_0 = 0.5/\sqrt{252}$	$\kappa = 1/\sqrt{252}$	$T_{IS} = 252 \cdot 3$	$\delta = 1$
$\varsigma_0^2 = 0.5/252$	$\ell = 100$	$T_{OOS} = 252 \cdot 3$	
$\nu = 10$			

Table 3.B.3: Prior, selection, sample and decay parameters for the subsampling experiments using the random selection bias of Appendix 3.B.1.

Ultimately, accurate parameter recovery matters only if it yields better estimates of the true Sharpe ratio SR. To test this, for each in-sample Sharpe ratio we compute the posterior mean estimate $\mathbb{E}[SR \mid \widehat{SR}_{IS}]$ using the parameters from the in-sample-only fit and, separately, from the joint fit. As a benchmark, we also compute the posterior mean using the true data-generating parameters. We then calculate the Mean Absolute Error (MAE) between these estimates and the true Sharpe ratio for each simulated data point.

Figure 3.B.5 plots the average of these MAEs over our 1,000 subsamples. We observe that, on average, fitting with in- and out-of-sample data yields parameters sufficiently close to the true parameters that inference using them is almost identical.

3.B Implementation & Robustness

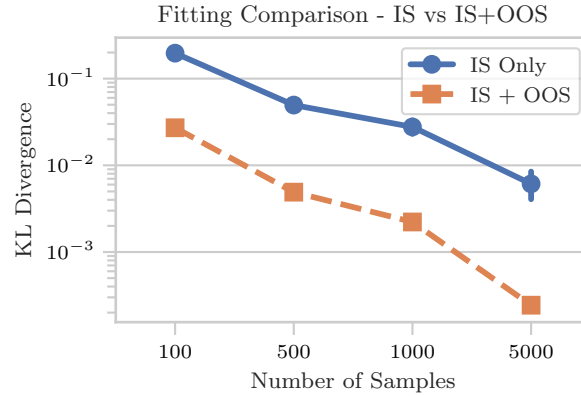


Figure 3.B.3: Kullback–Leibler divergence between the fitted and true models. The plot shows how fit quality varies with sample size when using only in-sample data versus using both in- and out-of-sample data.

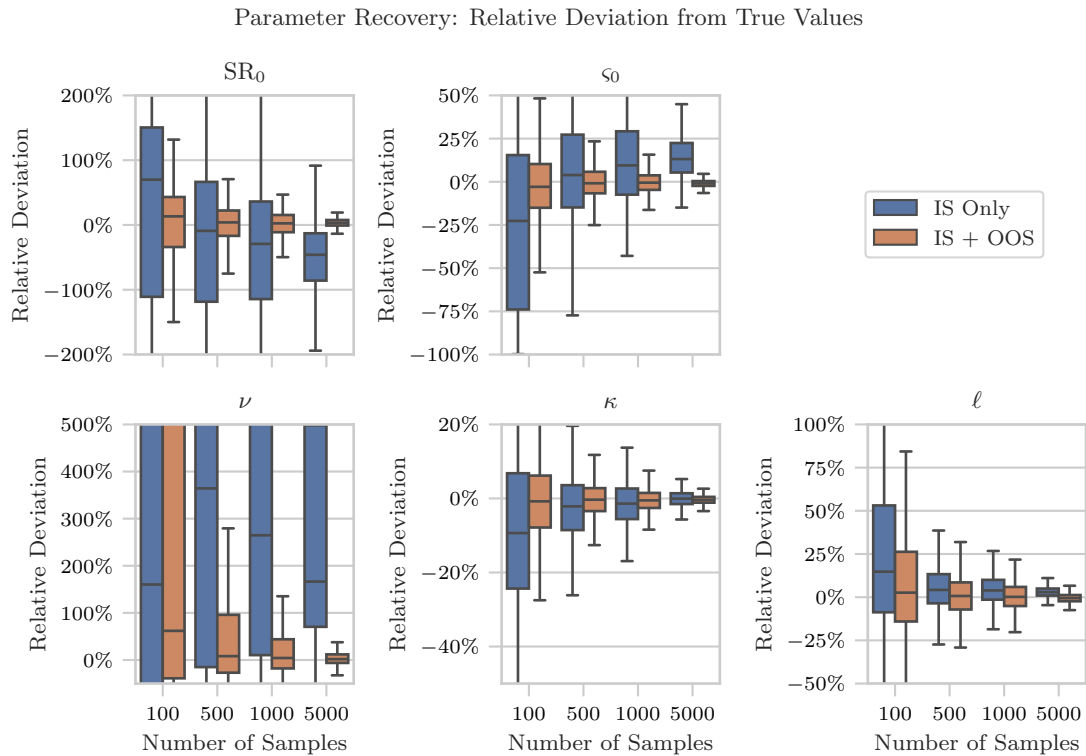


Figure 3.B.4: Relative errors in recovered parameters for the two fitting methods. The plots show how the bias and variance of the parameter estimates change with the number of samples used for fitting.

Fitting with in-sample data alone, however, produces a much higher error, up to 0.15 for small datasets, a significant amount when the true Sharpe ratios are of order 1.



Figure 3.B.5: Average Mean Absolute Error of the posterior mean $\mathbb{E}[\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}]$ vs the true Sharpe ratio SR per fit using in-sample only or in- and out-of-sample data.

3.B.2 Numerical Implementation Details

To use the model, one must fit its parameters via MLE and then compute the posterior. Both cases require computing a nested integral (up to a quadruple integral in the hierarchical case) as quickly as possible. Speed is particularly important for the fitting procedure, with potentially thousands of function evaluations required for convergence. The calculations must also be accurate and robust, especially against extreme parameter values, to prevent the optimiser from converging to a problematic fit, for example where the likelihoods explode due to numerical inaccuracy.

These two goals, speed and accuracy, are somewhat at odds. While standard adaptive quadrature methods, such as those in `scipy` [Vir+20], can accurately calculate the likelihood for a single sample, it is not fast enough for our purposes. To overcome this, we developed a specialised numerical package using the JAX library [FJL19] that combines several techniques to achieve high performance, accuracy, and robustness.

Using JAX provides four main benefits: Just-In-Time (JIT) compilation, GPU

computation, easy parallelisation and automatic differentiation. The first three are key to computing the likelihood rapidly, while the last allows us to pass an analytical gradient to the optimiser. For the optimiser, we utilise `iminuit` [Dem+25], a python wrapper of the C++ `Minuit2` library maintained by CERN [JR75]. We found this optimiser was much more successful in navigating the model’s difficult loss landscape than those available in `scipy`. Its success stems from its quasi-Newton algorithm, MIGRAD, which uses gradient information to intelligently approximate the local curvature (the Hessian matrix) of the loss surface. Supplying the precise analytical gradients computed by JAX allows this algorithm to perform an accurate line search, taking more efficient and reliable steps towards the minimum compared to methods that rely on numerical derivative estimates. This combination proves particularly effective for achieving rapid convergence in the complex, non-convex landscapes typical of statistical fitting problems. We also aid the optimiser by starting it at sensible parameter values; for instance, our initial guess for SR_0 is the mean of the out-of-sample Sharpe ratios.

Instead of adaptive quadrature, we use the Simpson rule for integration [Sim43]. While less sophisticated, this method is significantly faster and is easily compatible with automatic differentiation. However, we do dynamically adapt the integration grid to encompass the probability mass of the distribution being integrated. This approach, which uses the percentage point function (PPF), ensures the grid is always optimally placed to maximise accuracy. This is vital, as a wide, fixed grid would cause numerical issues for small values of ς_0 or values of ω close to zero or one. As JAX lacks a PPF for the t -distribution, we implement a minimal algorithm similar to Hill’s algorithm [Hil70] to compute the required quantiles. Specifically, we start with an initial guess from a Cauchy approximation when $\nu < 2$, otherwise we use a four-term Cornish-Fisher correction to the normal quantile [FC60]. We then apply a single Halley iteration [Hal94], which we find sufficient for our needs.

A significant optimisation comes from noting that the most computationally

intensive parts of the likelihood (e.g. $\mathbb{P}(\text{acc} \mid \text{SR})$, defined in eq. (3.2.2)) depend not on the sample Sharpe ratio $\widehat{\text{SR}}_{\text{IS}}$, but on the length of the in-sample period T_{IS} . Since a dataset typically contains only a few unique values of T_{IS} , we can precompute these expensive calculations once for each unique T_{IS} . This allows us to compute the marginal likelihoods for a very large number of samples almost as quickly as for one. The computation time thus scales with $O(N)$, where N is the number of unique T_{IS} values, not the total number of samples. We can often improve this further by rounding the $T_{\text{IS},i}$ values slightly to reduce the number of unique values. Conveniently, JAX enables easy parallel batch computation of these expensive components, significantly reducing the computation time per likelihood evaluation.

To avoid numerical underflow or overflow, we perform all calculations in the logarithmic domain. This technique improves numerical accuracy and is crucial when the divisors (e.g. $\mathbb{P}(\text{acc} \mid \text{SR})$) become very small. Without it, the likelihood could become numerically invalid, drawing the optimiser towards a problematic region. To prevent underflow or overflow when integrating our log-domain integrands (such as those in eqs. (3.2.2) to (3.2.4)), we use the log-sum-exp trick, factoring out the maximum value of the integrand before exponentiation. This ensures the optimiser receives stable, finite values even for likelihoods that would otherwise evaluate to zero in standard floating-point precision.

3.B.3 Hybrid Parameter Estimation

As introduced in Section 3.2.2, we found that for certain datasets the parameter set maximising the likelihood was “non-physical”, in the sense that the parameters lost their interpretability or became unintuitive. In particular, when the selection effect is weakly identified, for example when $\ell < 2$, the MLE parameter set Θ typically included an excessively large κ value. For instance, κ could be estimated at twenty times the prior location SR_0 , making the acceptance probability for all observed

Sharpe ratios, $p_{\text{acc}}(\widehat{\text{SR}}_{\text{IS}} \mid \kappa, \ell)$, effectively infinitesimal. This outcome does not align with an intuitive interpretation of the model. We attribute this issue to a combination of overfitting and the general difficulty of estimating weakly identified parameters, a well-known issue in Bayesian inference [Gel+13, Section 4.3][Gel+20, Section 5].

There are several ways to address this. One option is to add an L_1 or L_2 penalty to the optimisation problem, encouraging smaller parameter estimates. Alternatively, a traditional Bayesian approach could impose a weakly informative prior on parameters such as κ , acting as a soft penalisation on extreme values. Both approaches, however, require the choice of hyperparameters governing the degree of penalisation.

Instead, we adopt a hybrid strategy that combines Maximum Likelihood Estimation (MLE) and the Generalised Method of Moments (GMM). This is a form of two-step estimator [NM94, Section 6]: we split the parameter set Θ into those optimised via MLE, denoted $\widehat{\Theta}$, and those estimated via GMM, denoted $\widetilde{\Theta}$. At each step the MLE optimiser (Minuit in our case) proposes $\widehat{\Theta}$, which we then use to find $\widetilde{\Theta}$ satisfying the method-of-moments condition

$$m(\widetilde{\Theta} \mid \widehat{\Theta}) - \widehat{\mathbf{m}} = \mathbf{0}_K,$$

where $m(\widetilde{\Theta} \mid \widehat{\Theta}) \in \mathbb{R}^K$ denotes the model-implied moments and $\widehat{\mathbf{m}} \in \mathbb{R}^K$ denotes the empirical moments. This combines ideas of constrained maximum likelihood estimation [AS58] and profile likelihood [Mv00; CCG14]. In practice, we find this yields more realistic parameter estimates in cases where pure MLE produces extreme values. For example, in table 3.C.4 we see that the model fit using the MLE approach leads to the selection midpoint κ converging to the boundary set by the optimiser (10), while the MLE-GMM approach yields more reasonable parameter values.

To perform the GMM step we use the `optimistix` library [RLK24] to solve the

non-linear least-squares form of GMM,

$$\min_{\tilde{\Theta}} \left\| m(\tilde{\Theta} \mid \hat{\Theta}) - \hat{\mathbf{m}} \right\|_2^2.$$

We use the Levenberg-Marquardt solver as it is recommended as a good algorithm for many problems. The solver does not support bound constraints, so we apply a logistic reparameterisation to constrain parameters to reasonable regions, e.g. $\varsigma_0 > 0$. Specifically, to impose $a_i \leq \tilde{\Theta}_i \leq b_i$ we solve over unconstrained η_i with the mapping

$$\tilde{\Theta}_i(\eta_i) = a_i + (b_i - a_i) \sigma(\eta_i), \quad \sigma(x) = \frac{1}{1 + e^{-x}}.$$

As with MLE, we start each parameter at sensible, data-driven initialisations. Although Levenberg-Marquardt is theoretically more robust than Gauss-Newton, it did not converge well when supplied with very poor initial guesses (we initially used the midpoint of the bounded region before implementing data-driven initialisation). The starting values are

$$\begin{aligned} \text{SR}_0 &= \hat{\mathbb{E}}[\widehat{\text{SR}}_{\text{OOS}}], & \varsigma_0 &= \sqrt{\hat{\mathbb{V}}[\widehat{\text{SR}}_{\text{OOS}}] - 1/T_{\text{OOS}}}, & \delta &= \frac{\hat{\mathbb{E}}[\widehat{\text{SR}}_{\text{OOS}}]}{\hat{\mathbb{E}}[\widehat{\text{SR}}_{\text{IS}}]}, \\ \kappa &= \frac{\hat{\mathbb{E}}[\widehat{\text{SR}}_{\text{OOS}}] + \hat{\mathbb{E}}[\widehat{\text{SR}}_{\text{IS}}]}{2}, & \ell &= 5, & \nu &= 5, \end{aligned}$$

where $\hat{\mathbb{E}}, \hat{\mathbb{V}}$ denote the sample mean and variance respectively. If out-of-sample data is unavailable, in-sample data is used in its place, and ℓ and ν are started at uninformative initialisations due to difficulty in fitting these parameters via GMM. In practice, we recommend applying this methodology primarily to κ , δ and SR_0 .

For a chosen profile set $\tilde{\Theta}$ we include only the moments needed to inform that parameter. The mapping is shown in Table 3.B.4. In practice the dataset often contains many unique pairs $(T_{\text{IS}}, T_{\text{OOS}})$. To match empirical moments (computed over all samples) to analytical moments (which depend on T_{IS} and T_{OOS}), we compute each moment per pair and weight them by the observed frequencies of the pairs.

3.B Implementation & Robustness

Param	$\mathbb{E}[\widehat{\text{SR}}_{\text{IS}}]$	$\mathbb{E}[\widehat{\text{SR}}_{\text{OOS}}]$	$\mathbb{V}[\widehat{\text{SR}}_{\text{IS}}]$	$\mathbb{V}[\widehat{\text{SR}}_{\text{OOS}}]$	$\text{Cov}(\widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}})$	$\gamma_4(\widehat{\text{SR}}_{\text{IS}})$	$\gamma_4(\widehat{\text{SR}}_{\text{OOS}})$
SR_0	✓	✓	×	×	×	×	×
ς_0	×	×	✓	✓	×	×	×
κ	✓	✓	×	×	×	×	×
ℓ	×	×	✓	✓	×	×	×
δ	✓	✓	×	×	✓	×	×
ν	×	×	✓	✓	×	✓	✓

Table 3.B.4: Mapping between model parameters and empirical moments used in GMM estimation. When out-of-sample data is unavailable, all OOS moments are omitted. $\gamma_4(\cdot)$ denotes the kurtosis.

The hybrid MLE-GMM procedure stabilises weakly identified selection parameters and preserves interpretability without introducing additional penalisation hyperparameters. In our applications it consistently prevents extreme estimates for κ while retaining good in-sample fit for the likelihood-driven components. We therefore recommend profiling onto moment conditions for κ , δ and SR_0 , leaving the remaining parameters to be determined by MLE.

3.B.4 Effect of Correlated Sharpe Ratios

To see why independent samples are crucial for calibrating the prior, consider a simple model where returns are correlated through shared exposure to a market factor (for related discussion of cross-strategy dependence see [Lóp19; LL19; LLZ25]). We model the returns $r_{i,t}$ as a linear combination of this market factor $r_{M,t}$ and an idiosyncratic component $\epsilon_{i,t}$ representing idiosyncratic performance:

$$r_{i,t} = \rho r_{M,t} + (1 - \rho)\epsilon_{i,t},$$

where ρ controls the loading on the market, and the idiosyncratic returns are assumed to be independent. For $\rho = 0$, the returns are independent, and we can continue as normal, but for $\rho > 0$ the returns become correlated through their shared exposure to the market factor.

This correlation affects the cross-sectional distribution of sample Sharpe ratios used to fit the prior. Specifically, if the market performs unusually well (or poorly) relative to its true Sharpe ratio, the distribution of sample Sharpe ratios will be pushed upwards (or downwards), causing us to misestimate the location of the prior if we proceed naively. Additionally, because the returns are correlated, a naive estimate of the prior's variance will be too low; for high ρ , it can even yield negative estimates.

The true Sharpe ratio SR_i depends on the market and idiosyncratic components. Assuming the idiosyncratic variance is identical across assets, $\sigma_{\epsilon,i}^2 = \sigma_\epsilon^2$, the moments of a single return $r_{i,t}$ imply:

$$\begin{aligned}\mathbb{E}[r_{i,t}] &= \rho\mu_M + (1 - \rho)\mu_{\epsilon,i}, \\ \mathbb{V}[r_{i,t}] &= \rho^2\sigma_M^2 + (1 - \rho)^2\sigma_\epsilon^2, \\ SR_i &= \frac{\rho\mu_M + (1 - \rho)\mu_{\epsilon,i}}{\sqrt{\rho^2\sigma_M^2 + (1 - \rho)^2\sigma_\epsilon^2}}.\end{aligned}$$

Since μ_M is constant across assets, the cross-sectional variance $\varsigma_0^2 = \mathbb{V}[SR_i]$ depends solely on the variation in idiosyncratic means $\mu_{\epsilon,i}$. We parameterize the idiosyncratic Sharpe ratios $SR_{\epsilon,i} = \mu_{\epsilon,i}/\sigma_\epsilon$ with mean $SR_{\epsilon,0}$ and variance $\varsigma_{\epsilon,0}^2$, implying $\mathbb{V}[\mu_{\epsilon,i}] = \sigma_\epsilon^2\varsigma_{\epsilon,0}^2$.

Applying the variance operator to SR_i yields:

$$\varsigma_0^2 = \left(\frac{1 - \rho}{\sqrt{\rho^2\sigma_M^2 + (1 - \rho)^2\sigma_\epsilon^2}} \right)^2 \sigma_\epsilon^2\varsigma_{\epsilon,0}^2.$$

Simplifying this expression provides the final moments:

$$SR_0 = \frac{\rho\sigma_M SR_M + (1 - \rho)\sigma_\epsilon SR_{\epsilon,0}}{\sqrt{\rho^2\sigma_M^2 + (1 - \rho)^2\sigma_\epsilon^2}}, \quad \varsigma_0^2 = \frac{(1 - \rho)^2\sigma_\epsilon^2\varsigma_{\epsilon,0}^2}{\rho^2\sigma_M^2 + (1 - \rho)^2\sigma_\epsilon^2},$$

where $SR_M = \mu_M/\sigma_M$ is the market Sharpe ratio.

A naive way to estimate these parameters from a cross-section of N sample Sharpe ratios, $\left(\widehat{SR}_i\right)_{i \in [1, N]}$, uses the method of moments. We treat the observed

cross-sectional variance as the sum of the true parameter variance and the average estimation noise variance. To isolate the true variance, we subtract the noise component C_{Lo} :

$$\widehat{SR}_{0,N} = \widehat{\mathbb{E}} \left[\left(\widehat{SR}_i \right)_{i \in [1,N]} \right], \quad \text{and} \quad \widehat{\varsigma}_{0,N}^2 = \widehat{\mathbb{V}} \left[\left(\widehat{SR}_i \right)_{i \in [1,N]} \right] - C_{Lo},$$

where $\widehat{\mathbb{E}}[\cdot], \widehat{\mathbb{V}}[\cdot]$ represent the sample mean and variance of the sample Sharpe ratios $\left(\widehat{SR}_i \right)_{i \in [1,N]}$. We have assumed the returns are iid across time with no excess skewness or kurtosis to derive the correction term C_{Lo} , which accounts for the finite-sample variance of an individual sample Sharpe ratio, as given by Lo [Lo02, eq. 9],

$$C_{Lo} = \frac{1}{T_{IS}} \left(1 + \frac{1}{2} \widehat{SR}_{0,N}^2 \right).$$

These estimators perform poorly as they ignore cross-sectional correlation. We can instead construct “market-aware” estimators using observable market data. For the variance, we scale the correction factor to remove only the idiosyncratic portion of the estimation noise, since systematic error does not contribute to cross-sectional dispersion. Indeed, from the expression for ς_0^2 above,

$$\varsigma_0^2 = \frac{(1-\rho)^2 \sigma_\epsilon^2}{\rho^2 \sigma_M^2 + (1-\rho)^2 \sigma_\epsilon^2} \varsigma_{\epsilon,0}^2,$$

so the idiosyncratic share of total variance is $\frac{(1-\rho)^2 \sigma_\epsilon^2}{\rho^2 \sigma_M^2 + (1-\rho)^2 \sigma_\epsilon^2}$. We therefore subtract the same fraction of Lo’s correction:

$$\widehat{\varsigma}_{0,A}^2 = \widehat{\mathbb{V}} \left[\left(\widehat{SR}_i \right)_{i \in [1,N]} \right] - C_A, \quad \text{with} \quad C_A = \frac{(1-\widehat{\rho})^2 \widehat{\sigma}_\epsilon^2}{\widehat{\rho}^2 \widehat{\sigma}_M^2 + (1-\widehat{\rho})^2 \widehat{\sigma}_\epsilon^2} C_{Lo},$$

where $\widehat{\sigma}_M^2$ is the sample variance of market returns, $\widehat{\sigma}_\epsilon^2$ is the sample variance of idiosyncratic returns and $\widehat{\rho}$ an estimate for ρ . These parameters are readily estimable by regressing individual returns on the observed market returns.

For SR_0 we are not so fortunate. While a market aware correction can be calcu-

lated,

$$\widehat{\text{SR}}_{\text{IS},0,A} = \widehat{\text{SR}}_{\text{IS},0,N} - \frac{\rho\sigma_M(\widehat{\text{SR}}_M - \text{SR}_M)}{\sqrt{\rho^2\sigma_M^2 + (1-\rho)^2\sigma_\epsilon^2}},$$

it depends on the true market Sharpe ratio, SR_M , which is unknown. Using our best estimate, $\widehat{\text{SR}}_M$, reduces the correction term to zero, leaving us with the naive estimator.

