ECONOMETRIC METHODS FOR IMPLEMENTING DECISION FUNCTIONS

Filip Klimenka

Oriel College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2017

Acknowledgments

I would like to thank my advisor, James Wolter, without whom this thesis would have never happened. I am also thankful to my second advisor Kevin Sheppard, Bent Nielsen and seminar participants at Oxford for their helpful comments and discussion. I am also very grateful for the financial support from Oxford-Man Institute and the Economics Department at the University of Oxford. Finally, I would like to thank my family and Vitalina for their unconditional support throughout my time at Oxford.

Econometric methods for implementing decision functions

Filip Klimenka
Oriel College

Submitted for the degree of
Doctor of Philosophy
Trinity Term 2017

Abstract

This thesis develops econometric methods for implementing data-based decisions. Decisions are viewed as functions of parameters which are estimated from the data. Standard methods focus on providing precise estimates of parameters ignoring intention to use them in decisions. My thesis focuses on designing methods to minimize the expected error in decision functions. The first chapter develops model averaging estimators in multiple regressions that minimize the mean squared error (MSE) of a chosen decision function. Our motivating example is implementing a portfolio choice rule that depends on variables included in assets' returns specification. We characterize the asymptotic MSE of decisions functions based on different models and then describe model-selection and averaging estimators that enable improvements in the MSE. The performance of our method is demonstrated with extensive simulations and empirical applications to futures data. The second chapter describes the risk improvements for a model averaging using two models. This type of averaging is known as shrinkage. Since the risk improvement is over the function of parameters, this shrinkage is referred to as focused shrinkage. The estimator is a weighted average between unrestricted and restricted models. The latter is a minimum distance estimator and requires selecting a projection matrix. The risk improvement of our shrinkage estimator over maximum-likelihood for arbitrary projection matrices is derived. I then show in an application to portfolio choice, that for a specific choice of projection matrix, this improvement can be substantial. The third chapter considers an application of the focused shrinkage estimator to the Global Minimum Variance (GMV) portfolio. Implementing the GMV portfolio requires estimating a covariance matrix and the literature has offered several estimators. Focused shrinkage is particularly suitable here because it can be used to directly minimize the MSE of the GMV portfolio. We illustrate the benefits of our estimator by conducting extensive simulations and empirical applications.

Contents

Introduction **1**

1 Multiple model averaging and the focused information criterion with an application to portfolio choice **6**

1.1 Introduction . . . 6
1.2 The model . . . 9
1.2.1 Defining submodels . . . 11
1.2.2 OLS estimators . . . 13
1.3 Asymptotic framework . . . 14
1.3.1 Focus parameter and its limiting distribution . . . 15
1.4 FIC and Model Averaging . . . 17
1.4.1 Plug-in averaging estimator . . . 18
1.5 Simulations . . . 20
1.5.1 Optimal portfolio as a focus function . . . 20
1.5.2 Simulation specification and results . . . 21
1.6 Commodity Futures Trading Application . . . 25
1.7 Conclusion . . . 28
1.8 Appendix . . . 33
1.9 Appendix A . . . 34
1.9.1 Proof of Theorem A1 . . . 37
1.9.2 Proof of Theorem 1 . . . 44
1.9.3 Proof of Lemma A2 . . . 47
1.9.4 Proof of Theorem A3 . . . 49
1.10 Appendix B . . . 50
1.10.1 Asymptotic distributions of FIC and plug-in averaging estimator . . . 51
1.10.2 Valid confidence interval . . . 52
1.10.3 Proof of Theorem B1 . . . 54
1.10.4 Proof of Theorem B2 . . . 56
1.10.5 Proof of Lemma B3 . . . 58

2 Focused shrinkage with an application to portfolio choice **61**

2.1 Introduction . . . 61
2.2 Risk of Minimum Distance Estimators . . . 66

2.3 Focused Shrinkage . . . 69
2.3.1 The Choice of V . . . 72
2.3.2 The Choice of Θ_0 . . . 74
2.3.3 Focused Shrinkage with Unbalanced W . . . 75
2.4 Risk and Portfolio Choice . . . 76
2.4.1 Asset Dynamics and a Trading Rule . . . 76
2.4.2 Localization and Full Model Estimation . . . 79
2.4.3 Parameter Subspaces . . . 81
2.4.4 Baseline Specification . . . 82
2.4.5 Simulated Risk for the Trading Rule . . . 84
2.5 Portfolio Choice for Futures Contracts . . . 90
2.6 Conclusion . . . 92
2.7 Appendix . . . 92
2.7.1 Risk Simulations and Bounds . . . 92
2.7.2 Proof of Theorem 1 . . . 107
2.7.3 Proof of Theorem 2 . . . 109
2.7.4 Proof of Theorem 3 . . . 111

3 Focused shrinkage estimators for the global minimum variance portfolio 118
3.1 Introduction . . . 118
3.2 Global minimum variance problem and existing estimation approaches . . . 124
3.3 Focused GMV estimator . . . 126
3.3.1 Unrestricted estimator . . . 127
3.3.2 Restricted estimator . . . 130
3.3.3 Optimal focused shrinkage intensity w . . . 132
3.3.4 Computation of the Focused GMV estimator of the covariance matrix 133
3.4 Simulations . . . 135
3.4.1 Set up . . . 135
3.4.2 Competitors . . . 137
3.4.3 Simulation results . . . 139
3.5 Empirical application . . . 142
3.5.1 Data . . . 142
3.5.2 Methodology of performance evaluation . . . 143
3.5.3 Empirical results . . . 144
3.6 Conclusion . . . 148
3.7 Appendix . . . 148
3.7.1 Ledoit and Wolf (2003) estimator . . . 148
3.7.2 Frahm and Memmel (2010) estimator . . . 149
3.7.3 De Miguel et al. (2009) cross-validation strategies . . . 150

Introduction

Every statistician or data analysts has to make decisions. These decisions start with choosing between a simpler or a more complicated model for understanding a given phenomenon. This often leads to selecting variables to include in the model to produce the most precise estimates of its parameters. The cost of taking these decisions is the resulting error in the estimated parameters. Estimation error is omnipresent when working with economic data where typically only weak effects are present. A researcher always faces uncertainty about which model to use to capture relevant effects. Whether it is a problem of evaluating effects of economic policies or computing the optimal investment allocations, estimation error of the parameters of a chosen model can outweigh any benefit of its complicated structure developed to capture richer economic dynamics. More data, however, may not always be available. Furthermore, in cases, when it is available, it may not be of help because one needs to account for changing model parameters.

The price to pay for estimation error becomes apparent when decisions are made on the basis of a model. Central banks evaluate different characteristics of the economy and make decisions about interest rates on the basis of their models. These decisions have real implications for the subsequent course of the economy. Another example is financial institutions that decide on allocating investments into different assets. These assets' investment allocations often depend on estimates of expected returns and covariance structure. Financial data sets are known for having low signal-to-noise ratios. As a consequence, estimation error can be of considerable magnitude. Notably, DeMiguel et al. (2009) questions whether time-series data has any use in estimating the optimal investment allocations, comparing it to the equally weighted portfolio which does not use any data at all.

In an attempt to deal with estimation error, a researcher is often tempted to fine-tune specification of their model until they produce desirable results using the same data set. This results in a serious overfitting problem and drastic negative consequences for the performance

of their models on new data sets.

Fortunately, the statistical and econometric literature offers a plethora of methodology for reducing estimation error. These include standard information criteria such Akaike and Bayesian information criteria; see Burnham and Anderson (2002) for an excellent overview. Other methods include penalized regressions (Hastie et al., 2001) and shrinkage methods (Saleh, 2006) that help with finding the optimal trade-off between the bias and variance of the estimated parameters and are shown to achieve considerable improvements in the out-of-sample performance. Among shrinkage methods, model averaging estimators have been shown to be particularly successful (Timmerman, 2006). On a theoretical side, a number of recent papers show the desirable properties of model averaging. See Hansen (2007), Geweke and Amisano (2011), Liang et al. (2011), Elliot et al. (2013), Elliot et al. (2015).

Many of these methods are optimal in a sense of producing the lowest expected loss (i.e. mean squared error) in the parameter themselves. This is good enough when the objective is to obtain precise estimates of the model parameters. However, estimation of the model is often only a starting point of the analysis, and the final objective is to make decisions. Formally speaking, decisions are functions of the parameters and can be either ad hoc or obtained as a solution to an optimization problem. When decisions are to be made, a researcher is more interested in a method that directly reduces the estimation error in the decision function themselves than parameters. In that sense, existing methods generally ignore the intention to use them in any decision functions.

A more recent line of literature attempts to address this problem by developing methods that minimize the estimation error of decision functions. Claeskens and Hjort (2003) propose the so called focused information criterion (FIC) that takes into account the function of the parameters to be estimated. They subsequently apply this criterion to their model-averaging estimators (Claeskens and Hjort, 2003a). Liu (2014) studies the properties of model averaging with the focused information criterion for univariate least-squares regressions. My thesis further explores this line of research of tailoring statistical estimation of the parameters toward the decision function of interest.

The first paper considers model selection and model averaging using the FIC in multiple regressions. The motivating example is implementing a mean-variance portfolio - our chosen focus function - that depends on mean and covariances of the asset returns. Each asset return can be driven by a set of factors, and different models are summarized by a subset

of factors included into a multiple regression. Different models result in different estimates of the portfolio rule and create a need for model selection. An approach is developed that allows us to find a model with minimal expected error in the portfolio weights. This is done by formalizing the bias-variance trade-off of estimating the focus function. To obtain the bias and variance of different estimates of portfolio weights (focus functions) we need the asymptotic distributions of different submodel estimates. These are derived using a localization framework. The localization is done on both regressions coefficients describing asset returns and the error covariances. We then map the asymptotic distribution of each of the submodels to the focus functions based on these models. As a result, we are able to obtain expression for the asymptotic mean square error for each submodel and then select the submodel with the smallest asymptotic MSE.

Choosing a single model can, however, be suboptimal. It can be possible to decrease the expected error by averaging over submodels. We derive the limiting distribution of the model averaging estimator of the decision function for fixed weights and then provide an algorithm to select the weights which minimizes the expected error of our focus function.

We illustrate the merits of our focused model-averaging approach in a portfolio choice application where our focus function is the solution to a mean-variance portfolio choice problem. Extensive simulations and an empirical application to several futures data sets show considerable improvements over standard methods including standard OLS, AIC / BIC selected models and the equally-weighted model averaging estimator - a benchmark that is hard to beat (Elliott et al., 2013).

This paper is resubmitted to *Journal of Business and Economic Statistics* jointly with Dr. James Wolter.

The second paper considers a more detailed characterization of the risk (expected error) improvement for a model-averaging estimator. Building upon the work by Hansen (2016), we consider averaging between two models. This type of averaging is usually referred to as shrinkage. Following the focused information criterion approach, our shrinkage estimator is designed to minimize the mean square error of a chosen focus function.

The shrinkage estimator considered is expressed as a weighted average between unrestricted and restricted models. The unrestricted model is estimated by OLS (or maximum-likelihood) with all potential variables included into the specification. Our restricted model estimator is in the class of minimum distance estimators. This requires selecting a projection matrix.