To demonstrate these effects we perform a straightforward simulation. We generate 1,000 market paths and 10,000 idiosyncratic return series of length $T_{\text{IS}} = 252 \cdot 3$ from the distributions

$$r_{M,t} \sim \mathcal{N}(\text{SR}_M \cdot \sigma, \sigma^2), \quad \epsilon_{i,t} \sim \mathcal{N}(\text{SR}_{\epsilon,i} \cdot \sigma, \sigma^2),$$

where to reduce the number of parameters we have set the market and idiosyncratic volatilities to be equal $\sigma_M = \sigma_\epsilon = \sigma$, the distribution of idiosyncratic Sharpe ratios is $\text{SR}_{\epsilon,i} \sim \mathcal{N}(\text{SR}_{\epsilon,0}, \varsigma_{\epsilon,0}^2)$ and the parameters are set as:

$$\begin{aligned} \text{SR}_M &= 0.75/\sqrt{252}, & \sigma &= 0.1/\sqrt{252}, \\ \text{SR}_{\epsilon,0} &= 0, & \varsigma_{\epsilon,0}^2 &= 0.5/252. \end{aligned}$$

For each market path, we construct a set of $N = 10,000$ returns and estimate SR_0 and ς_0^2 using the naive and market-aware methods. The outcome is shown in Figure 3.B.6. We see that as ρ increases, the distribution of errors, $\widehat{\text{SR}}_0 - \text{SR}_0$, in our estimate of SR_0 increases, but across all the market paths, the naive estimate is unbiased. For ς_0^2 , we see that the naive estimate $\widehat{\varsigma}_{0,N}^2$ yields impossible negative estimates for the variance of the distribution of SR_i . But with the corrected estimator $\widehat{\varsigma}_{0,A}^2$ we perfectly estimate the true variance, and in fact the error per market path is very small, due to the large number of returns and reasonably long time series used to estimate the inputs to the corrected estimator.

Ultimately, this was the simplest possible case, and once we introduce selection bias the process of yielding good estimates for the prior parameters would become

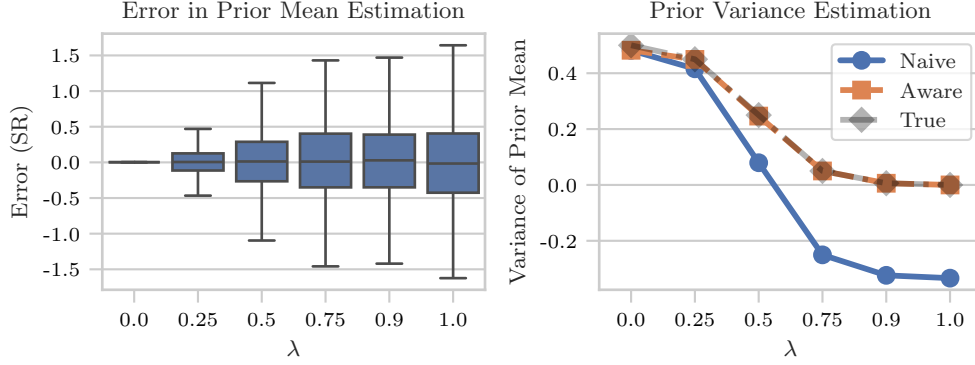


Figure 3.B.6: The impact of correlation, controlled by ρ , on the estimation of SR_0 and ς_0^2 .

much more complex, if not intractable. This difficulty, combined with the inability to produce a market-aware correction for the prior's location, suggests the most robust approach is to work with the idiosyncratic component directly. Should the overall Sharpe ratio be required, the systematic contribution can be added back to the corrected estimate of idiosyncratic performance. We can express the overall Sharpe ratio in terms of the component Sharpe ratios,

$$SR_i = \frac{\rho SR_M \sigma_M + (1 - \rho) SR_{\epsilon,i}^* \sigma_{\epsilon,i}}{\sqrt{\rho^2 \sigma_M^2 + (1 - \rho)^2 \sigma_{\epsilon,i}^2}},$$

where $SR_{\epsilon,i}^*$ denotes our corrected estimate for the idiosyncratic Sharpe ratio, and, as previously discussed, the remaining parameters are reliably estimated directly from the data.

3.C Supplementary Material for Section 3.3

3.C.1 Filter Criteria

To filter down the initial large dataset into a set of suitable funds for our approach we apply the following criteria to the dataset, yielding a final total of 7,303 funds:

1. Maximum AUM must be $> \$20m$,

2. At least 24 months of returns,
3. At least 6 months of returns in both pre- and post-listing periods,
4. Monthly volatility must be $> 0.5\%$,
5. Annualised Sharpe ratio must be < 10 ,
6. All monthly returns must be within $[-90\%, 500\%]$,
7. De-duplicate highly similar funds: for any pair with absolute correlation $> 95\%$, retain the fund with the longer history.

These filters are set to exclude very small hedge funds, potential outliers and funds with insufficient history.

3.C.2 Deriving Independent Returns

As described in Section 3.2.3, we require the samples used to fit the Bayesian model to be (mostly) independent to avoid issues with estimating the parameters correctly (see Appendix 3.B.4 for the impact of this assumption being violated). For hedge fund returns, this poses an immediate issue, as many funds are relatively correlated through exposure to shared risk factors. To resolve this, we split the returns for each fund into a “systematic” part and a “non-systematic” part.

We define systematic returns as the returns earned through passive (as opposed to active) exposure to known risk factors, and define all other returns as non-systematic. Specifically, we allow factor timing to be considered skill on the part of the fund manager, in addition to returns earned through manager specific non-factor explainable returns. To estimate these components we fit a single regression model for each fund,

$$r_{i,t} = \underbrace{\alpha_i + \beta_i^{\text{Timing}} \mathbf{F}_t^2}_{\text{Non-Systematic Part}} + \underbrace{\beta_i^{\text{Passive}} \mathbf{F}_t}_{\text{Systematic Part}},$$

where we have taken the approach of Treynor-Mazuy [TM66] by adding the squared factor returns to the regression to capture factor timing skill.

For our factors, we take a combination of the Fung-Hsieh seven factor model [FH04], a collection of global factors Barth et al. [Bar+23], and the factor set proposed by Chen et al. [Che+25], yielding a total of 20 factors. Once we add the squared factors and the intercept, this is a total of 41 possible parameters per regression.

This rules out two approaches. First, we cannot use simple OLS, as funds with short histories often have relatively few data points. Similarly, we do not take a walk-forward approach for the same reason as we wish to use as much data as possible for our regression and later analysis. However, by not using a walkforward approach we do introduce lookahead bias, as the non-systematic component represents perfectly hedged returns that would be impossible to capture in reality. Relatedly, in our later analysis we assume investors may trade in any fund in any amount, ignoring lockups, closures to new capital, and other constraints.

Our intention, however, is to provide a proof of concept: it is possible to fit our model to data exhibiting selection bias and realistic statistical effects, such as autocorrelation and fat tails. We therefore advise the reader to interpret Sharpe ratios derived herein with a grain of salt. The central aim is to demonstrate that the Bayesian method outperforms naive benchmarks (which also benefit from these same biases).

Returning to our model estimation problem, we take a similar approach to [Che+25], and estimate the parameters for each fund using an adaptive LASSO. This neatly resolves our issues of identifying a well fitting model for each fund.

Specifically, for each fund, we first fit a ridge regression using `sklearn` [Ped+11], tuning the degree of regularisation via Leave-One-Out Cross-Validation. We then use these coefficients for the penalty weights of the adaptive LASSO step, where the adaptive LASSO is performed using `asgl` [MAL21]. To tune the degree of regularisation in the adaptive LASSO, we minimise the extended regularized information criterion (ERIC) which has been shown to outperform more traditional information criteria (AIC/BIC) for adaptive LASSO [HWF15].

We find that this procedure works well, and yields reasonable parameter sets across the dataset. In particular, by examining which factors each fund loads on, we see an expected pattern: the majority of funds load on either the global or US market factor. In Figure 3.C.1 we see that more than 40% of funds load on the global market factor, and more than 30% of funds load on the US market factor, with 65% of funds loading on at least one market factor.

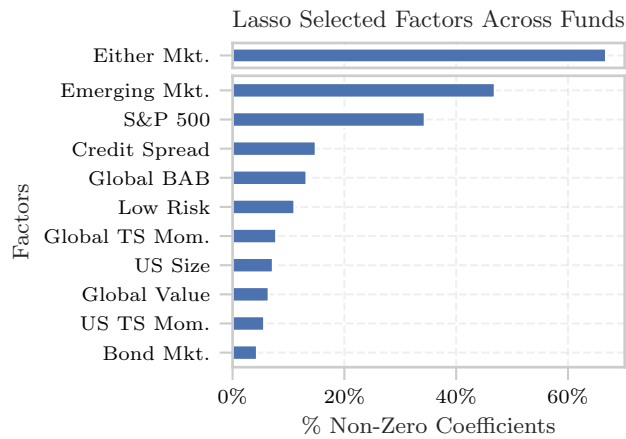


Figure 3.C.1: Proportion of funds with non-zero loading on the top 10 factors, and the proportion of funds that load on either of the two market factors. BAB denotes “Betting Against Beta” [FP14] and TS Mom. denotes “Time Series Momentum” [MOP12].

Importantly for our use case, we find this significantly reduces the average correlation between funds, with the mean correlation reducing from 28% for the total returns to 8% for the non-systematic returns, and the distribution of correlations significantly improving, as can be seen in Figure 3.C.2.

What is the impact of this procedure on the Sharpe ratios of the funds? Figure 3.C.3 shows the distributions of pre- and post-listing Sharpe ratios for the raw and non-systematic returns. On average, the annualised pre-listing Sharpe ratio falls from about 1.35 on the raw returns to 1.05 after hedging, so the impact on pre-listing performance is noticeable but relatively muted, with a small shift of the mass of the distribution to the left and the overall shape largely preserved. By contrast,

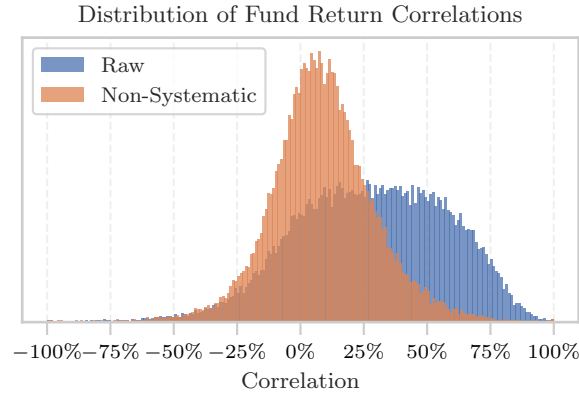


Figure 3.C.2: Pairwise correlations between fund returns, before and after removal of systematic factor contributions.

the mean post-listing Sharpe ratio falls from around 0.45 to 0.25 once we remove systematic components, and the distribution of non-systematic post-listing returns becomes significantly wider and much more tightly centred around zero. Quantitatively, the average reduction in Sharpe from pre- to post-listing is about 67% for the raw returns and 76% for the hedged returns, indicating that the deterioration in performance is larger once we strip out systematic exposures. This suggests that pre-listing returns tend to be less systematic, while post-listing returns are more so, which is consistent with the idea that funds scale up after listing and therefore take on more systematic risk due to capacity constraints on non-systematic strategies.

In Figure 3.C.4 we plot the performance of a simple inverse volatility weighting portfolio of the funds in the selected sample using either the non-systematic returns or the raw returns. The weights are computed by taking the inverse volatility of each fund over its entire history, and normalised such that the weights each month sum to 1, with the final performance normalised to unit volatility. Curiously, we see that the performance divides neatly into three periods: pre September 2001, between September 2001 and July 2008, and post July 2008. Pre-2008, the non-systematic portfolio significantly outperforms the raw portfolio, but post 2008 slightly underperforms. In the first period, the non-systematic portfolio yields an impressive 5.0

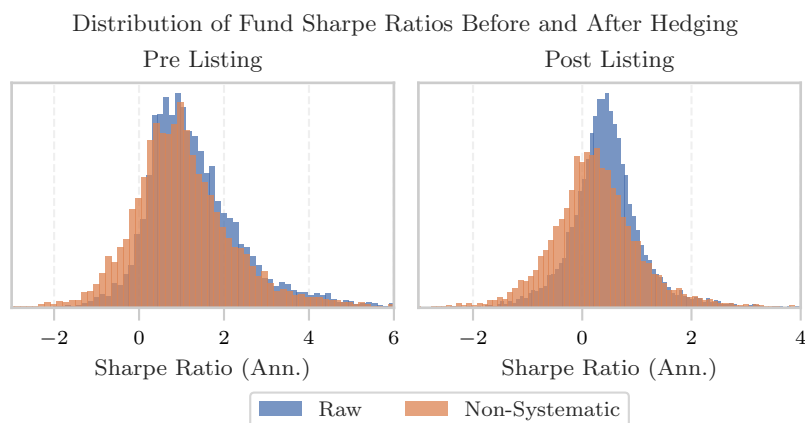


Figure 3.C.3: Distribution of Sharpe ratios before and after removal of systematic factor contributions, split into the pre- and post-listing period for each fund.

Sharpe ratio, falling to 2.9 in the second period and finally just 0.6 in the post-2008 period. By comparison, the raw portfolio has Sharpe ratios of 3.1, 2.4 and 0.7 respectively in the three periods.

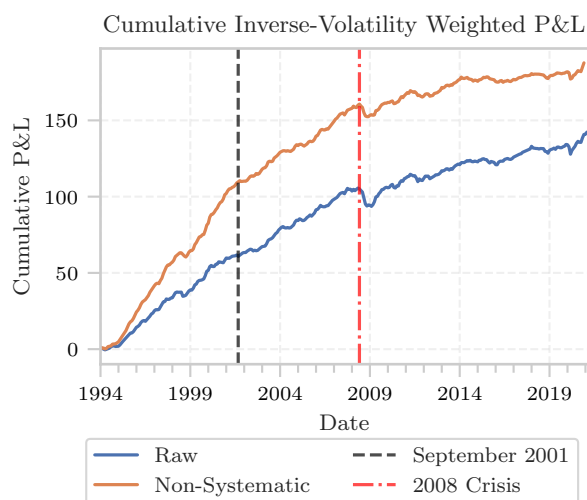


Figure 3.C.4: Cumulative P&L of inverse volatility weighted combination of raw and non-systematic returns.

3.C.3 Fitting Approach

For the non-walkforward analysis, we employed the hybrid MLE-GMM approach detailed in Appendix 3.B.3. We found that a pure MLE approach often converged to non-physical regions, where κ diverged to implausibly large values (e.g., $\kappa = 10$ for the first and second periods). This necessitated the hybrid approach, which identifies parameters of reasonable magnitude that still fit the data well. For example, for the first period, the MLE fitted model has a negative log likelihood (NLL) of 891. The MLE-GMM fitted models have NLLs of 894, 898, and 918 when profiling out κ , (κ, δ) , or $(\kappa, \delta, \text{SR}_0)$ respectively, a negligible difference in visual fit.

We ran the fitting procedure using $\{\kappa\}$, $\{\kappa, \delta\}$, or $\{\kappa, \delta, \text{SR}_0\}$ as the GMM profile set. We selected these parameters as they proved reliable for profiling, unlike ℓ , which is difficult to fit via GMM. Our experiments identified $\{\kappa, \delta\}$ as the most “stable” set; that is, it yielded the parameters closest (in L^2 distance) to $\bar{\Theta}$, the mean of the vectors derived from the three profile sets. Consequently, we use $\{\kappa, \delta\}$ as the profile set for the MLE-GMM analysis.

3.C.4 Fund Clustering

In addition to the returns being suitably independent, we also require that the samples used to fit the Bayesian model are representative of the samples we wish to perform inference on. Naturally, hedge funds are a good proxy for other hedge funds, so one approach is to simply fit the model to all the funds in our dataset, which is the approach taken in the main text. But we can also consider that not all funds behave the same with regard to pre- and post-listing performance. Therefore, we would like to somehow group similar funds together, and investigate if these groups behave differently within our model.

To do this, we cluster our funds together in a two-step process. First, we compute an extensive list of metrics (see Table 3.C.1), which we then pass through the factor

analysis of mixed data (FAMD) method to reduce the categorical and quantitative variables to a lower dimensional representation (FAMD is similar to PCA, but allows qualitative in addition to quantitative variables, we use the `prince` library [Hal]), selecting the first 34 components, which explain around 80% of the total variance, as we see in Figure 3.C.5.

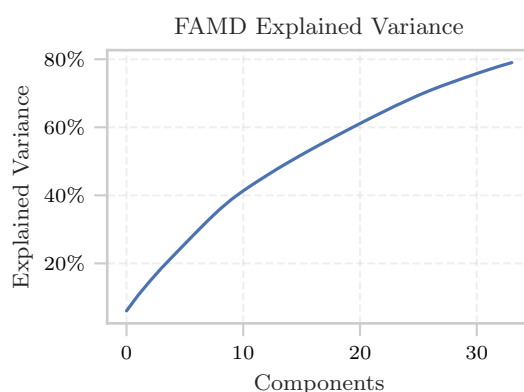


Figure 3.C.5: Explained variance by number of components for the FAMD analysis.

Fund Attributes	Performance	Factor Exposure
• Firm & fund domicile • Strategy style • Fees & high-water mark provision • Lockup, notice period, redemption frequency • Leverage (indicator) and log leverage • Launch, peak & close log AUM • Initial, listing & close year • Sample length: pre-listing, post-listing, total	• Pre-, post-listing & overall Sharpe ratio • Mean, standard deviation, skewness and kurtosis • Maximum drawdown • Replication ratio (OOS/IS SR) • Stability of Sharpe (“Sharpe-of-Sharpe”)	• Overall loading magnitude: $\log L_2$ of non-timing loadings; $\log L_2$ of timing loadings; ratio of non-timing to timing loadings • Signed exposures to systematic factors (with squared terms for timing)

Table 3.C.1: Qualities supplied to FAMD, grouped by theme. All performance qualities are computed using both the raw and non-systematic returns.

We then perform hierarchical clustering using Ward linkage [War63], and select the number of clusters which maximises the silhouette score [Rou87]. For our dataset, this is 26 clusters, seen in Figure 3.C.6.

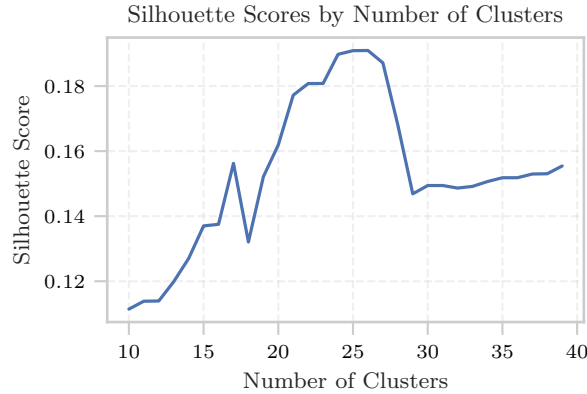


Figure 3.C.6: Silhouette score by number of clusters.

Many of these clusters are too small for our approach; without sufficient data, the Bayesian parameter estimates would be unreliable. To resolve this, we compute centroids for each cluster and merge those with fewer than 300 funds into the nearest cluster exceeding that threshold. This results in nine final clusters: three large and six small. Figure 3.C.7 displays the distribution of fund counts across clusters.

To identify which features are important in cluster allocation, we fit a boosting model to predict the cluster of each fund using the original mixed categorical and quantitative metrics (specifically, we use `catboost` [Pro+19] which natively supports categorical variables). We can then compute SHAP values [LL17; Lun+20] for the model to identify which original features are most important in cluster allocation. Figure 3.C.8 plots the features SHAP values in order of importance. Reassuringly, we observe that the fund style is the most important feature, followed by domicile, liquidity terms, leverage, vintage and fees.

As we only have a relatively small number of final clusters, we can easily check the composition of each cluster and identify roughly what each cluster corresponds to. Table 3.C.2 shows the breakdown of each cluster by investment style and fund

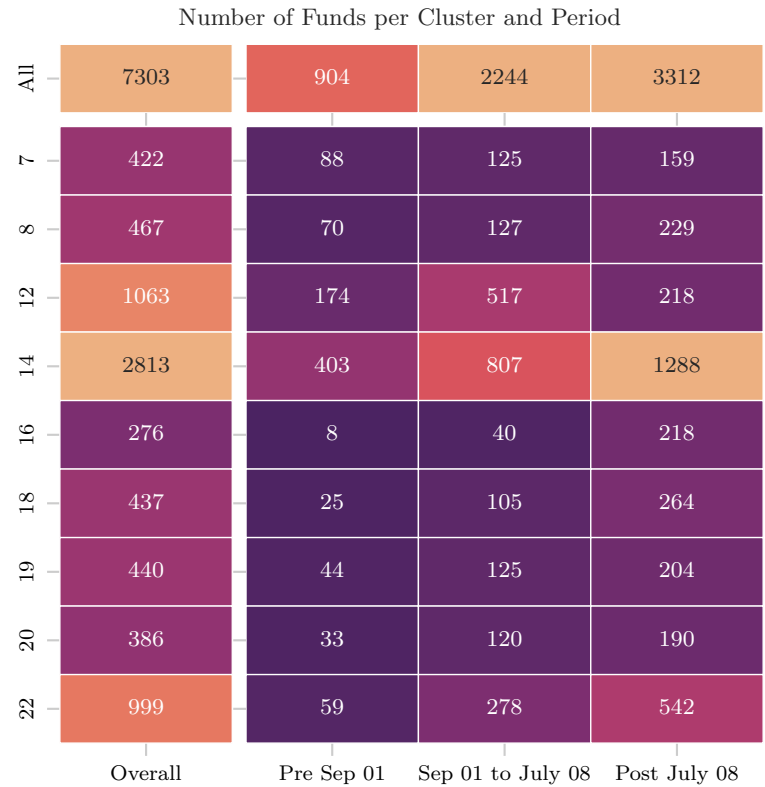


Figure 3.C.7: Number of funds per cluster, and per period

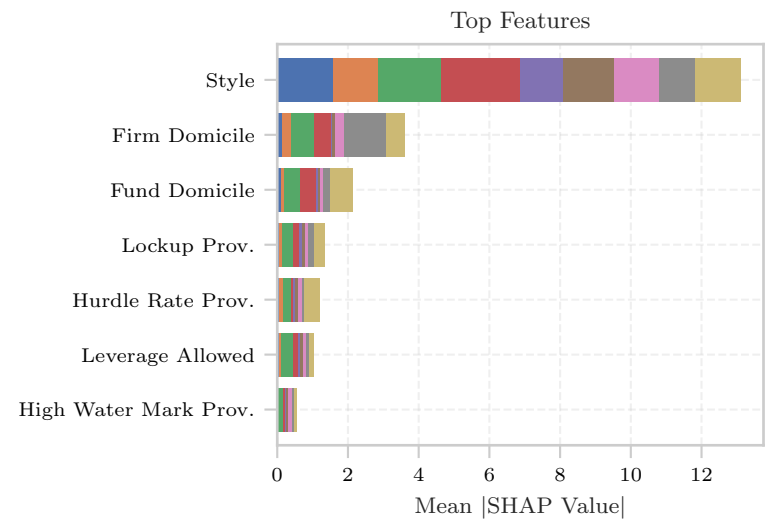


Figure 3.C.8: Top features for cluster selection, sorted by mean |SHAP Value|.

3.C Supplementary Material for Section 3.3

domicile. Most clusters consist primarily of a single style, while three clusters contain a mix of strategies, with equity being the dominant style in these clusters, as can be seen in Figure 3.C.9.

Cluster	Style	Fund Domicile
7	Event Driven	Caribbean/North America
8	Relative Value (RV)	Caribbean/NA/EU
12	Fund of Funds (FoF)	Caribbean/North America
14	Equity + Multi-Strat	Caribbean/North America
16	Credit	Caribbean/NA/EU
18	Macro	Caribbean/NA/EU
19	CTA	Caribbean/NA/EU
20	Equity + FoF + CTA + RV	Caribbean/EU
22	Equity + FoF	Europe

Table 3.C.2: Cluster composition. Note most clusters are primarily US firm domicile, bar cluster 22 which is European.

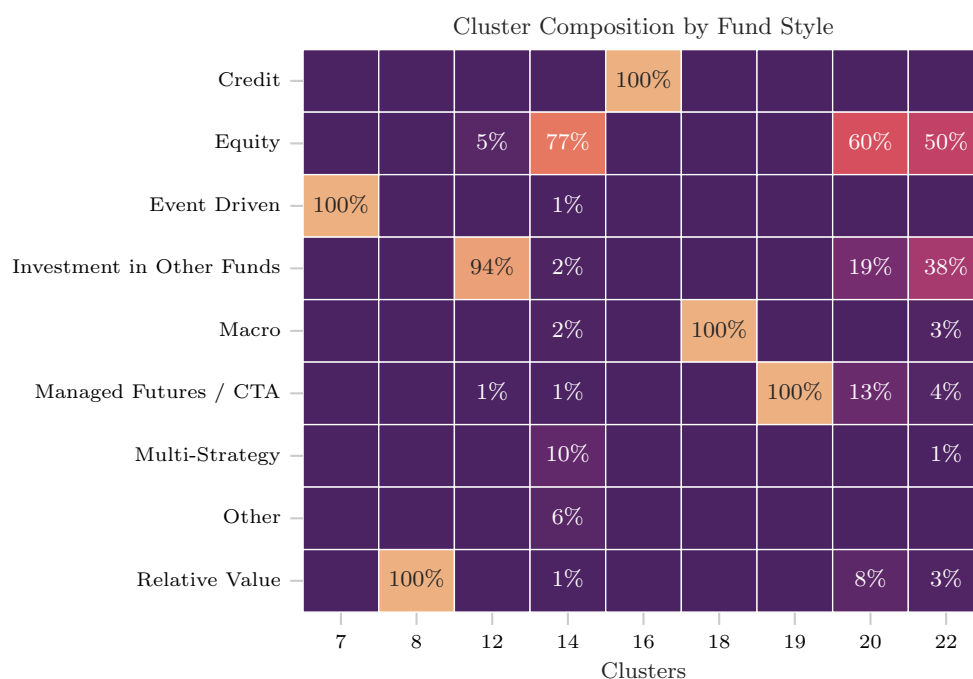


Figure 3.C.9: Cluster composition by style.

As our primary focus is the pre-listing and post-listing performance for each fund,

we should check if the clusters have different distributions, as if they simply had the same distribution we could fit one model and use it for all funds. In short, they do indeed have quite different distributions. Figure 3.C.10 shows the distributions of pre- and post-listing annualised Sharpe ratios for each cluster.

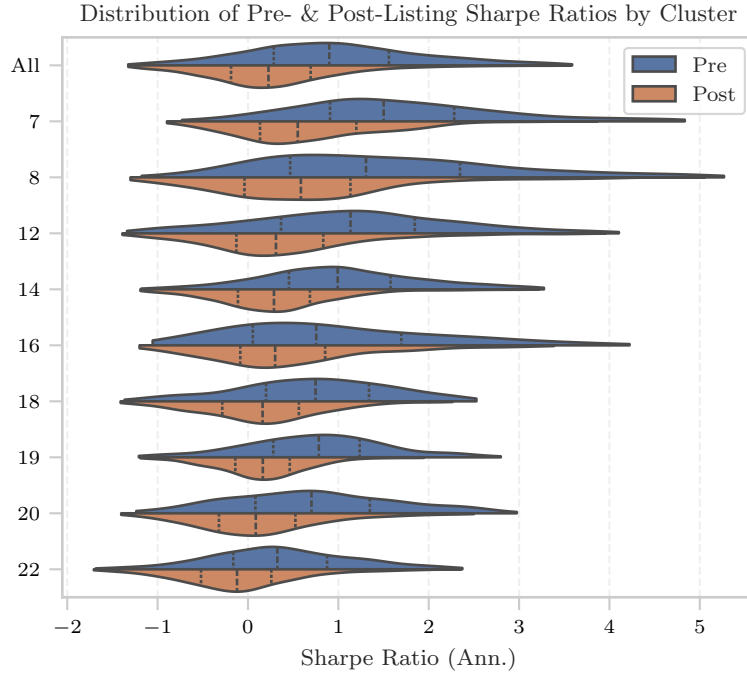


Figure 3.C.10: Pre- and Post-Listing Sharpe ratio distribution by cluster.

As discussed in the main text, we split the data into three periods defined by the regimes observed in Appendix 3.C.2: pre-September 2001, September 2001 to July 2008, and post-July 2008. The first two periods have insufficient sample sizes for cluster-specific fitting. Consequently, we analyse cluster-level fits only for the full time series and the post-2008 period.

We report the parameter estimates per cluster in the overall and post 2008 periods in Table 3.C.3. We can draw a few inferences directly from the estimated parameter estimates. Firstly, the amount of alpha decay is quite high across all clusters, with δ generally below 0.6. Secondly, the selection effect varies across clusters, but is generally quite soft, with $\ell < 10$. Thirdly, the distribution of true

Sharpe ratios is estimated to have fat tails across almost all clusters, with ν typically < 6 .

Whole Sample						
	SR_0	ς_0	ν	κ	ℓ	δ
All Funds	-0.07	1.00	3.44	1.73	4.34	0.39
Cluster 7	0.83	0.97	3.25	4.12	2.82	0.40
Cluster 8	0.31	1.14	5.78	6.48	2.19	0.53
Cluster 12	0.55	1.49	3.23	0.80	2.55	0.37
Cluster 14	0.28	0.66	4.09	3.60	3.39	0.41
Cluster 16	0.17	1.04	3.64	5.20	1.98	0.58
Cluster 18	0.00	0.73	14.25	3.60	2.85	0.54
Cluster 19	0.52	0.83	3.84	-0.24	4.32	0.30
Cluster 20	-0.35	0.69	4.31	2.32	4.37	0.61
Cluster 22	-0.17	0.73	2.86	2.88	2.47	0.49
Post 2008						
	SR_0	ς_0	ν	κ	ℓ	δ
All Funds	-0.03	0.73	5.21	4.12	2.50	0.55
Cluster 7	0.35	0.59	5.96	4.64	3.43	0.51
Cluster 8	-0.10	0.97	6.10	6.03	2.81	0.61
Cluster 12	0.31	1.39	4.71	-1.28	12.32	0.60
Cluster 14	-0.03	0.87	2.26	1.39	4.77	0.56
Cluster 16	0.35	2.36	2.43	-1.04	7.08	0.56
Cluster 18	-0.28	0.62	9.85	2.70	2.81	0.73
Cluster 19	0.35	0.73	3.93	6.27	1.15	0.26
Cluster 20	-0.17	0.69	3.71	1.70	3.27	0.65
Cluster 22	-0.24	0.52	2.94	5.65	1.21	0.83

Table 3.C.3: Posterior parameter estimates by cluster and period. SR_0 , ς_0 and κ are reported in annualised terms, divide by $\sqrt{12}$ to yield the original parameter estimates.

Finally, we can look at the posterior estimates $\mathbb{E}[SR_{\delta,i} | \widehat{SR}_{IS,i}, \widehat{SR}_{OOS,i}]$. These reflect our best Bayesian guess for the true Sharpe ratio for each fund, which we can decompose into a haircut due to Bayesian shrinkage and selection bias (these two effects are intertwined, as previously discussed) and alpha decay. Figure 3.C.11 shows this decomposition. For each fund we compute the posterior estimate for

the true Sharpe ratio $\mathbb{E}[\text{SR}_{\delta,i} | \widehat{\text{SR}}_{\text{IS},i}, \widehat{\text{SR}}_{\text{OOS},i}]$, and then apply the decay factor δ . We then compute the haircut vs the pre-listing return per fund, $\mathbb{E}[\text{SR}_{\delta,i}] / \widehat{\text{SR}}_{\delta,i}$ and $\delta \mathbb{E}[\text{SR}_{\delta,i}] / \widehat{\text{SR}}_{\text{IS},i}$, and plot the medians of these values. We observe that the split between initial shrinkage and alpha decay varies significantly by cluster, with some clusters being dominated by alpha decay (for example, Cluster 7, Event driven funds) while others are primarily shrinkage dominated with relatively little alpha decay (for example, Cluster 20).

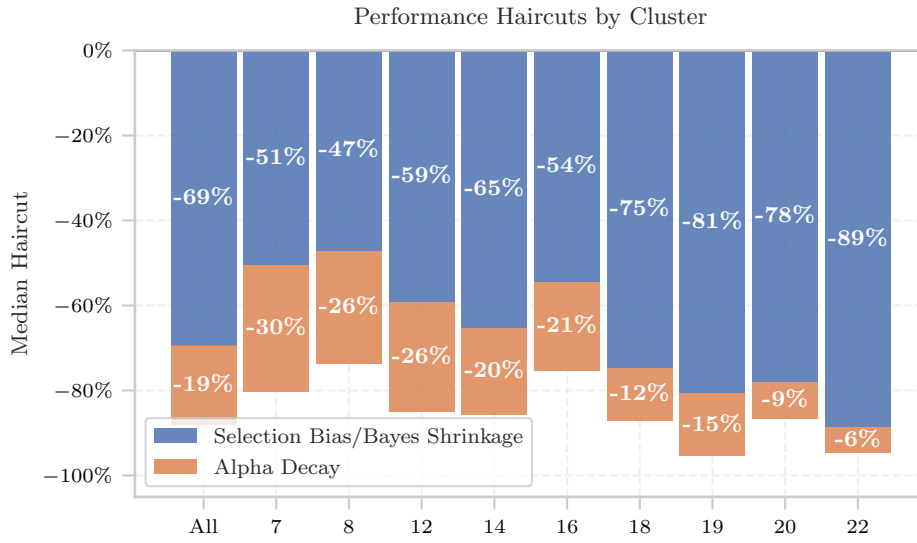


Figure 3.C.11: Median pre- to post-listing Sharpe ratio haircuts by cluster, decomposed into selection bias and alpha decay.

3.C.5 Comparator Models & Metrics

Point Prediction & Model Calibration Metrics

We start by defining $\text{SR}_i^{\text{pred}}$ as the prediction for the out-of-sample Sharpe ratio $\widehat{\text{SR}}_{\text{OOS},i}$ implied by the model. For the various Bayesian models, this is the alpha decayed posterior mean of the true Sharpe ratio,

$$\text{SR}_i^{\text{pred}} := \delta \mathbb{E} \left[\text{SR}_i \mid \widehat{\text{SR}}_{\text{IS},i} \right],$$

and for the James–Stein approach it is simply the James–Stein estimator $\widehat{\text{SR}}_{\text{IS},i}^{\text{JS}}$, as defined in Appendix 3.C.5. As the most naive estimator, we can simply take the in-sample Sharpe ratio directly $\widehat{\text{SR}}_{\text{IS},i}$.

The first measure of prediction quality is the mean squared error (MSE), which we compute for both our predictions $\text{SR}_i^{\text{pred}}$ and our naive estimate $\widehat{\text{SR}}_{\text{IS},i}$,

$$\text{MSE} := \frac{1}{N} \sum_{i=1}^N \left(\text{SR}_i^{\text{pred}} - \widehat{\text{SR}}_{\text{OOS},i} \right)^2, \quad \text{MSE}_{\text{base}} := \frac{1}{N} \sum_{i=1}^N \left(\widehat{\text{SR}}_{\text{IS},i} - \widehat{\text{SR}}_{\text{OOS},i} \right)^2,$$

with lower values indicating that the prediction better captures the central tendency of out-of-sample Sharpe ratios. To aid interpretation, we report the percentage improvement in MSE relative to the baseline,

$$\text{Improvement} := 1 - \frac{\text{MSE}}{\text{MSE}_{\text{base}}}.$$

But this disregards one of the main benefits of Bayesian approaches: the quantification of uncertainty. Specifically, a model which is confident but wrong should be penalised more than a model which correctly identifies its own uncertainty. To assess this we follow [Gel+13, Ch. 7] and use the log predictive density to measure probabilistic predictive accuracy.

For each fund we compute the predictive density of the out-of-sample Sharpe ratio at its observed value $\widehat{\text{SR}}_{\text{OOS},i}$. This is the posterior predictive density, obtained by marginalising the out-of-sample likelihood over the posterior distribution of the fund’s true Sharpe ratio:

$$f\left(\widehat{\text{SR}}_{\text{OOS},i} \mid \widehat{\text{SR}}_{\text{IS},i}\right) = \int_{\text{SR}_i} f\left(\widehat{\text{SR}}_{\text{OOS},i} \mid \text{SR}_i\right) f\left(\text{SR}_i \mid \widehat{\text{SR}}_{\text{IS},i}\right) d\text{SR}_i,$$

from which we form the log pointwise predictive density (LPPD) contribution

$$\text{LPPD}_i := \log f\left(\widehat{\text{SR}}_{\text{OOS},i} \mid \widehat{\text{SR}}_{\text{IS},i}\right),$$

and thus compute the mean LPPD (MLPPD)

$$\text{MLPPD} := \frac{1}{N} \sum_{i=1}^N \text{LPPD}_i.$$

Higher values of MLPPD indicate that the predictive distribution assigns more probability mass to the realised post-listing Sharpe ratios.

To provide a probabilistic benchmark (in the same sense that we use SR_i as a naive benchmark for the MSE) we also construct a baseline predictive distribution by treating the in-sample Sharpe ratio as a point estimate for the true Sharpe ratio and use Lo's approximation for the sampling distribution of Sharpe ratios [Lo02] to yield

$$\widehat{\text{SR}}_{\text{OOS},i} \mid \widehat{\text{SR}}_{\text{IS},i} \sim \mathcal{N} \left(\widehat{\text{SR}}_{\text{IS},i}, \frac{1 + \frac{1}{2} \widehat{\text{SR}}_{\text{IS},i}^2}{T_{\text{OOS},i}} \right).$$

Let $\text{LPPD}_i^{\text{base}}$ denote the corresponding log pointwise predictive density contribution from this baseline, and let

$$\text{MLPPD}_{\text{base}} := \frac{1}{N} \sum_{i=1}^N \text{LPPD}_i^{\text{base}}.$$

When presenting results we report the difference

$$\Delta \text{MLPPD} := \text{MLPPD} - \text{MLPPD}_{\text{base}},$$

so that larger values of ΔMLPPD indicate an improvement over the naive baseline in terms of probabilistic predictive accuracy.

As a complementary diagnostic, we examine the calibration of the predictive distributions. For each fund we compute the probability integral transform (PIT)

$$u_i := F_{\widehat{\text{SR}}_{\text{OOS},i} \mid \widehat{\text{SR}}_{\text{IS},i}}(\widehat{\text{SR}}_{\text{OOS},i}),$$

where $F_{\widehat{\text{SR}}_{\text{OOS},i} \mid \widehat{\text{SR}}_{\text{IS},i}}$ is the predictive cumulative distribution function of $\widehat{\text{SR}}_{\text{OOS},i}$ conditional on $\widehat{\text{SR}}_{\text{IS},i}$. If the model is correctly specified and well calibrated, then

the PIT values $\{u_i\}$ should behave approximately as independent draws from a $\text{Uniform}(0, 1)$ distribution [Gel+13, Ch. 6].

To assess deviations from uniformity we compute the Kolmogorov–Smirnov statistic

$$\text{KS} := \sup_{u \in [0,1]} \left| \widehat{F}(u) - u \right|,$$

where $\widehat{F}$ is the empirical distribution function of the PIT values. From this we compute the associated p -value under the null hypothesis of uniformity. Smaller values of KS and larger p -values are consistent with adequate model calibration.

For comparison we compute the same calibration diagnostics for the naive baseline predictive distribution. Let

$$u_{i,\text{base}} := F_{\widehat{\text{SR}}_{\text{OOS},i} | \widehat{\text{SR}}_{\text{IS},i}}^{\text{base}}(\widehat{\text{SR}}_{\text{OOS},i})$$

denote the PIT values obtained from the baseline predictive CDF, and let KS_{base} be the associated Kolmogorov–Smirnov statistic. In our summary plots we report the difference

$$\Delta\text{KS} := \text{KS}_{\text{base}} - \text{KS},$$

so that larger values of ΔKS correspond to a reduction in the KS distance relative to the naive baseline (i.e. better calibration).

Comparator Models

We apply the metrics above to a set of comparator models that are designed to disentangle the roles of shrinkage, selection bias and alpha decay.

James–Stein shrinkage As a robust benchmark we use the James–Stein estimator [JS60; EM75; Jor86], which performs cross-sectional shrinkage without explicitly

modelling selection bias or decay. The estimator is given by

$$\widehat{\text{SR}}_{\text{IS},i}^{\text{JS}} := \overline{\text{SR}} + \eta \left(\widehat{\text{SR}}_{\text{IS},i} - \overline{\text{SR}} \right),$$

where $\overline{\text{SR}}$ denotes the grand mean of the in-sample Sharpe ratios $\frac{1}{N} \sum_{i=1}^N \widehat{\text{SR}}_{\text{IS},i}$, and we use positive-part shrinkage [Bar70],

$$\eta = \max \left(0, 1 - \frac{(N-3)\bar{\sigma}^2}{\sum_{j=1}^N (\widehat{\text{SR}}_{\text{IS},j} - \overline{\text{SR}})^2} \right),$$

where $\bar{\sigma}^2$ is the average estimation variance of the in-sample Sharpe ratios, calculated as the arithmetic mean of the individual variances derived using Lo's approximation [Lo02]:

$$\bar{\sigma}^2 := \frac{1}{N} \sum_{i=1}^N \frac{1 + \frac{1}{2} \widehat{\text{SR}}_{\text{IS},i}^2}{T_{\text{IS},i}}.$$

Here, each term $\frac{1 + \frac{1}{2} \widehat{\text{SR}}_{\text{IS},i}^2}{T_{\text{IS},i}}$ approximates $\mathbb{V}(\widehat{\text{SR}}_{\text{IS},i} \mid \text{SR}_i)$ under iid Gaussian returns, with $\widehat{\text{SR}}_{\text{IS},i}$ used as a plug-in estimate for SR_i .

For the predictive density used to compute MLPPD and PIT/KS metrics, we take $\widehat{\text{SR}}_{\text{IS},i}^{\text{JS}}$ as a point estimator for SR_i and use the standard sample distribution for Sharpe ratios, again using Lo's approximation,

$$\widehat{\text{SR}}_{\text{OOS},i} \mid \widehat{\text{SR}}_{\text{IS},i}^{\text{JS}} \sim \mathcal{N} \left(\widehat{\text{SR}}_{\text{IS},i}^{\text{JS}}, \frac{1 + \frac{1}{2} \left(\widehat{\text{SR}}_{\text{IS},i}^{\text{JS}} \right)^2}{T_{\text{OOS},i}} \right).$$

Unbiased Bayesian Models The first set of Bayesian comparators do not take selection bias into consideration, so that pre- versus post-listing differences are driven entirely by shrinkage and alpha decay. To fit these models we simply proceed by

MLE as in Section 3.2.3, but using the unbiased marginal likelihood

$$f(\widehat{\text{SR}}_{\text{IS},i}, \widehat{\text{SR}}_{\text{OOS},i}) = \int_{\text{SR}}' f(\widehat{\text{SR}}_{\text{IS},i} \mid \text{SR}') f(\widehat{\text{SR}}_{\text{OOS},i} \mid \delta \text{SR}') \pi(\text{SR}') d\text{SR}',$$

rather than the bias-aware marginal likelihood we use for the standard model.

To further isolate the importance of alpha decay we fit two models, one where we fix $\delta = 1$ in the optimiser, and a second model where we allow δ to be optimised along with the other parameters.

Bias aware Bayesian models. These are the full models introduced in the main text, which explicitly account for selection bias and alpha decay. We consider three variations:

- *Random IS only.* The model is fit using only the in-sample (pre-listing) Sharpe ratios. This primarily functions to show that model estimation is difficult using the in-sample only data; as we see in the main text, this model collapses to a poor fit.
- *Random.* The baseline Bayesian model using both pre- and post-listing Sharpe ratios for the fitting process.
- *Random with clusters.* As a minor illustration of why fitting with similar strategies (in terms of performance) is a useful endeavour, we fit a model to each of the clusters created in Appendix 3.C.4. We will observe that this can provide minor improvements in model calibration.

3.C.6 Walkforward Analysis

We have also conducted a walkforward style analysis which avoids the lookahead bias in the fitting of the model parameters Θ . To do this, we walkforward a year at a time and use only the returns up to that year-end to fit the model, i.e. compute the pre- and post-listing Sharpe ratios for all funds listed before the end of that year. We then use these fitted parameters to compute the posterior and predictive

distributions for all the funds listed the following year, using only their pre-listing returns to estimate the posterior (as we did above). In Figure 3.C.18 we plot the evolution of the fitted parameters over time.

The issue with this walkforward approach lies in that there is a step change difference in the performance of funds pre- and post 2008. This both violates our assumptions of stationarity, and yields estimation issues whereby the parameters fit using, say, data from 1996 to 2009 are “stale” relative to the prevailing regime. To resolve this, we start our analysis in either 1996 or 2009 (Jan 31st), and end it in either 2008 or 2020 (December 31st), truncating the returns for each fund to this period. This yields three separate walkforward samples, corresponding to 1996 to 2008, 1996 to 2020 (our entire sample) and 2009 to 2020.