This is chosen as a matrix based on the derivatives of our focus function. We show this choice minimizes bias among estimators in this class. These two estimators are then incorporated into a shrinkage procedure. We derive a risk bound for this shrinkage estimator for an arbitrary projection matrix. This bound describes the risk improvement over maximum-likelihood of shrinkage estimators constructed this way. Using this result, we show that the proposed restricted model can lead to substantially lower risk compared to maximum-likelihood. The improvement is largest when the restricted model has nontrivial bias. This is because our proposed projection matrix minimizes the bias. We then apply this estimator in our motivating example: implementing a mean-variance portfolio choice rule with transaction costs. Extensive simulations demonstrate improved risk in this case. The estimator is also implemented in a portfolio choice application to futures data. Our approach outperforms standard procedures here as well.

An important difference from the model-averaging estimator in the first chapter is that this approach does not require estimation of the localization parameter. Instead it requires selecting a tuning parameter in the weight function (shrinkage intensity). In that sense it is similar to the soft-thresholding estimators studied in Judge and Bock (1976) and Saleh (2006) that are known for their improved risk properties compared to hard-thresholding (model-selection) estimators. Another difference is that we are able to describe an analytic risk bound of our estimator.

This chapter is a joint work with Dr. James Wolter.

In the third chapter I apply the focused shrinkage approach to the Global Minimum Variance (GMV) portfolio. The GMV portfolio chooses asset weights to minimize total portfolio variance. This has been shown to have a better performance compared to standard Markowitz portfolios in many situations (Clarke et al. (2006) and Clarke et al. (2011)). Markowitz portfolios can produce unstable weights because their implementation requires estimates of expected returns. These estimates often have significant estimation error (see Michaud, 1989). Implementing the GMV portfolio excludes expected returns from the optimization problem and only requires a covariance matrix estimate. In this way it can reduce estimation error. Standard choice is to use the sample covariance matrix. This can have a large estimation error when the number of assets is large. Because of this, the statistical and financial literature has offered a number of approaches to this problem. These include shrinkage estimators of the covariance matrix proposed by Ledoit and Wolf (2003, 2004)

and factor-model based covariance matrices (Fan et al. (2008)). A more recent literature directly treats portfolio weights as the final object of interest. Jagannathan and Ma (2003) suggested imposing short sale constraints on portfolio weights in the GMV problem. DeMiguel et al. (2009) extended this approach to a wider class of norm-constrained portfolios. For high-dimensional portfolios Fan et al. (2012) analyzed 1-Norm constraints on weights to reduce the estimation error inherited from estimating high dimensional covariance matrices. Finally, Frahm and Memmel (2010) propose estimators which minimize the out-of-sample variance of portfolio return by shrinking GMV weights towards the equally-weighted portfolio.

While the outlined studies target portfolio weights by restricting them using different constraints, they are in fact equivalent to solving the standard unconstrained GMV problem with the covariance matrix replaced by an appropriate shrinkage version. Some of these shrinkage schemes are shown to be optimal in a sense of minimizing the risk of the shrinkage version of covariance matrix, other are shown to have a bound on the resulting risk. These approaches, however, ignore the final object of interest: GMV portfolio weights and their associated risk. In contrast, the focused shrinkage estimator developed the second chapter can be applied to directly minimize the mean squared error of the estimated GMV portfolio.

The proposed estimator of the covariance matrix is based on a factor model. This is because shrinkage estimators show the best improvements with a sensible choice of a restricted model. While selecting a restricted model for the full covariance matrix is generally difficult, the residual covariance matrix from a factor model is expected to have an approximately diagonal structure that can be used as our restricted model. The resulting estimator weights between factor-model covariances estimated with unrestricted and restricted residual covariances. We illustrate the performance of our estimator by conducting extensive simulations designed to realistically represent U.S. stock market dynamics. The focused shrinkage estimator shows the best performance in terms of out-of-sample portfolio variances compared to nine standard competitors. Empirical applications to sorted Fama and French and industry portfolios show similar improvements.

Chapter 1

Multiple model averaging and the focused information criterion with an application to portfolio choice

Abstract. We consider multiple regression (MR) model averaging using the Focused Information Criterion (FIC). Our approach is motivated by the problem of implementing a mean-variance portfolio choice rule. The usual approach is to estimate parameters ignoring the intention to use them in portfolio choice. We develop an estimation method that focuses on the trading rule of interest. Asymptotic distributions of submodel estimators in the MR case are derived using a localization framework. The localization is of both regression coefficients and error covariances. Distributions of submodel estimators are used for model selection with the FIC. This allows comparison of submodels using the risk of portfolio rule estimators. FIC model averaging estimators are then characterized. This extension further improves risk properties. We show in simulations that applying these methods in the portfolio choice case results in improved estimates compared with several competitors. An application to futures data shows superior performance as well.

JEL: C31, C52, C53, C58.

1.1 Introduction

Since the seminal work of Markowitz (1952), mathematical descriptions of optimal portfolio selection have received great academic interest. In addition, mean-variance portfolio optimization has been implemented extensively in industry practice. Many institutions directly or indirectly use these methods. These include pension funds, mutual funds, university endowments, charitable foundations and hedge funds to name a few. A substantial portion of overall investments are made with these principles. See Garleanu and Pedersen (2013),

DeMiguel, Mei and Nogales (2016) and Collin-Dufresne, Daniel, Moallemi and Saglam (2015) for recent contributions in this area.