As above, we compare the estimated posterior and predictive distribution using the pre-listing Sharpe ratio to the post-listing Sharpe ratio for each fund, and compute the MSE, MLPPD and K-S statistic improvements relative to the naive baseline of the raw pre-listing Sharpe ratio. The outcome of this analysis is shown in Figure 3.C.12. We have excluded the by cluster fit in the interest of computation time, but otherwise include the same models as in the above analysis.

From an MSE perspective, the James–Stein estimator remains unchanged, as expected. However, the unbiased model without alpha decay and the bias adjusted model fitted to in-sample-only data suffer significantly. Crucially, the unbiased model with alpha decay continues to perform well, and the full bias adjusted model dominates across metrics and walkforward periods.

3.C.7 Additional Figures & Tables



Figure 3.C.12: Improvement in statistical metrics relative to a naive baseline when fitted with a walkforward approach.

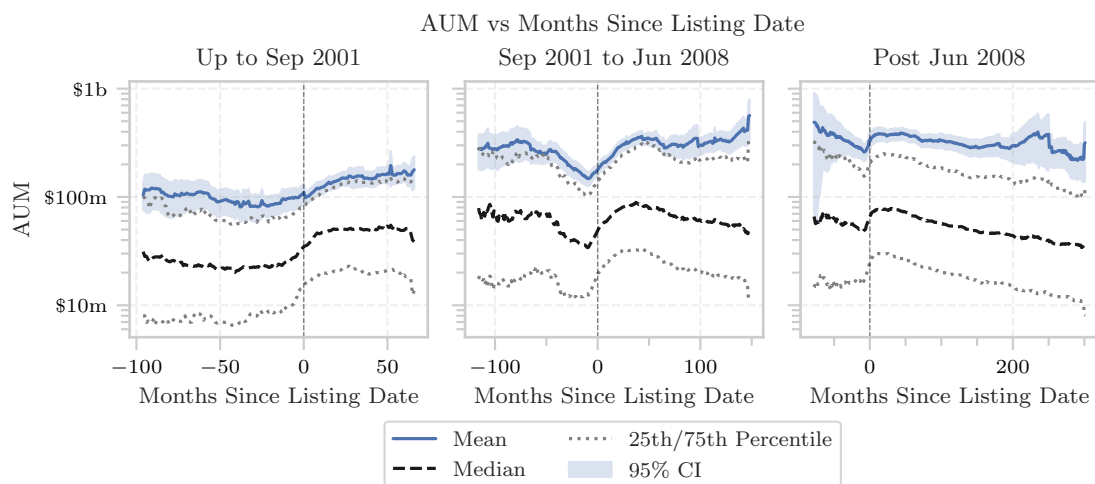


Figure 3.C.13: Average AUM in USD vs months pre- or post-listing in commercial databases.

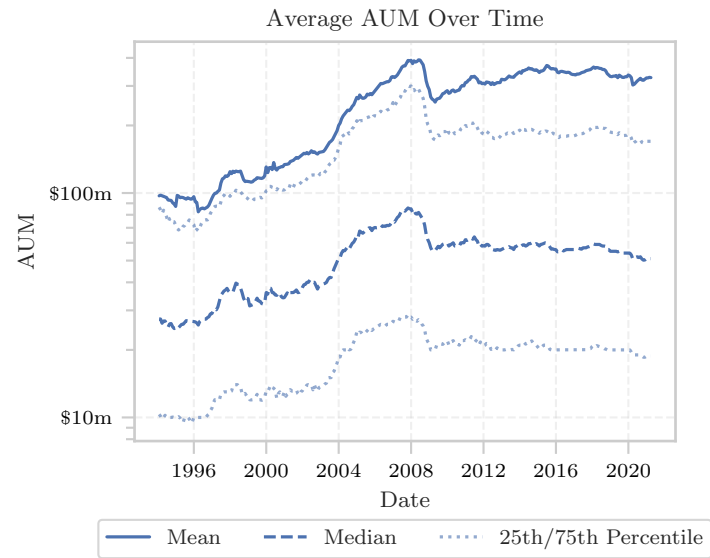


Figure 3.C.14: Average AUM in USD over time.

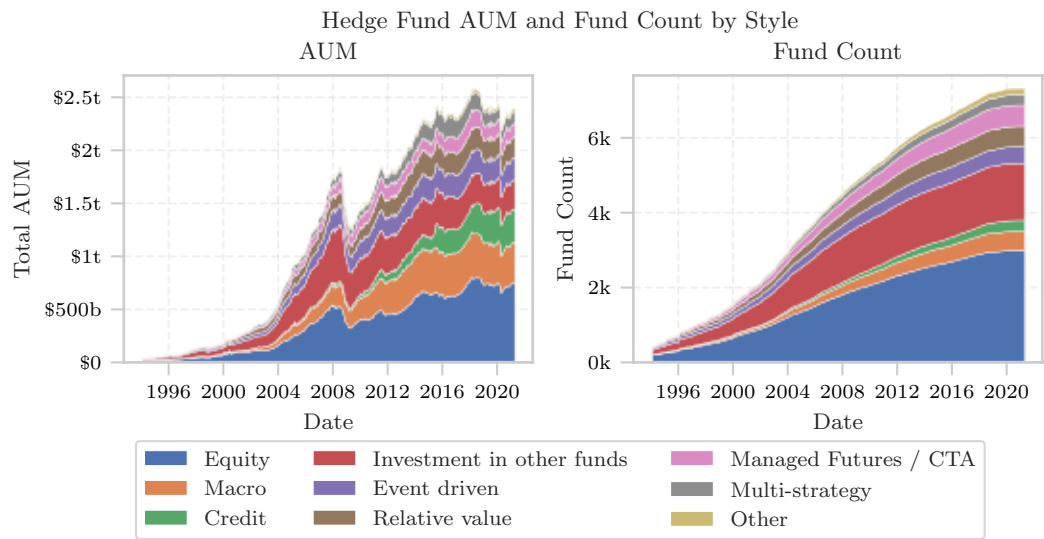


Figure 3.C.15: Total AUM in USD, and total fund count, by style over time.

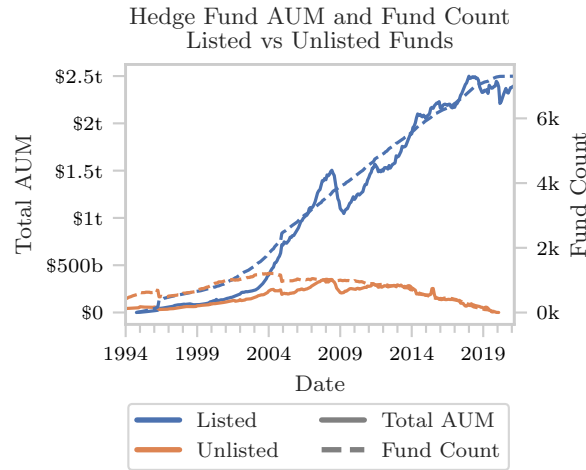


Figure 3.C.16: Total AUM in USD, and total fund count, by listing status over time.

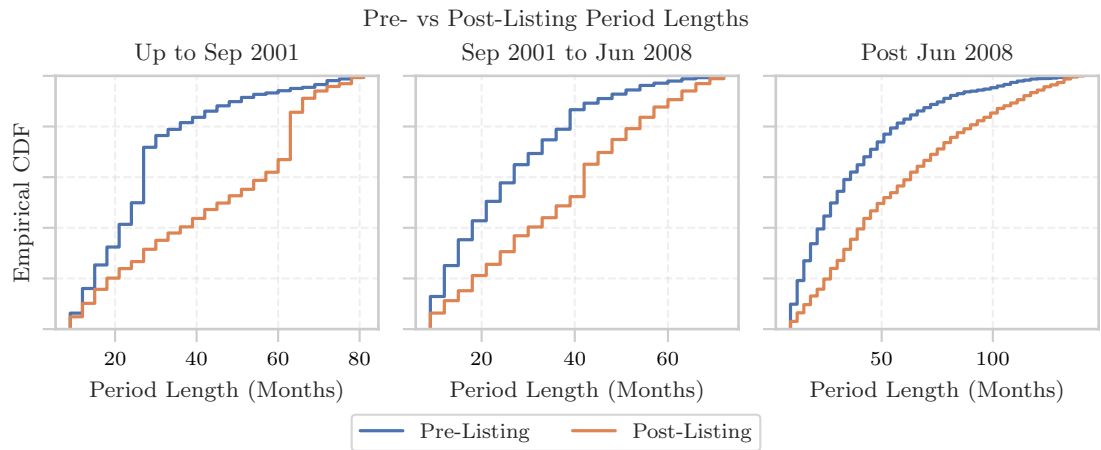


Figure 3.C.17: Cumulative distribution of period lengths.

3.C Supplementary Material for Section 3.3

Model / fit type	Param	Up to Sep 2001	Sep 2001 to July 2008	Post July 2008
Unbiased $\delta = 1$ (MLE)	SR_0	0.36 (1.26)	0.25 (0.87)	0.08 (0.27)
	ς_0	0.38 (1.32)	0.30 (1.03)	0.23 (0.78)
	ν	2.59	2.56	3.00
	κ	—	—	—
	ℓ	—	—	—
	δ	1.00	1.00	1.00
Bias Adjusted IS only (MLE)	SR_0	0.24 (0.84)	0.20 (0.69)	0.02 (0.08)
	ς_0	0.29 (1.01)	0.30 (1.03)	0.22 (0.77)
	ν	3.59	3.26	4.11
	κ	2.55 (8.85)	3.89 (13.48)	3.33 (11.52)
	ℓ	1.69	0.85	1.70
	δ	1.00	1.00	1.00
Unbiased (MLE)	SR_0	0.42 (1.47)	0.29 (1.01)	0.12 (0.41)
	ς_0	0.43 (1.50)	0.33 (1.15)	0.32 (1.12)
	ν	2.58	2.64	2.78
	κ	—	—	—
	ℓ	—	—	—
	δ	0.77	0.77	0.54
Bias Adjusted (MLE)	SR_0	0.26 (0.89)	0.24 (0.83)	-0.01 (-0.03)
	ς_0	0.25 (0.87)	0.27 (0.95)	0.22 (0.75)
	ν	4.36	3.04	3.95
	κ	9.65 (33.44)	10.00 (34.64)	1.69 (5.87)
	ℓ	1.83	0.60	2.13
	δ	0.91	0.83	0.68
Bias Adjusted (MLE-GMM κ, δ)	SR_0	0.30 (1.05)	0.27 (0.93)	-0.01 (-0.05)
	ς_0	0.30 (1.04)	1.79 (6.22)	0.21 (0.73)
	ν	3.26	2.01	5.21
	κ	2.33 (8.07)	-0.07 (-0.25)	1.19 (4.12)
	ℓ	1.24	0.86	2.50
	δ	0.87	0.68	0.55

Table 3.C.4: Fitted parameters by model and period. For SR_0 , ς_0 and κ , annualised equivalents are shown in parentheses. A dash (—) indicates that the parameter is not identified in the corresponding model.

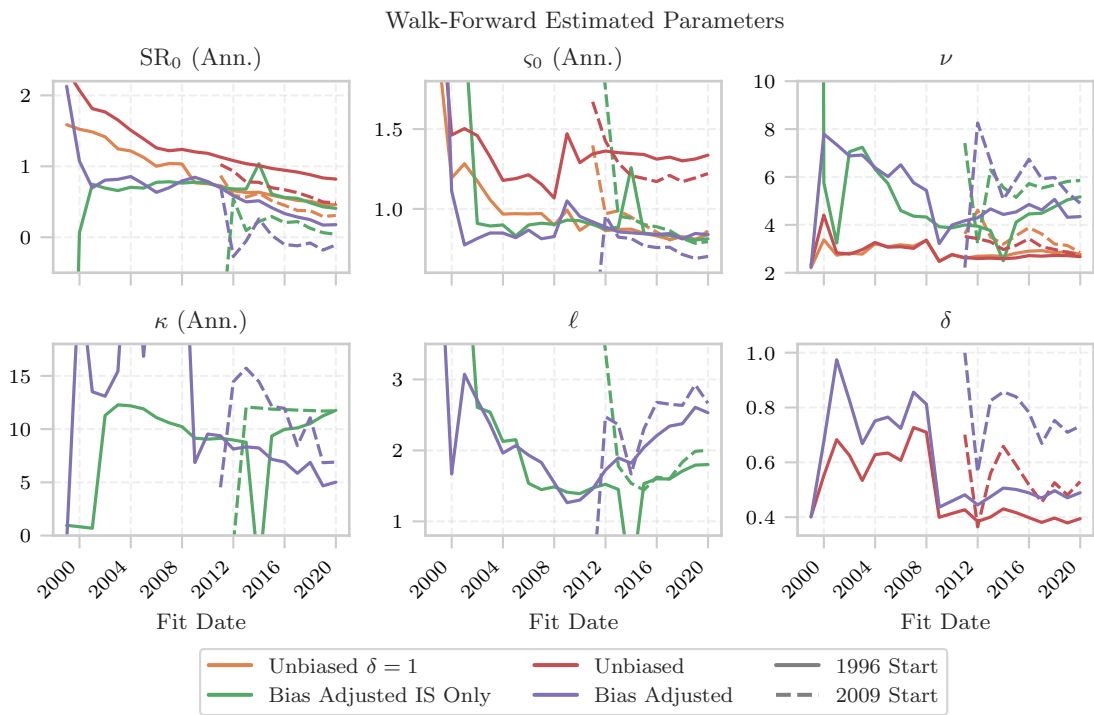


Figure 3.C.18: Fitted parameters for the walkforward analysis. Solid lines show the fitted parameters for the 1996 start date, and dashed lines for the 2009 start date. The y-axis is truncated for each plot for clarity.

3.D Illustrative Examples

The following examples are designed to build intuition for how selection bias distorts the in-sample Sharpe ratio, and how different parameters drive this effect. The results are intended to be illustrative, rather than prescriptive. The model parameters for this section are based on our empirical analysis and on what is plausible in practice.

The prior we use here is slightly pessimistic, as we assume the distribution of true Sharpe ratios is centred around zero. In many scenarios this would likely be shifted to have a more positive midpoint, but the pessimistic prior is useful to demonstrate how severe the issue of selection bias can be. The goal, as ever, is for researchers to apply this framework to their own data to yield insights specific to their strategies, but this section’s aim is to provide valuable intuition into how the selection process biases backtests.

Unless otherwise stated, we use the parameters in table 3.D.1 for all the examples in this section, and always use a Gaussian IID likelihood. We assume returns are measured daily with 252 trading days per year, and the parameters reflect this. For visualisations and discussions, however, we will work in annualised scales, as this is the scale most familiar to readers. For example, SR, θ or κ shown as 1, represents a daily value of $1/\sqrt{252}$. A useful first step is visualising what the prior and likelihoods look like for these parameter values.

Prior	Selection	Sample
$SR_0 = 0$	$\kappa = 1/\sqrt{252}$	$T_{IS} = 252 \cdot 3$
$\varsigma_0^2 = 0.5/252$	$\ell = 100$	$T_{OOS} = 252 \cdot 3$
$\nu = 10$		
$\nu_\theta = 10$		
$\omega = 0.9$		

Table 3.D.1: Prior, selection and sample parameters for the experiments of Appendix 3.D.

Figure 3.D.1 shows the density of the prior for the true Sharpe ratio in the unbiased, hierarchical and random cases. For the hierarchical and random cases, the figure shows the bias adjusted prior. We see that the unbiased prior is centred around SR_0 and the fixed prior is identically distributed, while the hierarchical and random priors have been shifted upwards by the selection process. As we allocate the majority of the variance to the fixed (θ) level of the prior, the hierarchical prior remains close to the unbiased one. As $\omega \rightarrow 0$, it would converge towards the random case as the total variance becomes fully allocated to the random SR layer.

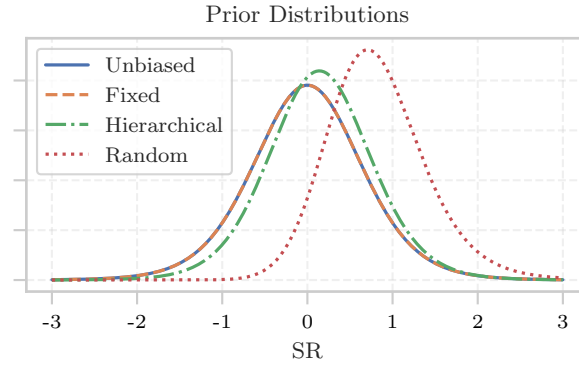


Figure 3.D.1: Prior distribution of true Sharpe ratios under unbiased, fixed, hierarchical and random selection.

Looking at the hierarchical prior a touch closer, fig. 3.D.2 plots the density of the hyperprior on θ and the conditional prior for SR given $\theta = 1$. These densities are not bias adjusted, as the intention is simply to show how much the true Sharpe ratio can vary once θ is fixed. A relatively small amount of change is possible; three standard deviations of the conditional prior correspond to about 0.66 in annualised Sharpe. This represents a realistic scenario where a researcher finds a core strategy and then tweaks it to yield incremental improvements.

Figure 3.D.3 shows the marginal likelihoods for the in- and out-of-sample Sharpe ratios under the four cases: unbiased, fixed, random, and hierarchical. In-sample, the likelihoods for the fixed, hierarchical, and random models are shifted substantially upwards. The random model yields the highest Sharpe ratios; because ω is close to

3.D Illustrative Examples

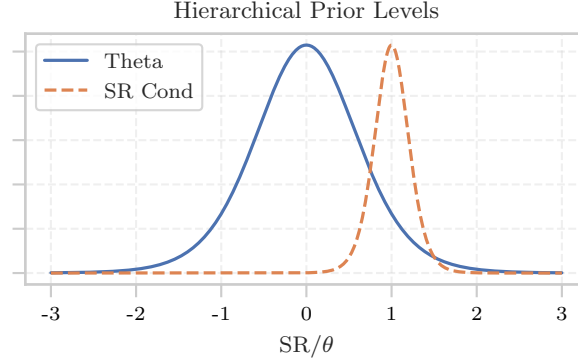


Figure 3.D.2: Densities for the hyperprior and conditional prior of the hierarchical case. Note these are not bias adjusted.

1, the distributions for the fixed and hierarchical models are nearly identical. Out-of-sample, however, the likelihoods for the unbiased, fixed, and hierarchical models are centred around zero, reflecting the prior. In contrast, the random model's likelihood is skewed positive, as the selection process influences the true Sharpe ratio upwards.

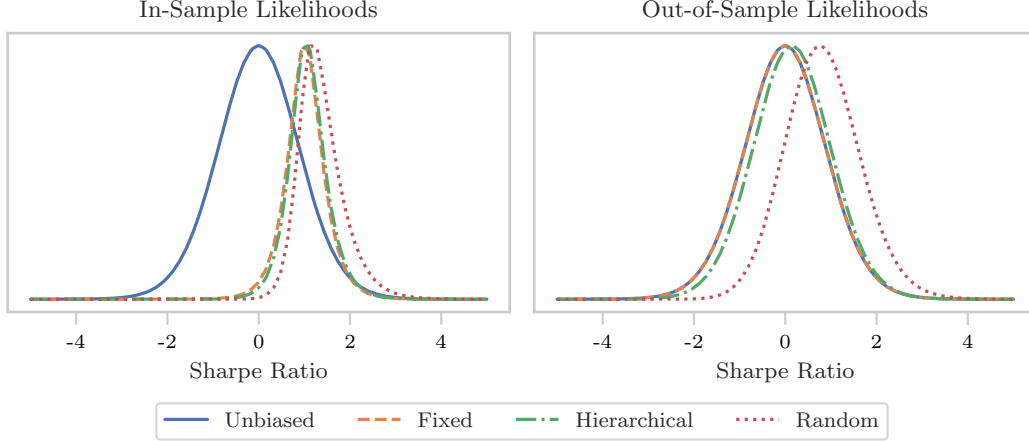


Figure 3.D.3: Densities for the in- and out-of-sample Sharpe ratios in the unbiased, fixed, hierarchical and random cases.

To demonstrate how the different selection bias types affect the posterior, we fix $\widehat{SR}_{IS} = 1.5$ and $T_{IS} = 252 \cdot 3$. The resulting posteriors are shown in Figure 3.D.4. First, Bayesian regularisation shrinks all posteriors towards the prior. The evidence (backtest length) is insufficient to support a true Sharpe ratio as high as the in-

3.D Illustrative Examples

sample estimate, so the posterior is naturally more cautious. Second, the fixed and hierarchical posteriors are even lower, while the random posterior is identical to the unbiased posterior. The intuition is that for the random case, the selection process shifts both the prior and the likelihood upwards, and the posterior is thus unbiased. In the fixed and hierarchical cases however, the selection process has little or no influence on the true Sharpe ratio, forcing a downward adjustment to the posterior.

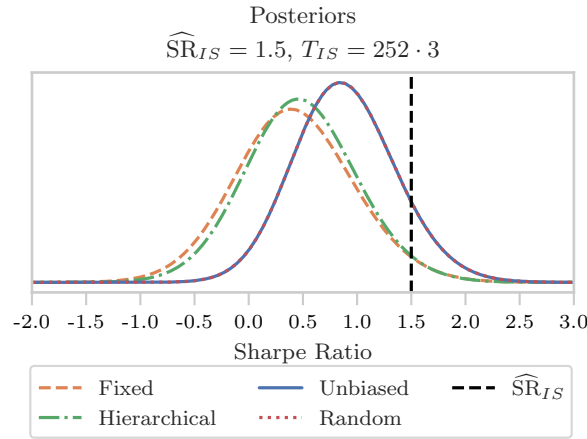


Figure 3.D.4: Posteriors for the fixed, hierarchical, random and unbiased cases; $\widehat{SR}_{IS} = 1.5, T_{IS} = 252 \cdot 3$.

How does this change for varying $\widehat{SR}_{IS}$ and T_{IS} ? Figure 3.D.5 plots the replication ratio $\mathbb{E}[SR \mid \widehat{SR}_{IS}] / \widehat{SR}_{IS}$ (the ratio of the posterior mean to the in-sample Sharpe ratio) for the unbiased/random cases, which are equal, and the fixed case. The hierarchical case is omitted because for $\omega = 0.9$ it is very similar to the fixed case. In the unbiased/random cases, the replication ratio rises slightly with the in-sample Sharpe ratio, but increases primarily with the length of the backtest T_{IS} . In the fixed (and hierarchical) case, the ratio depends strongly on the distance between the sample Sharpe ratio and the selection threshold. With $\kappa = 1$, an in-sample Sharpe ratio just above this level warrants scepticism; it is more likely to be a low true Sharpe ratio that had an unusually high in-sample realisation. As the in-sample Sharpe ratio moves further from the threshold, this possibility diminishes, and it becomes more likely that the true Sharpe ratio that yielded a large in-sample

3.D Illustrative Examples

Sharpe ratio is itself large, and therefore less distorted by selection bias. As in the unbiased/random case, longer backtests mitigate this effect by reducing the measurement error in the sample Sharpe ratio.