If an investor wishes to implement a mean-variance portfolio rule, she must estimate the underlying return structure. Standard approaches separate statistical estimation from implementation. The parameters of the model are first estimated. Then estimates are plugged into policy functions when making decisions. The first stage ignores the second.

In this paper, we propose a different methodology which uses policy functions in estimation. When trying to estimate a portfolio rule, the parameters of the underlying return structure are only tangentially relevant. Our main interest lies in the optimal policy, which is a transformation of these parameters. Because of this, we develop a procedure focused on the policy, not the parameters themselves. Estimating the parameters is only a means to an end.

Our results are motivated with a mean-variance portfolio rule assuming asset returns follow a multiple regression (MR) structure. In this setup, returns equations depend on arbitrary covariates. In applications, a number of potential covariates are usually available, but we are unsure which are important. Additionally, there is often limited historical data. Which covariates should be included to optimize bias-variance trade-off in finite samples is not obvious. In our setup, this trade-off is distorted by the policy function. The importance of a particular parameter depends on how much it influences the policy. Standard model selection methods such as AIC/BIC do not account for this. Our method specifically optimizes bias-variance trade-off for the function of interest.

In the sequel, a localization structure is used to formalize the bias-variance trade-off when estimating a policy. The resulting methods are a version of the focused information criterion (FIC). The FIC was first proposed by Claeskens and Hjort (2003) in the likelihood case. Since then, it has been further developed for regression cases. Our approach is most closely related to Liu (2014). Asymptotic theory is derived for an arbitrary MR and policy rule. The portfolio choice setup considered below is a special case.

In order to implement FIC model selection, a subset of the regression coefficients are assumed local-to-zero. Specifically, coefficients have the form $\delta/\sqrt{T}$ where T is the number of observed periods and δ is an arbitrary constant. Mean-variance portfolio optimization also depends on the covariance matrix of shocks in the MR specification. A local-to-zero structure on the off-diagonal components of this covariance matrix is also assumed.

If we set a local-to-zero parameter of the MR model exactly to zero, this produces a

submodel where that parameter is fixed. By considering subsets of local-to-zero parameters we define a class of submodels: the class where given subsets of parameters are fixed at zero. Defining a class of submodels is the starting point for the FIC. Which submodels are considered is arbitrary and user determined.

The first contribution of this paper is to derive asymptotic distributions of OLS estimators of MR submodels assuming the local-to-zero parameter structure described above. In the class of submodels whose distributions are characterized, both regression parameters and error covariance matrix components can be fixed at zero. We derive the joint asymptotic distribution of submodel estimators for both sets of parameters. This extends previous FIC regression results which do not consider MR models. Previous work has not considered error covariance matrix estimates or localization of covariance terms.

Because of the $\sqrt{T}$ localization, submodel estimators are asymptotically normal with non-zero mean and variance. The non-zero mean results from bias introduced by setting local-to-zero parameters exactly to zero. Parameter estimates are plugged into policy functions when making decisions. Parameters fixed at zero must also be plugged in. Both the estimated and fixed parameters influence the asymptotic distribution of the policy. Estimated parameters influence both bias and variance. Fixed parameters contribute only to bias. The final result is an asymptotically normal estimator of the policy with non-zero mean and variance.

The distributions described above formalize the bias-variance trade-off of submodel estimation. Different submodels and policy functions result in different mean and variance terms. Once these distributions are determined they can be used to rank submodels. In this paper, we rank submodels by the asymptotic mean square error (AMSE) of their estimated policy. Other loss functions are also possible. Estimated submodel AMSEs give the FIC ranking.

One possibility is to choose the single submodel with the best FIC ranking. However, this can be suboptimal. It may be possible to decrease AMSE by averaging submodels. In what follows, we average submodels to minimize AMSE as in Liu (2014). A closed form expression for AMSE of policy function estimates is derived for the model averaging case. This is used to construct an estimator with optimal weights. There have been a number of recent papers showing the desirable properties of model averaging. Clark and McCracken (2009) average models for forecasting using a similar setup as this paper. Timmermann (2006) surveys the forecast combination literature. See Hansen (2007); Geweke and Amisano (2011); Liang, Zou, Wan and Zhang (2011) and Elliot, Gargano and Timmermann (2013, 2015) for more examples

and citations.

Simulation studies were conducted to compare the performance of the proposed FIC model averaging methodology to standard approaches. This method is compared with policies estimated with OLS, models selected by AIC/BIC, an equally-weighed model averaging scheme and single model selection using the FIC. FIC model averaging emerges as the most robust method, consistently ranked as the best or second-best across various scenarios.

An empirical application of these methods to futures markets was also conducted. Trading in several sets of commodity futures was considered using FIC model averaging. Optimal portfolios derived in Garleanu and Pedersen (2013) were estimated with rolling windows. The results show the proposed method significantly outperforms a set of alternatives in both Sharpe ratio and absolute returns.

The paper is organized as follows. Section 2 presents the model, defines submodels and describes basic estimators. Section 3 describes the localization assumptions used for model selection. Asymptotic distributions of submodel estimators of focus functions are then derived. Section 4 shows how the results from Section 3 can be used for FIC model selection. FIC model averaging estimators and their properties are then presented. Finally, a plug-in model averaging estimator is proposed. Section 5 presents simulations examining the performance of the plug-in model averaging estimator against several competitors. Section 6 presents an empirical application to commodity futures data. Section 7 concludes.

1.2 The model

Consider an M-dimensional predictive MR model:

$$y_{t+1} \;=\; X_t'\beta + Z_t'\gamma + \epsilon_{t+1},$$

defined with the following matrices:

$$
\underset{M\times 1}{y_{t+1}} = \begin{bmatrix} y_{1t+1} \\ \vdots \\ y_{Mt+1} \end{bmatrix}, \underset{P\times M}{X_t} = \begin{bmatrix} x_{1t} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & x_{Mt} \end{bmatrix}, \underset{K\times M}{Z_t} = \begin{bmatrix} z_{1t} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & z_{Mt} \end{bmatrix}. \tag{1.1}
$$

$$
\underset{M\times 1}{\epsilon_{t+1}} = \begin{bmatrix} \epsilon_{1t+1} \\ \vdots \\ \epsilon_{Mt+1} \end{bmatrix}, \underset{P\times 1}{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_M \end{bmatrix}, \underset{K\times 1}{\gamma} = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_M \end{bmatrix}
$$

In this equation, covariates are divided into core X_t and auxiliary Z_t. For equation i, x_{it} is a $p_i \times 1$ vector of core variables. Similarly, z_{it} is a $k_i \times 1$ vector of auxiliary variables. The total number of core and auxiliary regressors in all equations are $P = \sum_{i=1}^{M} p_i$, $K = \sum_{i=1}^{M} k_i$. Core variables will always be included in estimation. The researcher is unsure which auxiliary covariates should be included. Our model selection and averaging procedures will decide this while focusing on a policy function of interest.

Further, let

$$
H_t = \begin{bmatrix} X_t \\ Z_t \end{bmatrix}, \ \theta = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \tag{1.2}
$$

and

$$
y_{t+1} = H_t'\theta + \epsilon_{t+1}. \tag{1.3}
$$

Finally, we stack observations with the following matrices

$$
\begin{aligned}
y &= \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}, \ X = \begin{bmatrix} X_0' \\ \vdots \\ X_{T-1}' \end{bmatrix}, \ Z = \begin{bmatrix} Z_0' \\ \vdots \\ Z_{T-1}' \end{bmatrix}, \\
\epsilon &= \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{bmatrix}, \ H = \begin{bmatrix} X & Z \end{bmatrix}.
\end{aligned} \tag{1.4}
$$

All observations of (1.3) can be compactly written as:

$$\begin{aligned} y &= X\beta + Z\gamma + \epsilon, \\ &= H\theta + \epsilon. \end{aligned} \tag{1.5}$$

The portfolio choice rule considered in the sequel depends on the covariance matrix of ϵ_t ($cov(\epsilon_t) = \Omega$). Therefore, an estimator of this covariance matrix is required. All parameters in Ω are classified as either core or auxiliary. It is also possible to *a priori* restrict some covariances to zero. We rule this out for simplicity. The elements of Ω are stacked into a vector Θ for use in asymptotic results that follow. Core elements (which must contain all variances) are stacked first, followed by auxiliary covariances:

$$\Theta = \begin{bmatrix} \Phi \\ \\ \Upsilon \end{bmatrix} \tag{1.6}$$

The notation G is used for the number of core elements and N for the number of auxiliary. The order of these parameters is not important. As with the regression coefficients, this division of parameters is used in model selection and averaging procedures described in the sequel.

1.2.1 Defining submodels

In order to conduct model selection and averaging, we first define submodels. Submodels are formed by taking the full model and restricting some subset of (γ, Υ) to zero. In the sequel, we consider an arbitrary set of submodels. In applications, those considered are user determined. These could come from economic theory, computational cost restrictions or other considerations. As a starting point, we assume a set of $\bar{M}$ distinct submodels. These are taken as given.

The subscript m is used to represent an arbitrary submodel. Let Π_m be the $K_m \times K$ selection matrix that selects the included auxiliary regressors in m. This can be constructed by removing those rows from the K-dimensional identity matrix whose order corresponds to

the column order of excluded auxiliary regressors in Z. Let

$$Z_m \;=\; Z\Pi_m^{'}$$

be the resulting matrix of retained auxiliary regressors. Submodel m thus includes all core covariates X and a subset of auxiliary covariates Z_m. This results in $P + K_m$ regressors. Define also the selection matrices

$$S_0 \;=\; \begin{bmatrix} 0_{P\times K} \\ I_K \end{bmatrix},\; S_m \;=\; \begin{bmatrix} I_P & 0_{P\times K_m} \\ 0_{K\times P} & \Pi_m^{'} \end{bmatrix}. \tag{1.7}$$

Note that

$$\begin{aligned} H_m \;&=\; HS_m \;=\; \begin{bmatrix} X & Z_m \end{bmatrix}, \\ \theta_m \;&=\; S_m^{'}\theta \;=\; \begin{bmatrix} \beta \\ \Pi_m\gamma \end{bmatrix} = \begin{bmatrix} \beta \\ \gamma_m \end{bmatrix}, \end{aligned} \tag{1.8}$$

where θ_m is the subset of regression coefficients θ which are not restricted to zero. These are the parameters that will be estimated when considering submodel m. Also note that the full model corresponds to the case $\Pi_m = I$, and the narrow model to Π_m being empty.