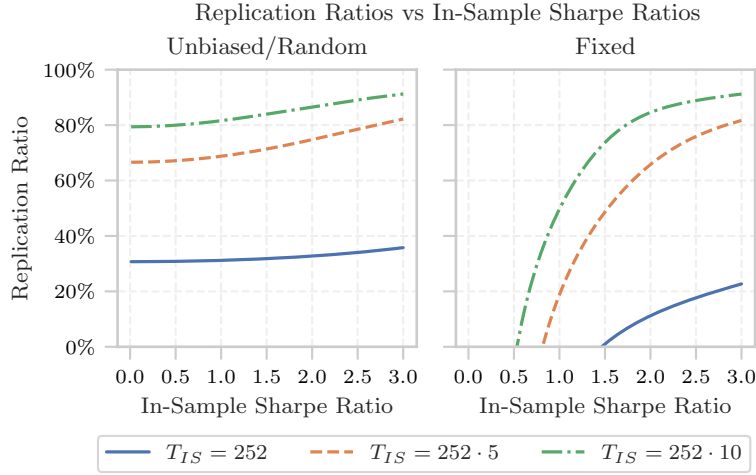


Figure 3.D.5: Replication ratios for unbiased/random selection bias vs fixed selection bias.

Next we examine how the variance split affects the replication ratio in the hierarchical case. Figure 3.D.6 shows that as the variance split ω approaches one, allocating more variance to the fixed (θ) level of the prior, the replication ratio decreases. The intuition is straightforward: if our selection process has less influence on the true Sharpe ratio, a high in-sample result is likely driven more by overfitting than by genuine improvement. We should therefore reduce our expectations for out-of-sample performance.

Rather than vary the observed in-sample Sharpe ratio, we can also examine how the replication ratio responds to changes in the prior and the selection bias parameters, as shown in Figure 3.D.7 and Figure 3.D.8 respectively. In each plot, we vary only one parameter while holding the others fixed to their values in Table 3.D.1.

In Figure 3.D.7, we plot the random and fixed cases as we vary the prior parameters. We omit the unbiased case, which is equivalent to the random case, and the hierarchical case, which is an interpolation of the two. In the fixed case, not only

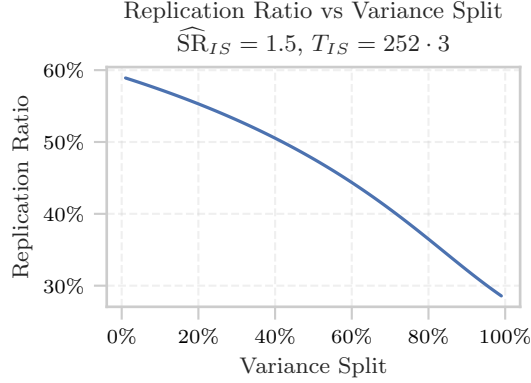


Figure 3.D.6: Replication ratio $\mathbb{E}[SR \mid \widehat{SR}_{IS}] / \widehat{SR}_{IS}$ by variance split ω .

is the replication ratio lower across the board, but its response to changing prior parameters is also different. As the prior mean SR_0 increases towards the in-sample Sharpe ratio ($\widehat{SR}_{IS} = 1.5$), the gap between the fixed and random cases closes. A higher prior mean for the true Sharpe ratio lessens the impact of selection bias, causing the posterior in the fixed case to converge towards the random one. Additionally, and naturally, as SR_0 increases we see that Bayesian shrinkage decreases, and once $SR_0 > \widehat{SR}_{IS}$ it pulls our estimate of the true Sharpe ratio above the in-sample value. Relatedly, but to a lesser degree, as the variance of the prior ς_0^2 increases we see that the replication ratio increases. This is caused by a higher likelihood of observing large true Sharpe ratios, and therefore our fairly conservative in-sample Sharpe ratio $\widehat{SR}_{IS} = 1.5$ becomes more probable. However, as the mean SR_0 is still less than κ , there is still a degree of selection bias, and thus the fixed replication ratio is much lower than the random one.

In Figure 3.D.8, we plot only the fixed case, as the replication ratio for the unbiased/random case is unaffected by the selection bias parameters. As in Figure 3.D.5, we observe that the replication ratio depends on the distance between the threshold κ and the in-sample Sharpe ratio $\widehat{SR}_{IS}$. As κ increases the replication ratio decreases, becoming negative for large κ . It is possible to have $\kappa > \widehat{SR}_{IS}$ because the selection threshold with $\ell = 100$ is “soft”. As this threshold gets sharper (i.e. as ℓ

3.D Illustrative Examples

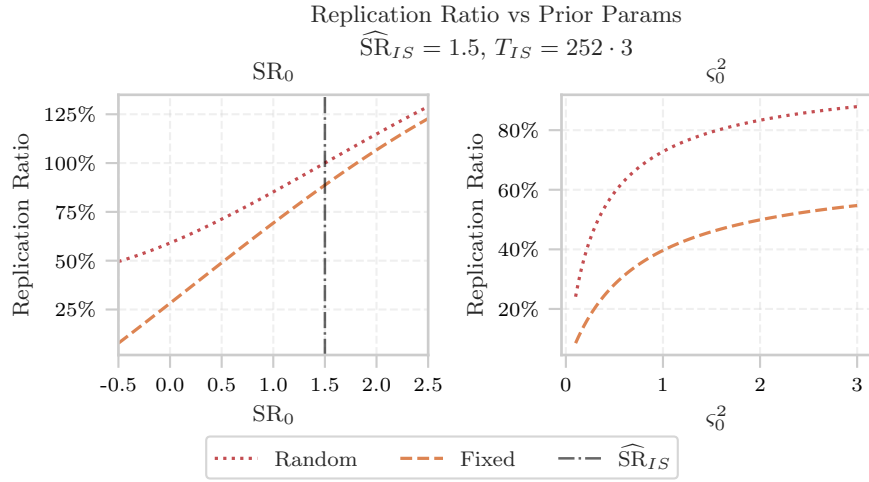


Figure 3.D.7: Replication ratio by SR_0 and ζ_0^2 .

increases), the replication ratio falls rapidly. This reflects the increasing strictness of the selection procedure. For large ℓ any strategy with $\widehat{SR}_{IS} < \kappa$ is discarded. Thus, the selection bias effect becomes stronger, and observed Sharpe ratios just above this threshold are more likely to be lucky than genuinely significant.

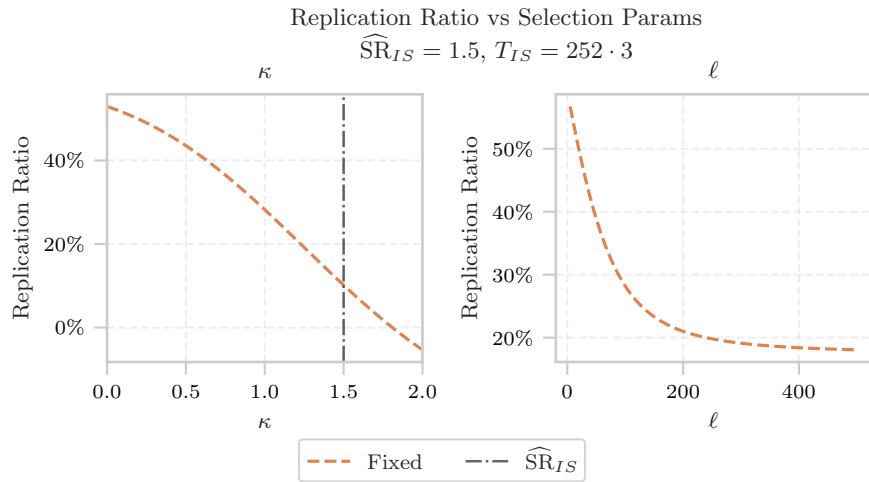


Figure 3.D.8: Replication ratio by κ and ℓ .

3.E Comparison to Chen and Zimmermann

Our approach is related to the study by Chen and Zimmermann [CZ20] (CZ), who develop a model-implied estimate of publication bias in the asset pricing literature. Here, we explore the theoretical differences between our models and assess how these differences affect fitting and inference when applied to the predictor data they kindly provide.

The primary distinction is one of objective. CZ take the view of an economist seeking to measure the magnitude of publication bias using only in-sample data. We take the view of an investor aiming to produce the most accurate forecast of an investment’s future performance. This goal motivates our use of all available data, both biased in-sample and unbiased out-of-sample returns, and informs the key methodological differences that follow.

First, our model explicitly incorporates alpha decay. This allows us to fit in-sample and out-of-sample data jointly, disentangling the distinct effects of selection bias and performance degradation over time. CZ acknowledge alpha decay but treat it as a residual between their bias adjusted estimates and the observed post-publication return. They choose to use in-sample data alone, arguing that out-of-sample periods are clouded by economic effects that would corrupt a pure measure of publication bias. By modelling alpha decay as a parameter, we can account for these effects directly. This joint-fitting also yields improved parameter estimates, a significant benefit for smaller datasets like the one here, with roughly 200 predictors (Appendix 3.B.1).

Second, we focus on different performance metrics. As our work is oriented toward capital allocation, we concentrate on the Sharpe ratio. Accordingly, our likelihood function is derived from its asymptotic distribution, which accommodates the non-Gaussian and autocorrelated nature of financial returns. CZ, in line with the academic literature they study, focus on mean returns and t-statistics, employing

a Normal distribution for the likelihood. While the metrics are related, our choice allows for a more direct and robust analysis of risk-adjusted performance.

Finally, our model addresses a broader set of selection mechanisms, which we explore in Appendix 3.A.2. CZ focus exclusively on random selection, a specific case that is well-suited for their inferential goal but is one of several scenarios our framework can handle.

With the methodological differences established, we can turn to studying their impact on the open-source asset pricing (OSA) dataset, which Chen and Zimmermann kindly maintain [CZ22]. The dataset contains reproduced monthly returns for 212 predictors from the asset pricing literature, of which we select 203 after dropping factors which have time series that end prior to 2023.

The OSA dataset includes returns for a variety of portfolio construction methods. We select the kind used by CZ, the “Original Cuts” dataset, which splits the assets by quantile and equally weights the assets within each quantile. The authors also calculate a long-short portfolio return by simply going long the top (or bottom) quantile, and short the bottom (or top) quantile. The dataset also includes the start and end year used by the original paper for each predictor, and the year that paper was published. Using these three dates we can define a few key periods:

- The In-Sample Period: The period used by the original paper for evaluating the predictor,
- The Pre-Publication Period: The period between the end of the in-sample period and the publication date,
- The Post-Publication Period: The period from the publication date till now,
- The Post-Sample Period: The pre-publication and post-publication period combined.

Chen and Zimmermann note that for some factors their reproduced t -statistics are lower than the original paper, for example the predictors by Haugen and Baker [HB96].

We found that this could be remedied by taking a Markowitz combination of the top and bottom quantiles. To construct the Markowitz portfolio we compute the mean and covariance of the top and bottom quantiles using the in-sample period, and use the calculated weights to compute the Markowitz Long-Short return for all the periods.

Figure 3.E.1 shows the cumulative returns for the cash flow yield variance (VarCF) predictor using CZ’s original long-short portfolio, and our Markowitz constructed long-short portfolio, normalised such that the monthly volatility of both returns time series is 1. We see that the Markowitz constructed portfolio performs significantly better than the original, both in-sample and post-sample. The in-sample t -statistic increases from 1.9 to 3.7, and the post-sample t -statistic from 0.3 to 3.1.

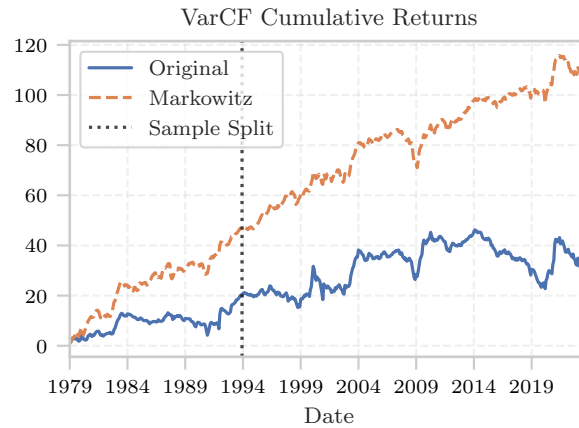


Figure 3.E.1: Cumulative long-short returns for the cash flow yield variance predictor [HB96] using the original or Markowitz construction.

In addition to the predictors CZ found problematic (those with lower than expected in-sample t -statistics), we find that the Markowitz construction improves t -statistics for all the predictors, both in- and post-sample. Figure 3.E.2 shows these distributions for the original and Markowitz portfolios. As expected, the in-sample t -statistics are significantly higher for the Markowitz portfolios, since they are built using in-sample moments, but the distribution of post-sample t -statistics improves too. The mean in-sample t -statistic increases from 4.0 to 6.2, and the

mean post-sample t -statistic increases from 1.7 to 3.4.

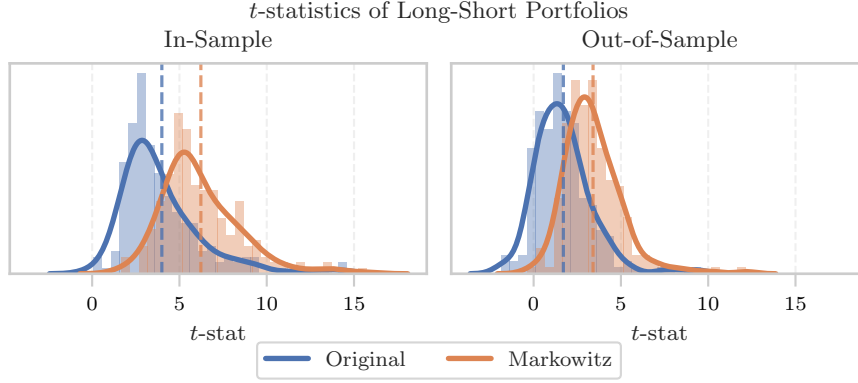


Figure 3.E.2: In- and post-sample long-short t -statistics for original and Markowitz portfolios, dashed lines denote means.

Thus, for our comparison we will use the Markowitz constructed portfolios, rather than the original portfolios, as we feel they more accurately capture the true performance of each predictor. Additionally, as our framework is grounded in Sharpe ratios, rather than t -statistics, we will conduct our experiments in terms of Sharpe ratios, but these could be converted back into t -statistics at the end by a simple rescaling according to sample length. This does of course mean any comparison to particular values from [CZ20] will not be an exact apples-to-apples comparison, but we aim to capture the spirit of their model versus our own.

We fit six models. The first three use in-sample data alone, in-sample with pre-publication data, or in-sample with post-sample data. The remaining three are variants where we fix $SR_0 = 0$, mirroring the approach of [CZ20]. Each model is fitted using a non-Gaussian AR(1) likelihood with moments estimated from the data: autocorrelation $\phi = 0.11$, skewness $\gamma_{3,r} = 0.77$, and kurtosis $\gamma_{4,r} = 15.52$. The parameters are estimated by taking the average across all the predictors, and by using the full time-series available for each predictor.

The fitted parameters are reported in Table 3.E.1. For the three models where SR_0 is fitted, its value is stable, varying by only 0.03 (0.1 annualised) around the

in-sample estimate of 0.27 (0.92 annualised). Notably, the alpha decay parameter is much lower for the in- and post-sample fit than for the in-sample and pre-publication fit, which aligns with the expectation that the pre-publication period is more insulated from alpha decay. Across all six fits, the selection function remains nearly identical, with $\kappa \approx 0.2$ (0.7 annualised) and $\ell \approx 45$.

The primary difference between fixing SR_0 and fitting it lies in the prior's standard deviation ς_0 and kurtosis ν parameters. When SR_0 is fitted, the prior is narrower ($\varsigma_0 \approx 0.15$) and has fatter tails ($\nu \approx 5$). In contrast, fixing $SR_0 = 0$ results in a wider prior ($\varsigma_0 = 0.24$) with little excess kurtosis, due to a large estimated ν . We believe this is because fixing the mean at zero requires a wider prior to capture the bulk of the in-sample distribution; the model then fits a larger ν to reduce kurtosis, preventing the tails from becoming excessively fat.

	IS Only		IS & Pre		IS & OOS	
	Fit SR_0	$SR_0 = 0$	Fit SR_0	$SR_0 = 0$	Fit SR_0	$SR_0 = 0$
SR_0	0.27	0.00	0.24	0.00	0.30	0.00
ν	5.67	100	4.37	25.83	3.24	100
ς_0	0.15	0.24	0.17	0.24	0.14	0.24
δ	1.00	1.00	0.93	0.94	0.61	0.62
κ	0.19	0.22	0.20	0.22	0.18	0.22
ℓ	45.43	45.10	42.93	44.54	46.64	44.68

Table 3.E.1: Fitted parameter values for Appendix 3.E. Note that the parameter values are in monthly terms.

To illustrate the difference between the model fits, we visualise their likelihoods and priors. We select two models for comparison: the in-sample only model with $SR_0 = 0$, representing CZ's approach, and our model fitted to both in-sample and pre-publication data. We focus for now on the pre-publication Sharpe ratios, rather than the post-sample Sharpe ratios, as they offer a closer comparison to the issue CZ studied.

Figure 3.E.3 shows that the in-sample likelihoods are nearly identical, perhaps

highlighting the difficulty of fitting these models on small, in-sample datasets alone. For the pre-publication Sharpe ratios, however, the difference between the fits is more apparent. Fitting to the in-sample and pre-publication data jointly yields a closer fit for these pre-publication ratios.

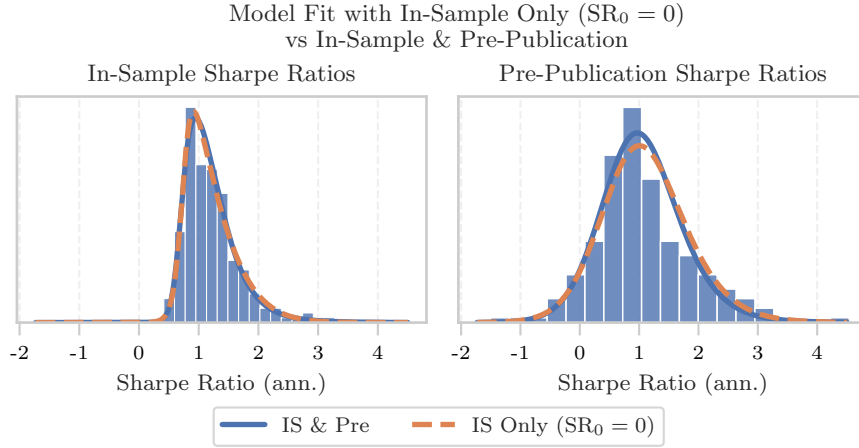


Figure 3.E.3: Model fit (marginal likelihoods) for in-sample only model (with $SR_0 = 0$) versus in- and pre-publication model.

Figure 3.E.4 plots the raw prior, $\pi(SR)$, and the bias adjusted prior, $\pi_B(SR)$, for both models (see Appendix 3.A.2). Since the in-sample only model fixes $SR_0 = 0$, its unbiased prior is centred at zero and is relatively wide to accommodate the observed in-sample Sharpe ratios. The prior for our jointly fitted model, in comparison, is considerably more optimistic. The bias adjusted priors for both models are, like the likelihoods, very similar, though the jointly fitted prior is slightly more peaked.

This result may seem slightly odd. It is natural that the prior for the Sharpe ratios conditional on acceptance has some reasonable positive expectation ($\mathbb{E}[\pi_B(SR)] = 1.14$ in annualised terms for the in- and pre-publication fitted model). But why is the expectation of the unconditional prior ($\mathbb{E}[\pi(SR)] = 0.84$) quite high? The nuance lies in that this is not the completely unconditional prior of any possible strategy, but rather it is only unconditional of the modelled selection effect. More explicitly, it is still conditional on the initial selection that the researcher implicitly performs

from their own economic priors on what they believe could work, known stylised facts, portfolio construction, etc. Naturally this leads to the researcher filtering out many obviously poor strategies, resulting in an overall positive distribution of possible strategies. We also note that the average out-of-sample Sharpe ratio, which we note is subject to not insignificant alpha decay, is 0.62. With this context, a gross Sharpe ratio of around 0.84 seems quite reasonable.

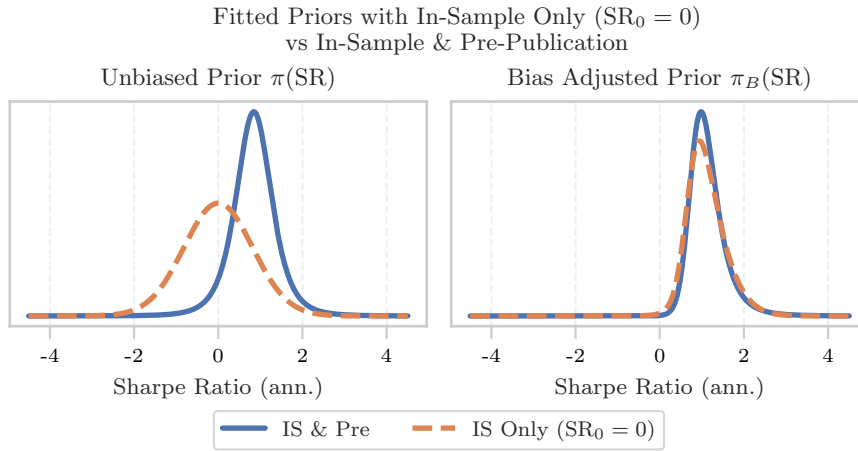


Figure 3.E.4: Unbiased and bias adjusted priors for in-sample only model (with $SR_0 = 0$) versus in- and pre-publication model.

How do these differences affect inference? Figure 3.E.5 plots the percentage change from the in-sample Sharpe ratio $\widehat{SR}_{IS}$ to the posterior mean $\mathbb{E}[SR \mid \widehat{SR}_{IS}]$: $\mathbb{E}[SR \mid \widehat{SR}_{IS,i}] / \widehat{SR}_{IS,i} - 1$. For the model where SR_0 is fixed at zero, all posterior means are shrunk towards the prior mean, resulting in adjustments of -2% to -34%. In the jointly fitted model, by contrast, the more optimistic prior leads to most posteriors being shrunk, but a few (those with low in-sample Sharpe ratios) are increased, with values ranging from +30% to -32%. This makes clear that the calculated adjustments strongly depend on the fitted prior, and the estimated adjustment per predictor is a combination of selection bias inferred from the fitting procedure, and standard Bayesian shrinkage.