Let Γ_m be an $N_m \times N$ matrix that selects the included error covariances from Υ. Similar to before, this can be constructed by removing the rows from the $N \times N$ identity matrix whose order corresponds to the row order of excluded auxiliary covariances in Υ. The mth submodel's covariances can be represented as

$$\Theta_{mm} \;=\; \begin{bmatrix} \Phi_m \\ \Upsilon_{mm} \end{bmatrix} \;=\; \begin{bmatrix} \Phi_m \\ \Gamma_m\Upsilon_m \end{bmatrix}$$

where $\Phi_m = \Phi$ and $\Upsilon_m = \Upsilon$ are vectors that denote all core and all auxiliary covariances. We write $\Theta_m = \begin{bmatrix} \Phi_m^{'} & \Upsilon_m^{'} \end{bmatrix}^{'}$. Covariance terms not in Θ_{mm} are set to zero in submodel m. Finally, let

$$F_0 \;=\; \begin{bmatrix} 0_{G\times N} \\ I_N \end{bmatrix},\; F_m \;=\; \begin{bmatrix} I_G & 0_{G\times N_m} \\ 0_{N\times G} & \Gamma_m^{'} \end{bmatrix} \tag{1.9}$$

be selection matrices such that

$$\Theta_{mm} \;=\; F_m^{'}\Theta_m \;=\; \begin{bmatrix} \Phi_m \\ \\ \Gamma_m\Upsilon_m \end{bmatrix}. \tag{1.10}$$

The full model corresponds to the case $F_m = I$.

In the context of estimation this notation is not redundant. The subscript m represents the restriction on regression parameters (through Π_m) which influences residuals. Residuals influence estimates of Θ. Because of this, in general, different submodels m_1 and m_2 will have $\hat{\Phi}_{m_1} \neq \hat{\Phi}_{m_2}$ and $\hat{\Upsilon}_{m_1} \neq \hat{\Upsilon}_{m_2}$. Υ_{mm} denotes a subset of auxiliary residual covariances Υ_m defined using Γ_m. In going from Υ_m to Υ_{mm}, certain covariance terms from Υ_m are removed. The removed terms are assumed to be zero and are not estimated. In the notation Υ_{mm}, the first subscript m relates to the restriction of parameters through Π_m which alters residuals, the second subscript m to the actual restriction of covariances through Γ_m.

To sum up, a given submodel m is characterized by selection matrices (S_m, F_m). Parameters selected will be estimated under the submodel restrictions. The remaining parameters are set to zero in estimation.

1.2.2 OLS estimators

The parameters in a given submodel are estimated with OLS. The estimators for submodel m are

$$\begin{aligned} \hat{\theta}_m \;&=\; \begin{bmatrix} \hat{\beta}_m \\ \\ \hat{\gamma}_m \end{bmatrix} = \left(H_m^{'}H_m\right)^{-1} H_m^{'}y \\ \\ \hat{\Theta}_{mm} \;&=\; \begin{bmatrix} \hat{\Phi}_m \\ \\ \hat{\Upsilon}_{mm} \end{bmatrix} = \begin{bmatrix} \hat{\Phi}_m \\ \\ \Gamma_m\hat{\Upsilon}_m \end{bmatrix} = F_m^{'} \begin{bmatrix} \frac{1}{T}\left(\hat{\epsilon}_{im}^{'}\hat{\epsilon}_{jm}\right)^{core} \\ \\ \vdots \\ \\ \frac{1}{T}\left(\hat{\epsilon}_{km}^{'}\hat{\epsilon}_{lm}\right)^{aux} \\ \\ \vdots \end{bmatrix}, \quad i,j,k,l \in \{1,\ldots,M\} \end{aligned} \tag{1.11}$$

where $\hat{\epsilon}_{im}$ denotes a vector of residuals related to a dependent variable y_i for submodel m. The coefficients of the excluded variables γ_{m^c} and covariances Υ_{mm^c} (where c denotes complement) are set to zero.

If $\Pi_m = I_K$ and $\Gamma_m = I_N$, the formulas give least squares estimators of the full model. If Π_m and Γ_m are empty, we have estimators of the narrow model where all auxiliary parameters are set to zero. Note that in submodel estimators the order of regressors and of covariances is always preserved by design.

1.3 Asymptotic framework

Submodels estimated by least squares (1.11) have omitted variable bias. If γ and Υ are nonzero and fixed, the limiting distributions of the parameters restricted to zero (multiplied by $\sqrt{T}$) will have bias that tends to infinity. This is not useful for understanding submodel estimation in finite samples. In order to have a meaningful asymptotic bias-variance trade-off we use localization. This is a standard requirement for all FIC methods. Throughout the sequel, the data generating process is assumed to follow the full model specification and satisfy Assumption 1 below.

Assumption 1 $\gamma = \delta/\sqrt{T}$, $\Upsilon = \Delta/\sqrt{T}$, where δ and Δ are constant vectors. β and Φ are constant vectors.

As we show below, Assumption 1 ensures that submodel estimators are asymptotically normal with finite bias. Full model estimators are asymptotically unbiased. Submodel estimation also changes asymptotic covariance compared with the full model. This results in a nontrivial bias-variance trade-off in the limit.

To proceed, let $i, j, k, l \in \{1, \ldots, M\}$ and define:

$$\lambda_t = \begin{bmatrix} \underset{G\times 1}{\begin{bmatrix} (\epsilon_{it}\epsilon_{jt} - \omega_{ij})^{core} \\ \vdots \end{bmatrix}} \\ \underset{N\times 1}{\begin{bmatrix} (\epsilon_{kt}\epsilon_{lt} - \omega_{kl})^{aux} \\ \vdots \end{bmatrix}} \end{bmatrix}, \; J_T = \sum_{t=1}^{T} \lambda_t \tag{1.12}$$

where $\omega_{ij} = E[\epsilon_{it}\epsilon_{jt}]$ are the unique elements of $cov(\epsilon_t) = \Omega$. To derive limiting distributions, we make standard high-level assumptions which are outlined in the appendix. Consequences of these assumptions useful for understanding the results that follow are given in Condition 3 below. Assumption 2 is only slightly stronger than Condition 3.