The two models yield average adjustments, $\frac{1}{N} \sum_{i=1}^N \mathbb{E}[SR \mid \widehat{SR}_{IS,i}] / \widehat{SR}_{IS,i} - 1$, of -

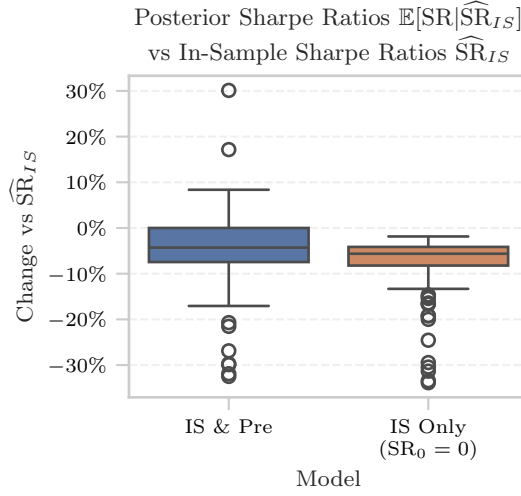


Figure 3.E.5: Percent change from the in-sample Sharpe ratio to the posterior mean $\mathbb{E}[\text{SR} \mid \widehat{\text{SR}}_{\text{IS},i}] / \widehat{\text{SR}}_{\text{IS},i} - 1$ for in-sample only model (with $\text{SR}_0 = 0$) versus in- and post-sample model.

7% and -4% respectively. For comparison, [CZ20, fig. A.5] reports a bias adjustment of -10% for their sample; but as noted, differences in the models and data treatment make this an inexact comparison.¹¹

Given the noted importance of the prior mean for the estimated posterior means, and therefore also for estimates of the magnitude of publication bias, we believe it is better to allow the data to fully inform the prior, rather than fix parameters at arbitrary values. But with a sample size of just 200 values to fit the parameters, any particular calculated posteriors and adjustments should be taken with a pinch of salt. Some comfort can be taken following our findings in Appendix 3.B.1 that jointly fitting the data improves parameter estimates, and this is further underscored by the similar values for SR_0 across all three unrestricted models.

Finally, we study the impact of alpha decay using two visualisations. The first, in Figure 3.E.6, plots the distributions of in-sample Sharpe ratios $\widehat{\text{SR}}_{\text{IS}}$, posterior means $\mathbb{E}[\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}}]$, decayed posterior means $\mathbb{E}[\text{SR}_\delta \mid \widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}}]$, and

¹¹ We note that these adjustments are much smaller than those we observed in the hedge fund dataset; however, the purpose of this section is to compare our approach against CZ, rather than hedge funds against anomalies.

the out-of-sample Sharpe ratios $\widehat{SR}_{OOS}$. The out-of-sample data is taken to be either the pre-publication or post-sample data. To obtain the best possible estimate for the true Sharpe ratio SR , we fit the model and compute the posterior using the combined in- and out-of-sample data. The second visualisation, in Figure 3.E.7, shows a waterfall breakdown of the average adjustment into components: selection bias ($\Delta_{SB} = \mathbb{E}[SR \mid \widehat{SR}_{IS}, \widehat{SR}_{OOS}] / \widehat{SR}_{IS} - 1$), alpha decay ($\Delta_{AD} = \mathbb{E}[SR_\delta \mid \widehat{SR}_{IS}, \widehat{SR}_{OOS}] / \widehat{SR}_{IS} - 1 - \Delta_{SB}$) and noise ($\Delta_\epsilon = \widehat{SR}_{OOS} / \widehat{SR}_{IS} - 1 - \Delta_{SB} - \Delta_{AD}$).

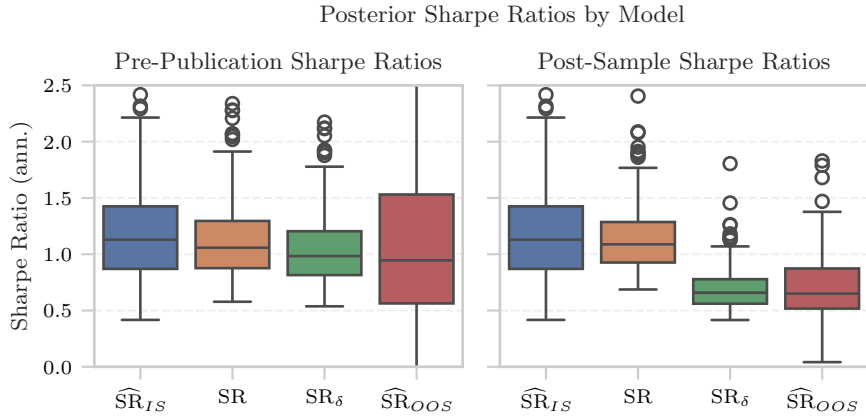


Figure 3.E.6: Distributions of in-sample Sharpe ratios $\widehat{SR}_{IS}$, posterior means $\mathbb{E}[SR \mid \widehat{SR}_{IS}, \widehat{SR}_{OOS}]$, decayed posterior means $\mathbb{E}[SR_\delta \mid \widehat{SR}_{IS}, \widehat{SR}_{OOS}]$, and the out-of-sample Sharpe ratios $\widehat{SR}_{OOS}$, for the pre-publication and post-sample models.

For both out-of-sample periods, the alpha-decayed Sharpe ratio $\mathbb{E}[SR_\delta \mid \widehat{SR}_{IS}, \widehat{SR}_{OOS}]$ better captures the location of the out-of-sample distribution $\widehat{SR}_{OOS}$ than the raw Sharpe ratio $\mathbb{E}[SR \mid \widehat{SR}_{IS}, \widehat{SR}_{OOS}]$. As expected, the distribution of posterior estimates is also much narrower than that of the out-of-sample Sharpe ratios. For the pre-publication period, the average adjustment decomposes into approximately 5% from selection bias (or Bayesian shrinkage) and 7% from alpha decay; noise adds back 1%, yielding a final out-of-sample performance of 90% of the in-sample figure. For the post-sample data, we attribute only 1% to selection bias but 39% to alpha decay, with noise adding back 2% for a final performance of 62% of the in-sample result.

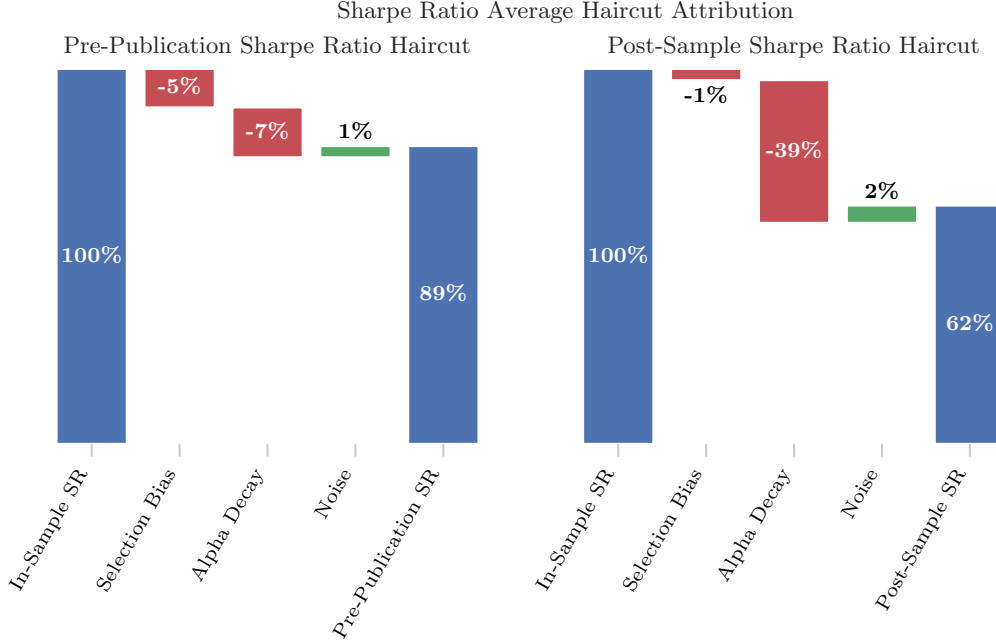


Figure 3.E.7: Attribution of average adjustments into selection bias ($\Delta_{SB} = \mathbb{E}[\text{SR} \mid \widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}}] / \widehat{\text{SR}}_{\text{IS}} - 1$), alpha decay ($\Delta_{AD} = \mathbb{E}[\text{SR}_\delta \mid \widehat{\text{SR}}_{\text{IS}}, \widehat{\text{SR}}_{\text{OOS}}] / \widehat{\text{SR}}_{\text{IS}} - 1 - \Delta_{SB}$) and noise ($\Delta_\epsilon = \widehat{\text{SR}}_{\text{OOS}} / \widehat{\text{SR}}_{\text{IS}} - 1 - \Delta_{SB} - \Delta_{AD}$) for the pre-publication and post-sample models.

These figures however are highly conditional on the estimated parameters Θ . Fixing the prior's location SR_0 at a lower value, as CZ do for instance, would attribute more of the haircut to selection bias and less to alpha decay. A dataset with thousands of predictors might yield more definitive parameter estimates, but with only 200 samples, the haircut attributions presented here should be interpreted with caution.

3.F Proofs

3.F.1 Proof of Proposition 3.A.1: Asymptotic Variance for the General Likelihood

The broad structure of the proof is similar to that in [LLZ25] and to the delta-method argument of Mertens [Mer02]. The only point that requires extra care is the asymptotic normality of the estimator $\widehat{\boldsymbol{\theta}}$; we justify this explicitly using [PS92, Thm. 3.4 and Thm. 3.8] before computing the entries of the long-run covariance matrix. Let $g(\boldsymbol{\theta}) = \mu/\sqrt{\sigma^2}$, where $\boldsymbol{\theta} = [\mu, \sigma^2]^\top$ and $\widehat{\boldsymbol{\theta}}$ is the sample estimator.

Since $y_t := r_t - \mu = \sum_{k \geq 0} \phi^k \epsilon_{t-k}$ is a causal linear process with absolutely summable coefficients ($|\phi| < 1$) and $\mathbb{E}[|\epsilon_t|^4] < \infty$, we apply [PS92, Thm. 3.4] to the sample mean $\bar{y}$ and [PS92, Thm. 3.8] to the sample mean of $y_t^2 - \sigma^2$, and use Cramér-Wold to yield the CLT for the vector $\boldsymbol{\theta}$,

$$\sqrt{T_{\text{IS}}} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{S}),$$

where $\mathbf{S}$ is the long-run covariance matrix of $H_t(\boldsymbol{\theta}) = \left[y_t, y_t^2 - \sigma^2 \right]^\top$. Since $g(\boldsymbol{\theta}) = \mu/\sqrt{\sigma^2}$ is C^1 on $\{\sigma^2 > 0\}$, we can therefore use this CLT and the delta method to yield

$$\sqrt{T_{\text{IS}}} \left(g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}) \right) \xRightarrow{d} \mathcal{N} \left(0, \nabla g(\boldsymbol{\theta})^\top \mathbf{S} \nabla g(\boldsymbol{\theta}) \right),$$

where the gradient is given by

$$\nabla g(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma} \\ -\frac{\mu}{2\sigma^3} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma} \\ -\frac{\text{SR}}{2\sigma^2} \end{bmatrix}.$$

Thus, the asymptotic variance V_{SR} is given by

$$\begin{aligned} V_{\text{SR}} &= \nabla g(\boldsymbol{\theta})^\top \cdot \mathbf{S} \cdot \nabla g(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma} & -\frac{\text{SR}}{2\sigma^2} \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} \\ -\frac{\text{SR}}{2\sigma^2} \end{bmatrix} \\ &= \frac{1}{\sigma^2} \mathbf{S}_{11} - \frac{\text{SR}}{\sigma^3} \mathbf{S}_{12} + \frac{\text{SR}^2}{4\sigma^4} \mathbf{S}_{22}, \end{aligned}$$

since $\mathbf{S}$ is symmetric. All that remains is to derive the entries of $\mathbf{S}$. Since $y_t = r_t - \mu = \phi y_{t-1} + \epsilon_t$,

$$\mathbf{S} = \begin{pmatrix} \sum_{\tau=-\infty}^{\infty} \text{Cov}(y_t, y_{t-\tau}) & \sum_{\tau=-\infty}^{\infty} \text{Cov}(y_t, y_{t-\tau}^2) \\ \sum_{\tau=-\infty}^{\infty} \text{Cov}(y_t^2, y_{t-\tau}) & \sum_{\tau=-\infty}^{\infty} \text{Cov}(y_t^2, y_{t-\tau}^2) \end{pmatrix}.$$

Starting with $\mathbf{S}_{11}$, the long-run variance of y_t , letting $\sigma^2 := \mathbb{V}[y_t] = \frac{\sigma_\epsilon^2}{1-\phi^2}$,

$$\begin{aligned} \mathbf{S}_{11} &= \sum_{\tau=-\infty}^{\infty} \text{Cov}(y_t, y_{t-\tau}) = \text{Cov}(y_t, y_t) + 2 \sum_{\tau=1}^{\infty} \text{Cov}(y_t, y_{t-\tau}) \\ &= \sigma^2 + 2 \sum_{\tau=1}^{\infty} \phi^\tau \sigma^2 \\ &= \sigma^2 \frac{1+\phi}{1-\phi}, \end{aligned}$$

where we used the symmetry of the autocovariance for a stationary process, the fact that $\text{Cov}(y_t, y_{t-\tau}) = \phi^\tau \sigma^2$ and the formula for the sum of an infinite geometric series.

Next we turn to $\mathbf{S}_{12}$ (and by symmetry $\mathbf{S}_{21}$), the long-run covariance between the asset's return and its volatility,

$$\mathbf{S}_{12} = \sum_{\tau=-\infty}^{\infty} \text{Cov}(y_t, y_{t-\tau}^2) = \sum_{\tau=-\infty}^{\infty} \mathbb{E}[y_t y_{t-\tau}^2].$$

To proceed, consider the three cases for τ . Starting with $\tau = 0$,

$$\mathbb{E}[y_t y_t^2] = \mathbb{E}[y_t^3] = \mathbb{E} \left[\left(\sum_{k=0}^{\infty} \phi^k \epsilon_{t-k} \right)^3 \right] = \sum_{k=0}^{\infty} \phi^{3k} \mathbb{E}[\epsilon_{t-k}^3] = \frac{\gamma_{3,\epsilon} \sigma_\epsilon^3}{1-\phi^3},$$

where we have used the $\text{MA}(\infty)$ representation for the $\text{AR}(1)$ process $y_t = \sum_{k=0}^{\infty} \phi^k \epsilon_{t-k}$, and the fact that the sequence $(\epsilon_t)_t$ is iid and $\mathbb{E}[\epsilon_t] = 0$, so all mixed terms vanish.

Next for $\tau > 0$,

$$\mathbb{E}[y_t y_{t-\tau}^2] = \mathbb{E} \left[\left(\phi^\tau y_{t-\tau} + \sum_{k=0}^{\tau-1} \phi^k \epsilon_{t-k} \right) y_{t-\tau}^2 \right] = \phi^\tau \mathbb{E}[y_{t-\tau}^3] = \phi^\tau \mathbb{E}[y_t^3],$$

since the sequence $(\epsilon_t)_t$ is iid and centred and using the stationarity of y_t . Therefore

$$\sum_{\tau=1}^{\infty} \mathbb{E}[y_t y_{t-\tau}^2] = \mathbb{E}[y_t^3] \sum_{\tau=1}^{\infty} \phi^\tau = \frac{\phi}{1-\phi} \mathbb{E}[y_t^3].$$

Finally, for $\tau < 0$, let $\tau = -k$ and establish the recursive relationship

$$\begin{aligned} \mathbb{E}[y_t y_{t+k}^2] &= \mathbb{E} \left[y_t (\phi y_{t+k-1} + \epsilon_{t+k})^2 \right] \\ &= \mathbb{E} \left[y_t (\phi^2 y_{t+k-1}^2 + 2\phi y_{t+k-1} \epsilon_{t+k} + \epsilon_{t+k}^2) \right] \\ &= \phi^2 \mathbb{E}[y_t y_{t+k-1}^2], \end{aligned}$$

where the last two terms are null by independence of ϵ_{t+k} and all variables at or before $t+k-1$. Applying this recurrence k times yields $\mathbb{E}[y_t y_{t+k}^2] = \phi^{2k} \mathbb{E}[y_t^3]$ and the sum over all lags reads

$$\sum_{k=1}^{\infty} \mathbb{E}[y_t y_{t+k}^2] = \frac{\phi^2}{1-\phi^2} \mathbb{E}[y_t^3].$$

Putting the three cases together we have

$$\begin{aligned} \mathbf{S}_{12} &= \mathbb{E}[y_t y_t^2] + \sum_{\tau=1}^{\infty} \mathbb{E}[y_t y_{t-\tau}^2] + \sum_{k=1}^{\infty} \mathbb{E}[y_t y_{t+k}^2] \\ &= \mathbb{E}[y_t^3] \left(1 + \frac{\phi}{1-\phi} + \frac{\phi^2}{1-\phi^2} \right) \\ &= \frac{\gamma_{3,\epsilon} \sigma_\epsilon^3}{(1-\phi^2)(1-\phi)} = \gamma_{3,\epsilon} \sigma^3 \sqrt{\frac{1+\phi}{1-\phi}}. \end{aligned}$$

Finally, we compute $\mathbf{S}_{22}$, the autocovariance of the squared process y_t^2 ,

$$\begin{aligned}\mathbf{S}_{22} &= \sum_{\tau=-\infty}^{\infty} \text{Cov}(y_t^2, y_{t-\tau}^2) = \mathbb{V}[y_t^2] + 2 \sum_{\tau=1}^{\infty} \text{Cov}(y_t^2, y_{t-\tau}^2) \\ &= \mathbb{V}[y_t^2] + 2 \sum_{\tau=1}^{\infty} \phi^{2\tau} \mathbb{V}[y_t^2] = \mathbb{V}[y_t^2] \frac{1 + \phi^2}{1 - \phi^2},\end{aligned}$$

since the autocovariance of the squared process is $\text{Cov}(y_t^2, y_{t-\tau}^2) = \phi^{2\tau} \mathbb{V}[y_{t-\tau}^2] = \phi^{2\tau} \mathbb{V}[y_t^2]$ by stationarity. We simply need to compute the variance $\mathbb{V}[y_t^2] = \mathbb{E}[y_t^4] - \mathbb{E}[y_t^2]^2 = \mathbb{E}[y_t^4] - \sigma^4$. Calculating the fourth moment,

$$\begin{aligned}\mathbb{E}[y_t^4] &= \mathbb{E}[(\phi y_{t-1} + \epsilon_t)^4] \\ &= \phi^4 \mathbb{E}[y_{t-1}^4] + 6\phi^2 \mathbb{E}[y_{t-1}^2] \mathbb{E}[\epsilon_t^2] + \mathbb{E}[\epsilon_t^4] \\ &= \phi^4 \mathbb{E}[y_t^4] + 6\phi^2 \sigma^2 \sigma_\epsilon^2 + \gamma_{4,\epsilon} \sigma_\epsilon^4 \\ &= \frac{6\phi^2 \sigma^2 \sigma_\epsilon^2 + \gamma_{4,\epsilon} \sigma_\epsilon^4}{1 - \phi^4},\end{aligned}$$

and therefore

$$\mathbb{V}[y_t^2] = \frac{6\phi^2 \sigma^4 (1 - \phi^2) + \gamma_{4,\epsilon} \sigma^4 (1 - \phi^2)^2}{1 - \phi^4} - \sigma^4 = \frac{\sigma^4}{1 + \phi^2} ((\gamma_{4,\epsilon} - 1) + (5 - \gamma_{4,\epsilon})\phi^2).$$

So, finally,

$$\mathbf{S}_{22} = \mathbb{V}[y_t^2] \frac{1 + \phi^2}{1 - \phi^2} = \frac{\sigma^4}{1 - \phi^2} ((\gamma_{4,\epsilon} - 1) + (5 - \gamma_{4,\epsilon})\phi^2).$$

And by substituting the values for $\mathbf{S}$ into the expression for the asymptotic variance we obtain

$$V_{\text{SR}} = \frac{1}{\sigma^2} \left(\sigma^2 \frac{1 + \phi}{1 - \phi} \right) - \frac{\text{SR}}{\sigma^3} \left(\gamma_{3,\epsilon} \sigma^3 \sqrt{\frac{1 + \phi}{1 - \phi}} \right) + \frac{\text{SR}^2}{4\sigma^4} \left(\frac{\sigma^4}{1 - \phi^2} [(\gamma_{4,\epsilon} - 1) + (5 - \gamma_{4,\epsilon})\phi^2] \right).$$

Simplifying this expression yields the desired result:

$$V_{\text{SR}} = \frac{1 + \phi}{1 - \phi} - \gamma_{3,\epsilon} \text{SR} \sqrt{\frac{1 + \phi}{1 - \phi}} + \frac{\text{SR}^2}{4(1 - \phi^2)} ((\gamma_{4,\epsilon} - 1) + (5 - \gamma_{4,\epsilon})\phi^2).$$

While the expression for V_{SR} is correct, in practice the moments of the returns r_t

are more practical to use. Therefore, let us translate our result in terms of these moments (rather than the moments of the innovations).

For the skewness, using the linear representation $y_t = \sum_{k=0}^{\infty} \phi^k \epsilon_{t-k}$, the third central moment is $\mathbb{E}[y_t^3] = \sum_{k=0}^{\infty} \phi^{3k} \mathbb{E}[\epsilon_{t-k}^3] = \frac{\gamma_{3,\epsilon} \sigma_\epsilon^3}{1-\phi^3}$. Since the return volatility is $\sigma = \sigma_\epsilon / \sqrt{1-\phi^2}$, the return skewness is given by

$$\gamma_{3,r} = \frac{\mathbb{E}[y_t^3]}{\sigma^3} = \frac{\gamma_{3,\epsilon} \sigma_\epsilon^3}{1-\phi^3} \frac{(1-\phi^2)^{3/2}}{\sigma_\epsilon^3},$$

so with a bit of algebra we have

$$\gamma_{3,\epsilon} = \gamma_{3,r} \frac{1-\phi^3}{(1-\phi^2)^{3/2}}.$$

Substituting this relation into the second term of V_{SR} yields:

$$\begin{aligned} \gamma_{3,\epsilon} \sqrt{\frac{1+\phi}{1-\phi}} &= \gamma_{3,r} \frac{(1-\phi)(1+\phi+\phi^2)}{(1-\phi^2)\sqrt{1-\phi^2}} \frac{\sqrt{1+\phi}}{\sqrt{1-\phi}} \\ &= \gamma_{3,r} \frac{1+\phi+\phi^2}{1-\phi^2}. \end{aligned}$$

For the kurtosis, we can use the previously derived expression for $\mathbb{E}[y_t^4]$:

$$\mathbb{E}[y_t^4] = \frac{\sigma_\epsilon^4}{1-\phi^4} \left(\gamma_{4,\epsilon} + \frac{6\phi^2}{1-\phi^2} \right).$$

Normalizing by $\sigma^4 = \sigma_\epsilon^4 / (1-\phi^2)^2$, we obtain the return kurtosis:

$$\begin{aligned} \gamma_{4,r} &= \frac{\mathbb{E}[y_t^4]}{\sigma^4} = \frac{(1-\phi^2)^2}{1-\phi^4} \left(\gamma_{4,\epsilon} + \frac{6\phi^2}{1-\phi^2} \right) \\ &= \frac{1-\phi^2}{1+\phi^2} \gamma_{4,\epsilon} + \frac{6\phi^2}{1+\phi^2}. \end{aligned}$$

Solving for $\gamma_{4,\epsilon}$ yields

$$\gamma_{4,\epsilon} = \frac{\gamma_{4,r}(1+\phi^2) - 6\phi^2}{1-\phi^2}.$$

Substituting this into V_{SR} :

$$\frac{1}{4(1-\phi^2)} [(\gamma_{4,\epsilon} - 1) + (5 - \gamma_{4,\epsilon})\phi^2] = \frac{1+\phi^2}{1-\phi^2} \frac{\gamma_{4,r} - 1}{4}.$$

Combining these yields the final asymptotic variance in terms of observed return statistics:

$$V_{\text{SR}} = \frac{1+\phi}{1-\phi} - \frac{1+\phi+\phi^2}{1-\phi^2} \gamma_{3,r} \text{SR} + \frac{1+\phi^2}{1-\phi^2} \frac{\gamma_{4,r} - 1}{4} \text{SR}^2.$$

4 | Conclusion

With all that said, what have we learned? What can we recommend to help both minimise and deal with the phenomenons observed in this thesis? Here are our recommendations:

- Use the simplest model you can. Financial data is notoriously noisy, and in most cases we have limited data on which to test our hypotheses.¹ The broad conclusions from both the financial literature and our own work is that it is very easy to overfit your model, so reducing the degrees of freedom is a prudent approach to minimise this risk. As the famous quip goes: “if you torture the data enough, nature will always confess”.²
- Use as much data as you can. This is challenging due to the constantly changing regimes in finance. The factor that reliably delivered positive alpha for decades can seemingly suddenly flip direction, and if you’re using decades of data, it can take far too long for you to react to these changes. This can be mitigated somewhat through various approaches such as: exponential weighting schemes [LS96], time-varying loadings [AF09], Bayesian changepoint detection [AM07], etc.
- If you can’t use a longer lookback, borrow strength from similar series. While each individual time series might need to be short due to changing regimes, different assets, or strategies, or funds, often behave similarly. Using pooling in either simple forms (panel regression) or more advanced forms (Bayesian hierarchical models) often results in much better estimates than naive approaches.

¹ In high-frequency settings this is typically less of an issue, with data intensive machine learning approaches working well [IL25].

² Incidentally, this is generally attributed to the economist Ronald Coase in a critique on empirical work in economics [Coa88]. This is not a new problem.

- Above all else, be sceptical. Building a backtest with a Sharpe ratio that looks fantastic is a very easy task. Building a strategy which will deliver that Sharpe ratio out-of-sample, taking into account all the subtleties of the real world is a very difficult task. A healthy dose of scepticism goes far in this regard.

Bibliography

- [AF09] Tobias Adrian and Francesco Franzoni. “Learning about Beta: Time-varying Factor Loadings, Expected Returns, and the Conditional CAPM”. *Journal of Empirical Finance* 16.4 (Sept. 2009), pp. 537–556. ISSN: 0927-5398. DOI: [10.1016/j.jempfin.2009.02.003](https://doi.org/10.1016/j.jempfin.2009.02.003).
- [AM07] Ryan Prescott Adams and David J. C. MacKay. “Bayesian Online Change-point Detection”. Oct. 2007. DOI: [10.48550/arXiv.0710.3742](https://doi.org/10.48550/arXiv.0710.3742). arXiv: [0710.3742](https://arxiv.org/abs/0710.3742) [stat].
- [Arn+19] Rob Arnott et al. “Alice’s Adventures in Factorland: Three Blunders That Plague Factor Investing”. *The Journal of Portfolio Management* 45.4 (Apr. 2019), pp. 18–36. ISSN: 0095-4918, 2168-8656. DOI: [10.3905/jpm.2019.45.4.018](https://doi.org/10.3905/jpm.2019.45.4.018).
- [AS58] J. Aitchison and S. D. Silvey. “Maximum-Likelihood Estimation of Parameters Subject to Restraints”. *The Annals of Mathematical Statistics* 29.3 (1958), pp. 813–828. ISSN: 0003-4851. JSTOR: [2237265](https://www.jstor.org/stable/2237265).
- [Bai+14] David H. Bailey et al. “Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance”. *Notices of the American Mathematical Society* 61.5 (May 2014), p. 458. ISSN: 0002-9920, 1088-9477. DOI: [10.1090/noti1105](https://doi.org/10.1090/noti1105).
- [Bar+23] Daniel Barth et al. “A Bias-Free Assessment of the Hedge Fund Industry: A New Evaluation of Total Assets, Alphas, and the Flow-Performance Relation”. SSRN Scholarly Paper. Rochester, NY, June 2023. DOI: [10.2139/ssrn.3544181](https://doi.org/10.2139/ssrn.3544181). Social Science Research Network: [3544181](https://ssrn.com/abstract=3544181).
- [Bar70] A. J. Baranchik. “A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution”. *The Annals of Mathematical Statistics* 41.2 (Apr. 1970), pp. 642–645. ISSN: 0003-4851, 2168-8990. DOI: [10.1214/aoms/1177697104](https://doi.org/10.1214/aoms/1177697104).
- [BBP16] David H. Bailey, Jonathan M. Borwein, and Marcos López de Prado. “The Probability of Backtest Overfitting”. *Journal of Computational Finance* (Sept. 2016).
- [BBP17] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. “Cleaning Large Correlation Matrices: Tools from Random Matrix Theory”. *Physics Reports*. Cleaning Large Correlation Matrices: Tools from Random Matrix Theory 666 (Jan. 2017), pp. 1–109. ISSN: 0370-1573. DOI: [10.1016/j.physrep.2016.10.005](https://doi.org/10.1016/j.physrep.2016.10.005).

- [Bec+16] Noah Beck et al. “Will Your Factor Deliver? An Examination of Factor Robustness and Implementation Costs”. *Financial Analysts Journal* 72.5 (Sept. 2016), pp. 58–82. ISSN: 0015-198X. DOI: [10.2469/faj.v72.n5.6](https://doi.org/10.2469/faj.v72.n5.6).
- [Ben+23] Itzhak Ben-David et al. “Competition for Attention in the ETF Space”. *The Review of Financial Studies* 36.3 (Mar. 2023), pp. 987–1042. ISSN: 0893-9454. DOI: [10.1093/rfs/hhac048](https://doi.org/10.1093/rfs/hhac048).
- [BGP99] Stephen J. Brown, William N. Goetzmann, and James Park. *Conditions for Survival: Changing Risk and the Performance of Hedge Fund Managers and CTAs*. Working Paper. June 1999.
- [BGR14] Geetesh Bhardwaj, Gary B. Gorton, and K. Geert Rouwenhorst. “Fooling Some of the People All of the Time: The Inefficient Performance and Persistence of Commodity Trading Advisors”. *The Review of Financial Studies* 27.11 (Nov. 2014), pp. 3099–3132. ISSN: 0893-9454. DOI: [10.1093/rfs/hhu040](https://doi.org/10.1093/rfs/hhu040).
- [BGS22] Laurent Barras, Patrick Gagliardini, and Olivier Scaillet. “Skill, Scale, and Value Creation in the Mutual Fund Industry”. *The Journal of Finance* 77.1 (2022), pp. 601–638. ISSN: 1540-6261. DOI: [10.1111/jofi.13096](https://doi.org/10.1111/jofi.13096).
- [BH95] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 0035-9246. JSTOR: [2346101](https://www.jstor.org/stable/2346101).
- [BJK21] Nicolas P.B. Bollen, Juha Joenväärä, and Mikko Kauppila. “Hedge Fund Performance: End of an Era?” *Financial Analysts Journal* 77.3 (July 2021), pp. 109–132. ISSN: 0015-198X, 1938-3312. DOI: [10.1080/0015198X.2021.1921564](https://doi.org/10.1080/0015198X.2021.1921564).
- [BKW04] David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity: 546*. 1st edition. Hoboken, N.J: Wiley-Interscience, July 2004. ISBN: 978-0-471-69117-4.
- [BL12] David H. Bailey and Marcos Lopez de Prado. “The Sharpe Ratio Efficient Frontier”. *Journal of Risk* 15.2 (Nov. 2012), pp. 3–44. DOI: [10.21314/JOR.2012.255](https://doi.org/10.21314/JOR.2012.255).
- [BL14] David H. Bailey and Marcos López De Prado. “The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality”. *The Journal of Portfolio Management* 40.5 (Sept. 2014), pp. 94–107. ISSN: 0095-4918, 2168-8656. DOI: [10.3905/jpm.2014.40.5.094](https://doi.org/10.3905/jpm.2014.40.5.094).
- [BL15] Chris Brightman and Feifei Li. *Chasing Performance with ETFs*. Research Affiliates, Nov. 2015.
- [BMW01] Klaas P. Baks, Andrew Metrick, and Jessica Wachter. “Should Investors Avoid All Actively Managed Mutual Funds? A Study in Bayesian Performance Evaluation”. *The Journal of Finance* 56.1 (2001), pp. 45–85. ISSN: 0022-1082. JSTOR: [222463](https://www.jstor.org/stable/222463).

- [Bon+19] Maxime Bonelli et al. “The Capacity of Trading Strategies”. SSRN Scholarly Paper. Rochester, NY, Feb. 2019. DOI: [10.2139/ssrn.2585399](https://doi.org/10.2139/ssrn.2585399).
- [Bon35] Carlo Emilio Bonferroni. “Il Calcolo Delle Assicurazioni Su Gruppi Di Teste”. *Studi in Onore del Professore Salvatore Ortu Carboni* (1935), pp. 13–60.
- [Bre15] Ernest Breitschwerdt. *Point-In-Time vs. Lagged Fundamentals*. 2015-08. S&P Capital IQ, Aug. 2015.
- [Bro20] Mike Brookes. “The Matrix Reference Manual”. <http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html>. 2020.
- [BY01] Yoav Benjamini and Daniel Yekutieli. “The Control of the False Discovery Rate in Multiple Testing under Dependency”. *The Annals of Statistics* 29.4 (Aug. 2001), pp. 1165–1188. ISSN: 0090-5364, 2168-8966. DOI: [10.1214/aos/1013699998](https://doi.org/10.1214/aos/1013699998).
- [Car+25] Álvaro Cartea et al. “Bottom-up Capacity Constraints and the Limits of Anomaly Profitability”. SSRN Scholarly Paper. Rochester, NY, Nov. 2025. DOI: [10.2139/ssrn.5797502](https://doi.org/10.2139/ssrn.5797502). Social Science Research Network: [5797502](https://ssrn.com/abstract=5797502).
- [CCG14] Stephen R. Cole, Haitao Chu, and Sander Greenland. “Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer”. *American Journal of Epidemiology* 179.2 (Jan. 2014), pp. 252–260. ISSN: 0002-9262. DOI: [10.1093/aje/kwt245](https://doi.org/10.1093/aje/kwt245).
- [CCZ17] Yong Chen, Michael Cliff, and Haibei Zhao. “Hedge Funds: The Good, the Bad, and the Lucky”. *Journal of Financial and Quantitative Analysis* 52.3 (June 2017), pp. 1081–1109. ISSN: 0022-1090, 1756-6916. DOI: [10.1017/S0022109017000217](https://doi.org/10.1017/S0022109017000217).
- [CH88] Samprit Chatterjee and Ali S. Hadi. *Sensitivity Analysis in Linear Regression: 190*. 1st edition. New York: Wiley, Apr. 1988. ISBN: 978-0-471-82216-5.
- [Che+25] Yong Chen et al. “Anomalies as New Hedge Fund Factors”. *Journal of Financial and Quantitative Analysis* (Oct. 2025), pp. 1–34. ISSN: 0022-1090, 1756-6916. DOI: [10.1017/S0022109025101270](https://doi.org/10.1017/S0022109025101270).
- [CJS25] Álvaro Cartea, Qi Jin, and Yuantao Shi. “The Limited Virtue of Complexity in a Noisy World”. SSRN Scholarly Paper. Rochester, NY, Apr. 2025. DOI: [10.2139/ssrn.5202064](https://doi.org/10.2139/ssrn.5202064). Social Science Research Network: [5202064](https://ssrn.com/abstract=5202064).
- [CL23] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge: Cambridge University Press, 2023.
- [Coa88] Ronald Coase. “How Should Economists Choose”. *Ideas, Their Origins and Their Consequences: Lectures to Commemorate the Life and Work of G. Warren Nutter* (Jan. 1988), p. 63.
- [Con01] R. Cont. “Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues”. *Quantitative Finance* 1.2 (Feb. 2001), pp. 223–236. ISSN: 1469-7688. DOI: [10.1080/713665670](https://doi.org/10.1080/713665670).
- [Cow33] Alfred Cowles. “Can Stock Market Forecasters Forecast?” *Econometrica* 1.3 (1933), pp. 309–324. ISSN: 0012-9682. DOI: [10.2307/1907042](https://doi.org/10.2307/1907042). JSTOR: [1907042](https://www.jstor.org/stable/1907042).

- [CV23a] Andrew Y. Chen and Mihail Velikov. “Zeroing In on the Expected Returns of Anomalies”. *Journal of Financial and Quantitative Analysis* 58.3 (May 2023), pp. 968–1004. ISSN: 0022-1090, 1756-6916. DOI: [10.1017/S0022109022000874](https://doi.org/10.1017/S0022109022000874).
- [CV23b] Dujava Cyril and Radovan Vojtko. “In-Sample vs. Out-Of-Sample Analysis of Trading Strategies”. SSRN Scholarly Paper. Rochester, NY, June 2023. DOI: [10.2139/ssrn.4486423](https://doi.org/10.2139/ssrn.4486423). Social Science Research Network: [4486423](https://ssrn.com/abstract/4486423).
- [CZ20] Andrew Y Chen and Tom Zimmermann. “Publication Bias and the Cross-Section of Stock Returns”. *The Review of Asset Pricing Studies* 10.2 (June 2020). Ed. by Jeffrey Pontiff, pp. 249–289. ISSN: 2045-9920, 2045-9939. DOI: [10.1093/rapstu/raz011](https://doi.org/10.1093/rapstu/raz011).
- [CZ22] Andrew Y. Chen and Tom Zimmermann. “Open Source Cross-Sectional Asset Pricing”. *Critical Finance Review* 11.2 (May 2022), pp. 207–264. ISSN: 2164-5744, 2164-5760. DOI: [10.1561/104.00000112](https://doi.org/10.1561/104.00000112).
- [Daw94] A. P. Dawid. “Selection Paradoxes of Bayesian Inference”. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*. Hayward, CA: Institute of Mathematical Statistics, 1994, pp. 211–220. ISBN: 978-0-940600-35-5. DOI: [10.1214/lnms/1215463797](https://doi.org/10.1214/lnms/1215463797).
- [Dem+25] Hans Dembinski et al. “Scikit-Hep/Iminuit”. Zenodo. Apr. 2025. DOI: [10.5281/zenodo.15157028](https://doi.org/10.5281/zenodo.15157028).
- [DL18] Hadrien De March and Charles-Albert Lehalle. “Optimal Trading Using Signals”. Nov. 2018. arXiv: [1811.03718](https://arxiv.org/abs/1811.03718) [q-fin].
- [EKM97] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events: For Insurance and Finance*. 1st ed. 1997, Corr. 10th printing 2012 edition. New York: Springer, June 1997. ISBN: 978-3-540-60931-5.
- [EM75] Bradley Efron and Carl Morris. “Data Analysis Using Stein’s Estimator and Its Generalizations”. *Journal of the American Statistical Association* 70.350 (1975), pp. 311–319. ISSN: 0162-1459. DOI: [10.2307/2285814](https://doi.org/10.2307/2285814). JSTOR: [2285814](https://www.jstor.org/stable/2285814).
- [FC60] Sir Ronald A. Fisher and E. A. Cornish. “The Percentile Points of Distributions Having Known Cumulants”. *Technometrics* 2.2 (May 1960), pp. 209–225. ISSN: 0040-1706. DOI: [10.1080/00401706.1960.10489895](https://doi.org/10.1080/00401706.1960.10489895).
- [FF93] Eugene F. Fama and Kenneth R. French. “Common Risk Factors in the Returns on Stocks and Bonds”. *Journal of Financial Economics* 33.1 (Feb. 1993), pp. 3–56. ISSN: 0304-405X. DOI: [10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5).
- [FH00] William Fung and David A. Hsieh. “Performance Characteristics of Hedge Funds and Commodity Funds: Natural vs. Spurious Biases”. *Journal of Financial and Quantitative Analysis* 35.3 (Sept. 2000), pp. 291–307. ISSN: 1756-6916, 0022-1090. DOI: [10.2307/2676205](https://doi.org/10.2307/2676205).
- [FH04] William Fung and David A. Hsieh. “Hedge Fund Benchmarks: A Risk-Based Approach”. *Financial Analysts Journal* 60.5 (2004), pp. 65–80. ISSN: 0015-198X. JSTOR: [4480604](https://www.jstor.org/stable/4480604).

- [FH09] William Fung and David A. Hsieh. “Measurement Biases in Hedge Fund Performance Data: An Update”. *Financial Analysts Journal* 65.3 (May 2009), pp. 36–38. ISSN: 0015-198X. DOI: [10.2469/faj.v65.n3.6](https://doi.org/10.2469/faj.v65.n3.6).
- [FIM12] Andrea Frazzini, Ronen Israel, and Tobias J. Moskowitz. “Trading Costs of Asset Pricing Anomalies”. SSRN Scholarly Paper. Rochester, NY, Dec. 2012. DOI: [10.2139/ssrn.2294498](https://doi.org/10.2139/ssrn.2294498).
- [FJL19] Roy Frostig, Matthew James Johnson, and Chris Leary. “Compiling Machine Learning Programs via High-Level Tracing”. *SysML Conference 2018*. Stanford, United States, Mar. 2019.
- [FP14] Andrea Frazzini and Lasse Heje Pedersen. “Betting against Beta”. *Journal of Financial Economics* 111.1 (Jan. 2014), pp. 1–25. ISSN: 0304-405X. DOI: [10.1016/j.jfineco.2013.10.005](https://doi.org/10.1016/j.jfineco.2013.10.005).
- [FRT22] Antoine Falck, Adam Rej, and David Thesmar. “When Do Systematic Strategies Decay?” *Quantitative Finance* 22.11 (Nov. 2022), pp. 1955–1969. ISSN: 1469-7688. DOI: [10.1080/14697688.2022.2098810](https://doi.org/10.1080/14697688.2022.2098810).
- [FS86] Peter A. Frost and James E. Savarino. “An Empirical Bayes Approach to Efficient Portfolio Selection”. *Journal of Financial and Quantitative Analysis* 21.3 (Sept. 1986), pp. 293–305. ISSN: 1756-6916, 0022-1090. DOI: [10.2307/2331043](https://doi.org/10.2307/2331043).
- [Gel+13] Andrew Gelman et al. *Bayesian Data Analysis, Third Edition*. CRC Press, Nov. 2013. ISBN: 978-1-4398-4095-5.
- [Gel+20] Andrew Gelman et al. “Bayesian Workflow”. Nov. 2020. DOI: [10.48550/arXiv.2011.01808](https://doi.org/10.48550/arXiv.2011.01808). arXiv: [2011.01808 \[stat\]](https://arxiv.org/abs/2011.01808).
- [GLX21] Stefano Giglio, Yuan Liao, and Dacheng Xiu. “Thousands of Alpha Tests”. *The Review of Financial Studies* 34.7 (July 2021), pp. 3456–3496. ISSN: 0893-9454. DOI: [10.1093/rfs/hhaa111](https://doi.org/10.1093/rfs/hhaa111).
- [GP13] Nicolae Gârleanu and Lasse Heje Pedersen. “Dynamic Trading with Predictable Returns and Transaction Costs”. *The Journal of Finance* 68.6 (Dec. 2013), pp. 2309–2340. ISSN: 00221082. DOI: [10.1111/jofi.12080](https://doi.org/10.1111/jofi.12080).
- [Gup00] A. K. Gupta. *Matrix Variate Distributions*. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics 104. Boca Raton, FL: Chapman & Hall, 2000. ISBN: 978-1-58488-046-2.
- [GWZ24] Amit Goyal, Ivo Welch, and Athanasse Zafirov. “A Comprehensive 2022 Look at the Empirical Performance of Equity Premium Prediction”. *The Review of Financial Studies* 37.11 (Nov. 2024), pp. 3490–3557. ISSN: 0893-9454. DOI: [10.1093/rfs/hhae044](https://doi.org/10.1093/rfs/hhae044).
- [Hal] Max Halford. “Prince”.
- [Hal94] Edmond Halley. “Methodus Nova Accurata & Facilis Inveniendi Radices Aequationum Quarumcumque Generaliter, Sine Praviæ Reductione”. *Philosophical Transactions of the Royal Society of London* 18.210 (May 1694), pp. 136–148. DOI: [10.1098/rstl.1694.0029](https://doi.org/10.1098/rstl.1694.0029).