Condition 3 As $T \to \infty$:

$$\hat{Q} \;=\; T^{-1}H'H \;\xrightarrow{p}\; Q \tag{1.13}$$

and

$$\begin{bmatrix} \frac{1}{\sqrt{T}}H'\epsilon \\ \frac{1}{\sqrt{T}}J_T \end{bmatrix} \;\xrightarrow{d}\; \begin{bmatrix} R \\ J \end{bmatrix} \;\sim\; N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

In our setup, we are concerned with estimating the unconditional covariance of the errors $cov\left(\epsilon_t\right) = \Omega$. This is a common approach in systematic trading contexts where small amounts of data are used for estimation with rolling windows. The rolling window setup captures changing dynamics of regression coefficients and volatility. However, it is well known that financial data has conditional heteroskedasticity. Many other models, for example GARCH specifications, have these dynamics. Similar results to those in the sequel could be derived with this type of specification. This would require replacing the convergence of $1/\sqrt{T}J_T$ in Condition 3 with corresponding asymptotics for GARCH parameters. From this, an FIC approach could be developed following similar steps. We leave a specific description of this to future research.

Under the above assumptions, the asymptotic distribution of submodel estimators is derived in Theorem A1 in the appendix. Similarly to Hjort and Claeskens (2003a) and Liu (2014), these distributions can be used to conduct model selection using a focus function. A focus function is a function of the estimated parameters. For example, in the portfolio choice case considered below, the focus is the portfolio weights. In general, the focus is the object of interest. Given a particular function, the limiting distribution of a plug-in focus estimator based on submodel m can be used as the basis of model selection. This is the idea of the FIC. This program is developed for the MR case below.

1.3.1 Focus parameter and its limiting distribution

Consider a N_{FIC}-dimensional function of the estimated model parameters (θ, Θ):

$$\begin{aligned} \mu \;&:\; R^{[(P+K)+(G+N)]} \to R^{N_{FIC}}, \\ \mu \;&=\; \mu\left(\theta, \Theta\right) = \mu\left(\beta, \gamma, \Phi, \Upsilon\right). \end{aligned} \tag{1.14}$$

This is the focus function, the object of interest in estimation. It is assumed to be twice continuously differentiable. Let

$$\hat{\mu}_m \;=\; \mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) \tag{1.15}$$

denote a submodel estimator of μ. Here, estimated parameters are plugged into μ. Elements which are set to zero in submodel m are set to zero in μ. This is suppressed in the notation. Let:

$$D_{\theta,\Theta} \;=\; \begin{bmatrix} D_\theta \\ \\ D_\Theta \end{bmatrix} \tag{1.16}$$

be a matrix of partial derivatives evaluated at the null points $\left(\beta', 0', \Phi', 0'\right)'$. The null points are the limit points of the localized parameters.

The following theorem derives the asymptotic distribution of $\hat{\mu}_m$. This distribution depends on both the asymptotic distribution of estimated parameters and which parameters are fixed at zero. It is therefore not an immediate result of Theorem A1. See the proof in the appendix for more details.

Theorem 1. *Suppose that Assumptions 1-2 hold. As $T \to \infty$, we have*

$$\begin{aligned}
\sqrt{T}\left(\mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) - \mu\left(\theta, \Theta\right)\right) \;&\xrightarrow{d}\; \Lambda_m \\
&= \begin{bmatrix} D_\theta \\ \\ D_\Theta \end{bmatrix}' \begin{bmatrix} C_m\delta + P_m R \\ \\ L_m\Delta + U_m J \end{bmatrix} \\
&\sim \begin{bmatrix} D_\theta \\ \\ D_\Theta \end{bmatrix}' N\left(\begin{bmatrix} C_m\delta \\ \\ L_m\Delta \end{bmatrix}, \begin{bmatrix} P_m\Sigma_{11}P_m & P_m\Sigma_{12}U_m \\ \\ U_m\Sigma_{21}P_m & U_m\Sigma_{22}U_m \end{bmatrix}\right)
\end{aligned} \tag{1.17}$$

where C_m, P_m, L_m and U_m are matrices related to submodel m:

$$\begin{aligned}
C_m \;&=\; \left(P_m Q - I\right) S_0 \\
P_m \;&=\; S_m Q_m^{-1} S_m' = S_m\left(S_m' Q S_m\right)^{-1} S_m' \\
L_m \;&=\; -F_0\left(I - \Gamma_m'\Gamma_m\right) \\
U_m \;&=\; F_m F_m'
\end{aligned}$$

Theorem 1 shows joint convergence of all submodels as any Λ_m can be expressed in terms of R and J. The size of the bias and variance in Λ_m are determined by $D_{\theta,\Theta}$ as well as δ, Δ, C_m, P_m, L_m and U_m. The parameters C_m, P_m, L_m and U_m depend on the submodel m and derivatives $D_{\theta,\Theta}$ depend on the focus μ. As a result, different submodels and focus functions result in a different mean and variance for Λ_m. The difference in distributions Λ_m formalizes bias-variance trade-off when estimating $\mu(\theta,\Theta)$ with different m. This gives the basis for a model selection criteria when estimating $\mu(\theta,\Theta)$. We now describe model selection and averaging using Λ_m.

1.4 FIC and Model Averaging

From Theorem 1, a loss metric for estimating $\mu(\theta,\Theta)$ using submodel m can be defined using Λ_m. We choose this loss metric to be the sum of asymptotic mean square errors (AMSE) of the N_{FIC} focus parameters:

$$SAMSE(\hat{\mu}_m) = \sum_{i=1}^{N_{FIC}} AMSE(\hat{\mu}_m^i) \tag{1.18}$$

where $\hat{\mu}_m^i$ is an element of the vector $\hat{\mu}_m$. The goal is to choose the submodel estimator $\hat{\mu}_m$ which minimizes SAMSE. This metric was chosen for simplicity. Another possible choice would be $AMSE\left(\sum_{i=1}^{N_{FIC}} \hat{\mu}_m^i\right)$. This would involve accounting for cross terms.