- [Har06] Frank E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2nd ed. edition. New York, NY: Springer-Verlag New York Inc., Jan. 2006. ISBN: 978-0-387-95232-1.
- [HB96] Robert A. Haugen and Nardin L. Baker. “Commonality in the Determinants of Expected Stock Returns”. *Journal of Financial Economics* 41.3 (July 1996), pp. 401–439. ISSN: 0304-405X. DOI: [10.1016/0304-405X\(95\)00868-F](https://doi.org/10.1016/0304-405X(95)00868-F).
- [Hil70] G. W. Hill. “Algorithm 396: Student’s t-Quantiles”. *Commun. ACM* 13.10 (Oct. 1970), pp. 619–620. ISSN: 0001-0782. DOI: [10.1145/355598.355600](https://doi.org/10.1145/355598.355600).
- [HL14] Campbell R. Harvey and Yan Liu. “Evaluating Trading Strategies”. *The Journal of Portfolio Management* 40.5 (Sept. 2014), pp. 108–118. ISSN: 0095-4918, 2168-8656. DOI: [10.3905/jpm.2014.40.5.108](https://doi.org/10.3905/jpm.2014.40.5.108).
- [HL15] Campbell R. Harvey and Yan Liu. “Backtesting”. *The Journal of Portfolio Management* 42.1 (Oct. 2015), pp. 13–28. ISSN: 0095-4918, 2168-8656. DOI: [10.3905/jpm.2015.42.1.013](https://doi.org/10.3905/jpm.2015.42.1.013).
- [HL18] Campbell R Harvey and Yan Liu. “Detecting Repeatable Performance”. *The Review of Financial Studies* 31.7 (July 2018), pp. 2499–2552. ISSN: 0893-9454. DOI: [10.1093/rfs/hhy014](https://doi.org/10.1093/rfs/hhy014).
- [HL20] Campbell R. Harvey and Yan Liu. “False (and Missed) Discoveries in Financial Economics”. *The Journal of Finance* 75.5 (2020), pp. 2503–2553. ISSN: 1540-6261. DOI: [10.1111/jofi.12951](https://doi.org/10.1111/jofi.12951).
- [HL21a] Campbell R. Harvey and Yan Liu. “Lucky Factors”. *Journal of Financial Economics* 141.2 (Aug. 2021), pp. 413–435. ISSN: 0304-405X. DOI: [10.1016/j.jfineco.2021.04.014](https://doi.org/10.1016/j.jfineco.2021.04.014).
- [HL21b] Campbell R. Harvey and Yan Liu. “Uncovering the Iceberg from Its Tip: A Model of Publication Bias and p-Hacking”. *SSRN Electronic Journal* (2021). ISSN: 1556-5068. DOI: [10.2139/ssrn.3865813](https://doi.org/10.2139/ssrn.3865813).
- [HLS20] Campbell R Harvey, Yan Liu, and Alessio Saretto. “An Evaluation of Alternative Multiple Testing Methods for Finance Applications*”. *The Review of Asset Pricing Studies* 10.2 (June 2020), pp. 199–248. ISSN: 2045-9920. DOI: [10.1093/rapstu/raaa003](https://doi.org/10.1093/rapstu/raaa003).
- [HLZ16] Campbell R. Harvey, Yan Liu, and Heqing Zhu. “... and the Cross-Section of Expected Returns”. *The Review of Financial Studies* 29.1 (Jan. 2016), pp. 5–68. ISSN: 0893-9454. DOI: [10.1093/rfs/hhv059](https://doi.org/10.1093/rfs/hhv059).
- [Hol79] Sture Holm. “A Simple Sequentially Rejective Multiple Test Procedure”. *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70. JSTOR: [4615733](https://www.jstor.org/stable/4615733).
- [HWD08] Herodotus, Robin Waterfield, and Carolyn Dewald. *The Histories*. Oxford World’s Classics. Oxford, New York: Oxford University Press, Apr. 2008. ISBN: 978-0-19-953566-8.

- [HWF15] Francis K. C. Hui, David I. Warton, and Scott D. Foster. “Tuning Parameter Selection for the Adaptive Lasso Using ERIC”. *Journal of the American Statistical Association* 110.509 (Jan. 2015), pp. 262–269. ISSN: 0162-1459. DOI: [10.1080/01621459.2014.951444](https://doi.org/10.1080/01621459.2014.951444).
- [HXZ20] Kewei Hou, Chen Xue, and Lu Zhang. “Replicating Anomalies”. *The Review of Financial Studies* 33.5 (May 2020), pp. 2019–2133. ISSN: 0893-9454, 1465-7368. DOI: [10.1093/rfs/hhy131](https://doi.org/10.1093/rfs/hhy131).
- [IL25] Anna Irrera and Justina Lee. “Billionaire Alex Gerko’s XTX to Build €1 Billion Data Hub in Machine-Learning Bet”. *Bloomberg.com* (Jan. 2025).
- [Ioa05] John P. A. Ioannidis. “Why Most Published Research Findings Are False”. *PLOS Medicine* 2.8 (Aug. 2005), e124. ISSN: 1549-1676. DOI: [10.1371/journal.pmed.0020124](https://doi.org/10.1371/journal.pmed.0020124).
- [JKP23] Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. “Is There a Replication Crisis in Finance?” *The Journal of Finance* 78.5 (2023), pp. 2465–2518. ISSN: 1540-6261. DOI: [10.1111/jofi.13249](https://doi.org/10.1111/jofi.13249).
- [Joe+21] Juha Joenväärä et al. “Hedge Fund Performance: Are Stylized Facts Sensitive to Which Database One Uses?” *Critical Finance Review* 10.2 (June 2021), pp. 271–327. ISSN: 2164-5744. DOI: [10.1561/104.00000104](https://doi.org/10.1561/104.00000104).
- [Jor86] Philippe Jorion. “Bayes-Stein Estimation for Portfolio Analysis”. *The Journal of Financial and Quantitative Analysis* 21.3 (1986), pp. 279–292. ISSN: 0022-1090. DOI: [10.2307/2331042](https://doi.org/10.2307/2331042). JSTOR: [2331042](https://www.jstor.org/stable/2331042).
- [Jou+24] Jacques Joubert et al. “The Three Types of Backtests”. SSRN Scholarly Paper. Rochester, NY, July 2024. DOI: [10.2139/ssrn.4897573](https://doi.org/10.2139/ssrn.4897573).
- [JR75] F. James and M. Roos. “Minuit - a System for Function Minimization and Analysis of the Parameter Errors and Correlations”. *Computer Physics Communications* 10.6 (Dec. 1975), pp. 343–367. ISSN: 0010-4655. DOI: [10.1016/0010-4655\(75\)90039-9](https://doi.org/10.1016/0010-4655(75)90039-9).
- [JS19] Philippe Jorion and Christopher Schwarz. “The Fix Is In: Properly Backing out Backfill Bias”. *The Review of Financial Studies* 32.12 (Dec. 2019), pp. 5048–5099. ISSN: 0893-9454. DOI: [10.1093/rfs/hhz024](https://doi.org/10.1093/rfs/hhz024).
- [JS60] W. James and Charles Stein. “Estimation with Quadratic Loss”. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Statistical Laboratory University of California: University of California Press, July 1960, pp. 361–379.
- [KF19] Adriano Koshiyama and Nick Firoozye. “Avoiding Backtesting Overfitting by Covariance-Penalties: An Empirical Investigation of the Ordinary and Total Least Squares Cases”. May 2019. arXiv: [1905.05023](https://arxiv.org/abs/1905.05023) [q-fin].
- [KMZ24] Bryan Kelly, Semyon Malamud, and Kangying Zhou. “The Virtue of Complexity in Return Prediction”. *The Journal of Finance* 79.1 (Feb. 2024), pp. 459–503. ISSN: 1540-6261. DOI: [10.1111/jofi.13298](https://doi.org/10.1111/jofi.13298).

- [KNT07] Robert Kosowski, Narayan Y. Naik, and Melvyn Teo. “Do Hedge Funds Deliver Alpha? A Bayesian and Bootstrap Analysis”. *Journal of Financial Economics* 84.1 (Apr. 2007), pp. 229–264. ISSN: 0304-405X. DOI: [10.1016/j.jfineco.2005.12.009](https://doi.org/10.1016/j.jfineco.2005.12.009).
- [KWZ24] Raymond Kan, Xiaolu Wang, and Xinghua Zheng. “In-Sample and out-of-Sample Sharpe Ratios of Multi-Factor Asset Pricing Models”. *Journal of Financial Economics* 155 (May 2024), p. 103837. ISSN: 0304-405X. DOI: [10.1016/j.jfineco.2024.103837](https://doi.org/10.1016/j.jfineco.2024.103837).
- [LB21] Marcos López de Prado and David H. Bailey. “The False Strategy Theorem: A Financial Application of Experimental Mathematics”. *The American Mathematical Monthly* 128.9 (Oct. 2021), pp. 825–831. ISSN: 0002-9890. DOI: [10.1080/00029890.2021.1965068](https://doi.org/10.1080/00029890.2021.1965068).
- [LKJ09] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. “Generating Random Correlation Matrices Based on Vines and Extended Onion Method”. *Journal of Multivariate Analysis* 100.9 (Oct. 2009), pp. 1989–2001. ISSN: 0047-259X. DOI: [10.1016/j.jmva.2009.04.008](https://doi.org/10.1016/j.jmva.2009.04.008).
- [LL17] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.
- [LL19] Marcos López de Prado and Michael J. Lewis. “Detection of False Investment Strategies Using Unsupervised Learning Methods”. *Quantitative Finance* 19.9 (Sept. 2019), pp. 1555–1565. ISSN: 1469-7688. DOI: [10.1080/14697688.2019.1622311](https://doi.org/10.1080/14697688.2019.1622311).
- [LLZ25] Marcos Lopez de Prado, Alexander Lipton, and Vincent Zoonekynd. “How to Use the Sharpe Ratio”. SSRN Scholarly Paper. Rochester, NY, Sept. 2025. DOI: [10.2139/ssrn.5520741](https://doi.org/10.2139/ssrn.5520741). Social Science Research Network: [5520741](https://ssrn.com/abstract=5520741).
- [Lo02] Andrew W Lo. “The Statistics of Sharpe Ratios”. *Financial Analysts Journal* (2002), p. 17.
- [Lóp19] Marcos López de Prado. “A Data Science Solution to the Multiple-Testing Crisis in Financial Research”. *The Journal of Financial Data Science* 1.1 (Jan. 2019), pp. 99–110. ISSN: 2640-3943. DOI: [10.3905/jfds.2019.1.099](https://doi.org/10.3905/jfds.2019.1.099).
- [LS96] Jacques Longerstaeey and Martin Spencer. *RiskMetrics*. New York: J.P.Morgan / Reuters, Dec. 1996.
- [Lun+20] Scott M. Lundberg et al. “From Local Explanations to Global Understanding with Explainable AI for Trees”. *Nature Machine Intelligence* 2.1 (Jan. 2020), pp. 56–67. ISSN: 2522-5839. DOI: [10.1038/s42256-019-0138-9](https://doi.org/10.1038/s42256-019-0138-9).
- [MAL21] Alvaro Mendez-Civieta, M. Carmen Aguilera-Morillo, and Rosa E. Lillo. “Adaptive Sparse Group LASSO in Quantile Regression”. *Advances in Data Analysis and Classification* 15.3 (Sept. 2021), pp. 547–573. ISSN: 1862-5355. DOI: [10.1007/s11634-020-00413-8](https://doi.org/10.1007/s11634-020-00413-8).
- [MAN19] MAN. “Crowding | Man Institute”. <https://www.man.com/maninstitute/crowding>. (accessed 2022-03-15). Feb. 2019.

- [Mar06] Howard Marks. “Memo from Howard Marks to Oaktree Clients on Risk”. Memo. Jan. 2006.
- [Mar52] Harry Markowitz. “Portfolio Selection”. *The Journal of Finance* 7.1 (1952), pp. 77–91. ISSN: 0022-1082. DOI: [10.2307/2975974](https://doi.org/10.2307/2975974). JSTOR: [2975974](https://www.jstor.org/stable/2975974).
- [MCW25] Xuran Meng, Yuan Cao, and Weichen Wang. “Estimation of Out-of-Sample Sharpe Ratio for High Dimensional Portfolio Optimization”. Jan. 2025. DOI: [10.48550/arXiv.2406.03954](https://doi.org/10.48550/arXiv.2406.03954). arXiv: [2406.03954](https://arxiv.org/abs/2406.03954) [math].
- [Mer02] Elmar Mertens. *Variance of the IID Estimator in Lo (2002)*. University of Basel, Nov. 2002.
- [MJM25] Joseph Mulligan, Antoine (Jack) Jacquier, and Johannes Muhle-Karbe. “In-Sample and Out-of-Sample Sharpe Ratios for Linear Predictive Models”. SSRN Scholarly Paper. Rochester, NY, July 2025. DOI: [10.2139/ssrn.5086171](https://doi.org/10.2139/ssrn.5086171). Social Science Research Network: [5086171](https://ssrn.com/abstract=5086171).
- [MOP12] Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen. “Time Series Momentum”. *Journal of Financial Economics*. Special Issue on Investor Sentiment 104.2 (May 2012), pp. 228–250. ISSN: 0304-405X. DOI: [10.1016/j.jfineco.2011.11.003](https://doi.org/10.1016/j.jfineco.2011.11.003).
- [MP16] R. David Mclean and Jeffrey Pontiff. “Does Academic Research Destroy Stock Return Predictability?” *The Journal of Finance* 71.1 (2016), pp. 5–32. ISSN: 1540-6261. DOI: [10.1111/jofi.12365](https://doi.org/10.1111/jofi.12365).
- [Mul+26] Joseph Mulligan et al. “Bayes Is All You Need: Better Estimates of True Sharpe Ratios”. SSRN Scholarly Paper. Rochester, NY, Feb. 2026. Social Science Research Network: [6300138](https://ssrn.com/abstract=6300138).
- [Mul01] P. Muller. “Proprietary Trading: Truth and Fiction”. *Quantitative Finance* 1.1 (Jan. 2001), pp. 6–8. ISSN: 1469-7688. DOI: [10.1080/713665546](https://doi.org/10.1080/713665546).
- [Mv00] S. A. Murphy and A. W. van der Vaart. “On Profile Likelihood”. *Journal of the American Statistical Association* 95.450 (2000), pp. 449–465. ISSN: 0162-1459. DOI: [10.2307/2669386](https://doi.org/10.2307/2669386). JSTOR: [2669386](https://www.jstor.org/stable/2669386).
- [MWW24] Johannes Muhle-Karbe, Zexin Wang, and Kevin Webster. “Stochastic Liquidity as a Proxy for Nonlinear Price Impact”. *Operations Research* 72.2 (Mar. 2024), pp. 444–458. ISSN: 0030-364X. DOI: [10.1287/opre.2022.0627](https://doi.org/10.1287/opre.2022.0627).
- [NM94] Whitney K. Newey and Daniel McFadden. “Chapter 36 Large Sample Estimation and Hypothesis Testing”. *Handbook of Econometrics*. Vol. 4. Elsevier, Jan. 1994, pp. 2111–2245. DOI: [10.1016/S1573-4412\(05\)80005-4](https://doi.org/10.1016/S1573-4412(05)80005-4).
- [Nov15] Robert Novy-Marx. *Backtesting Strategies Based on Multiple Signals*. Working Paper 21329. National Bureau of Economic Research, July 2015. DOI: [10.3386/w21329](https://doi.org/10.3386/w21329).
- [NV16] Robert Novy-Marx and Mihail Velikov. “A Taxonomy of Anomalies and Their Trading Costs”. *The Review of Financial Studies* 29.1 (Jan. 2016), pp. 104–147. ISSN: 0893-9454. DOI: [10.1093/rfs/hhv063](https://doi.org/10.1093/rfs/hhv063).

- [Par95] James M. Park. “Managed Futures as an Investment Class Asset”. PhD thesis. United States – New York: Columbia University, 1995. ISBN: 9798209518426.
- [Pat16] Nick Patterson. “Talking Machines: AI Safety and the Legacy of Bletchley Park, with Nick Patterson - Robohub”. <https://robohub.org/talking-machines-ai-safety-and-the-legacy-of-bletchley-park-with-nick-patterson/>. Feb. 2016.
- [Pav21] Steven E. Pav. *The Sharpe Ratio: Statistics and Applications*. New York: Chapman and Hall/CRC, Sept. 2021. ISBN: 978-1-003-18105-7. DOI: [10.1201/9781003181057](https://doi.org/10.1201/9781003181057).
- [Ped+11] Fabian Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928.
- [PP12] Kaare Brandt Petersen and Michael Syskind Pedersen. “Matrix Cookbook” (Nov. 2012).
- [Pro+19] Liudmila Prokhorenkova et al. “CatBoost: Unbiased Boosting with Categorical Features”. Jan. 2019. DOI: [10.48550/arXiv.1706.09516](https://doi.org/10.48550/arXiv.1706.09516). arXiv: [1706.09516 \[cs\]](https://arxiv.org/abs/1706.09516).
- [PS02] Ľuboš Pástor and Robert F. Stambaugh. “Mutual Fund Performance and Seemingly Unrelated Assets”. *Journal of Financial Economics* 63.3 (Mar. 2002), pp. 315–349. ISSN: 0304-405X. DOI: [10.1016/S0304-405X\(02\)00064-8](https://doi.org/10.1016/S0304-405X(02)00064-8).
- [PS20] Dirk Paulsen and Jakob Söhl. “Noise Fit, Estimation Error and a Sharpe Information Criterion”. *Quantitative Finance* 20.6 (June 2020), pp. 1027–1043. ISSN: 1469-7688. DOI: [10.1080/14697688.2020.1718746](https://doi.org/10.1080/14697688.2020.1718746).
- [PS92] Peter C. B. Phillips and Victor Solo. “Asymptotics for Linear Processes”. *The Annals of Statistics* 20.2 (1992), pp. 971–1001. ISSN: 0090-5364. JSTOR: [2241993](https://www.jstor.org/stable/2241993).
- [Pv03] Nolke Posthuma and Pieter Jelle van der Sluis. *A Reality Check on Hedge Funds Returns*. 2003-17. Amsterdam: Vrije Universiteit Amsterdam, Faculty of Economics and Business Administration, 2003.
- [Rao09] C. Radhakrishna Rao. *Linear Statistical Inference and Its Applications*. John Wiley & Sons, Sept. 2009. ISBN: 978-0-470-31714-3.
- [RLK24] Jason Rader, Terry Lyons, and Patrick Kidger. “Optimistix: Modular Optimisation in JAX and Equinox”. Feb. 2024. DOI: [10.48550/arXiv.2402.09983](https://doi.org/10.48550/arXiv.2402.09983). arXiv: [2402.09983 \[math\]](https://arxiv.org/abs/2402.09983).
- [Rob56] Herbert Robbins. “An Empirical Bayes Approach to Statistics”. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Vol. 3.1. University of California Press, Jan. 1956, pp. 157–164. ISBN: 978-0-520-31388-0. DOI: [10.1525/9780520313880-015](https://doi.org/10.1525/9780520313880-015).

- [Ros79] Robert Rosenthal. “The File Drawer Problem and Tolerance for Null Results”. *Psychological Bulletin* 86.3 (1979), pp. 638–641. ISSN: 1939-1455. DOI: [10.1037/0033-2909.86.3.638](https://doi.org/10.1037/0033-2909.86.3.638).
- [Rou87] Peter J. Rousseeuw. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”. *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. ISSN: 0377-0427. DOI: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [Sec02] Securities and Exchange Commission. *Proposed Amendments to Investment Company Advertising Rules*. Securities and Exchange Commission, May 2002.
- [Sha66] William F. Sharpe. “Mutual Fund Performance”. *The Journal of Business* 39.1 (1966), pp. 119–138. ISSN: 0021-9398. DOI: [10.1086/294846](https://doi.org/10.1086/294846). JSTOR: [2351741](https://www.jstor.org/stable/2351741).
- [Sha94] William F. Sharpe. “The Sharpe Ratio”. *The Journal of Portfolio Management* 21.1 (Oct. 1994), pp. 49–58. ISSN: 0095-4918, 2168-8656. DOI: [10.3905/jpm.1994.409501](https://doi.org/10.3905/jpm.1994.409501).
- [Sim43] Simpson. *Mathematical Dissertations on a Variety of Physical and Analytical Subjects*. 1743.
- [SLP17] Antti Suhonen, Matthias Lennkh, and Fabrice Perez. “Quantifying Backtest Overfitting in Alternative Beta Strategies”. *The Journal of Portfolio Management* 43.2 (Jan. 2017), pp. 90–104. ISSN: 0095-4918, 2168-8656. DOI: [10.3905/jpm.2017.43.2.090](https://doi.org/10.3905/jpm.2017.43.2.090).
- [SW91] David J. Saville and Graham R. Wood. *Statistical Methods: The Geometric Approach*. Ed. by Stephen Fienberg and Ingram Olkin. Springer Texts in Statistics. New York, NY: Springer, 1991. ISBN: 978-1-4612-6965-6 978-1-4612-0971-3. DOI: [10.1007/978-1-4612-0971-3](https://doi.org/10.1007/978-1-4612-0971-3).
- [Tho16] Peter Thonemann. “Croesus and the Oracles”. *The Journal of Hellenic Studies* 136 (2016), pp. 152–167. ISSN: 0075-4269. JSTOR: [44157499](https://www.jstor.org/stable/44157499).
- [TM66] Jack L. Treynor and Kay K. Mazuy. “Can Mutual Funds Outguess the Market?” *Harvard Business Review* 44.4 (1966/07//Jul/Aug1966), pp. 131–136. ISSN: 00178012.
- [Vir+20] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2).
- [War63] Joe H. Ward Jr. “Hierarchical Grouping to Optimize an Objective Function”. *Journal of the American Statistical Association* 58.301 (Mar. 1963), pp. 236–244. ISSN: 0162-1459. DOI: [10.1080/01621459.1963.10500845](https://doi.org/10.1080/01621459.1963.10500845).
- [Web23] Kevin T. Webster. *Handbook of Price Impact Modeling*. Boca Raton: Chapman and Hall/CRC, 2023. ISBN: 978-1-032-32822-5.

- [Wie+16] Thomas Wiecki et al. *All That Glitters Is Not Gold: Comparing Backtest and Out-of-Sample Performance on a Large Cohort of Trading Algorithms*. SSRN Scholarly Paper ID 2745220. Rochester, NY: Social Science Research Network, Mar. 2016. DOI: [10.2139/ssrn.2745220](https://doi.org/10.2139/ssrn.2745220).
- [WS02] Graham R. Wood and David J. Saville. “A New Angle on the T-Test”. *Journal of the Royal Statistical Society. Series D (The Statistician)* 51.1 (2002), pp. 99–104. ISSN: 0039-0526. JSTOR: [3650393](https://www.jstor.org/stable/3650393).
- [WS05] Graham R. Wood and David J. Saville. “The Ubiquitous Angle”. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.1 (2005), pp. 95–107. ISSN: 1467-985X. DOI: [10.1111/j.1467-985X.2004.00338.x](https://doi.org/10.1111/j.1467-985X.2004.00338.x).
- [Yek12] Daniel Yekutieli. “Adjusted Bayesian Inference for Selected Parameters”. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 74.3 (2012), pp. 515–541. ISSN: 1369-7412. JSTOR: [41674641](https://www.jstor.org/stable/41674641).