To express $AMSE(\hat{\mu}_m^i)$, we define the $(1 \times N_{FIC})$ vector S^i that selects $\hat{\mu}_m^i$ from $\hat{\mu}_m$. Consequently, from Theorem 1

$$\sqrt{T}\left(\mu^i\left(\hat{\theta}_{m,},\hat{\Theta}_{mm}\right)-\mu^i(\theta,\Theta)\right) = S^i\sqrt{T}\left(\mu\left(\hat{\theta}_m,\hat{\Theta}_{mm}\right)-\mu(\theta,\Theta)\right), \tag{1.19}$$

and

$$\begin{aligned} AMSE(\hat{\mu}_m^i) &= S^i\left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix}\begin{bmatrix} C_m\delta\delta'C_m' & C_m\delta\Delta'L_m' \\ L_m\Delta\delta'C_m' & L_m\Delta\Delta'L_m' \end{bmatrix}\begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}\right)S^{i'} \\ &+ S^i\left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix}\begin{bmatrix} P_m\Sigma_{11}P_m & P_m\Sigma_{12}U_m \\ U_m\Sigma_{21}P_m & U_m\Sigma_{22}U_m \end{bmatrix}\begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}\right)S^{i'} \end{aligned}. \tag{1.20}$$

FIC estimates SAMSE for different ms and selects the submodel with the smallest value.

Consistent estimation of SAMSE is not possible as there are no consistent estimators of δ and Δ. This is due to the local asymptotic framework. However, it is possible to construct asymptotically unbiased estimators. This is done using a plug-in method with (1.20). Lemma A2 in the appendix gives details.

1.4.1 Plug-in averaging estimator

Above we considered SAMSE of single submodels individually. However, the estimator obtained by weighting (averaging) across $\hat{\mu}_m$ for different submodels m can improve SAMSE compared with the best single submodel. This is because the set of convex combinations of $\hat{\mu}_m$ always contains each individual $\hat{\mu}_m$. A model averaging estimator can improve on individual $\hat{\mu}_m$ if the weights are chosen to minimize SAMSE. We now describe such an estimator rigorously.

Let

$$w = \left[\begin{array}{ccc} w_1 & \cdots & w_{\bar{M}} \end{array} \right]', \quad w_m \geq 0, \quad \sum_{m=1}^{\bar{M}} w_m = 1 \tag{1.21}$$

be a weight vector, where $\bar{M}$ is the number of submodels. The weights are assumed to be positive and sum to one. The model averaging estimator of μ is

$$\bar{\mu}(w) = \sum_{m=1}^{\bar{M}} w_m \hat{\mu}_m. \tag{1.22}$$

Theorem A3 in the supplemental material gives the asymptotic distribution of $\bar{\mu}(w)$. This distribution implies that SAMSE of the averaging estimator $\bar{\mu}(w)$ is

$$SAMSE(\bar{\mu}(w)) = \sum_{i=1}^{N_{FIC}} AMSE(\bar{\mu}^i) = w' \psi w \tag{1.23}$$

where $\bar{\mu}^i$ is ith element of $\bar{\mu}$. ψ is defined as:

$$\psi = \sum_{i=1}^{N_{FIC}} \psi^i \tag{1.24}$$

where ψ^i are $\bar{M} \times \bar{M}$ matrices corresponding to μ^i. The (m,l)th element of ψ^i is given by

$$\begin{aligned} \psi^i_{m,l} &= S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} C_m \delta \delta' C_l' & C_m \delta \Delta' L_l' \\ L_m \Delta \delta' C_l' & L_m \Delta \Delta' L_l' \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'} \\ &+ S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} P_m \Sigma_{11} P_l & P_m \Sigma_{12} U_l \\ U_m \Sigma_{21} P_l & U_m \Sigma_{22} U_l \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'} \end{aligned} \tag{1.25}$$

and therefore $\psi_{m,l} = \sum_{i=1}^{N_{FIC}} \psi^i_{m,l}$. Note that for $m = l$, $\psi^i_{m,m}$ is simply the AMSE for μ^i and submodel m. The optimal weight vector is the value that minimizes

$$\begin{aligned} w^0 &= \arg\min \; w' \psi w \\ &\text{s.t. } \textstyle\sum_{m=1}^{\bar{M}} w_m = 1, \quad w_m \geq 0 \end{aligned} \tag{1.26}$$

The estimator $\bar{\mu}(w^0)$ weakly dominates individual submodel estimators with respect to SAMSE.

There is no closed form solution to (1.26) when more than two submodels are present. Fortunately, the optimal weight vector can be found numerically via quadratic programming algorithms. The optimal weights are infeasible because they depend on unknown parameters δ, Δ, D_θ, D_Θ, C_m, P_m, Σ_{11}, Σ_{12} and Σ_{22}. A straightforward solution is to use sample counterparts as described in Lemma A2. This leads to data-dependent weights obtained by minimizing the sample analog of SAMSE (i.e. minimizing $w' \hat{\psi} w$ where $\hat{\psi}$ is the sample analogue of ψ). The plug-in estimator is defined as:

$$\begin{aligned} \bar{\mu}(\hat{w}) &= \textstyle\sum_{m=1}^{\bar{M}} \hat{w}_m \hat{\mu}_m \\ \hat{w} &= \arg\min \; w' \hat{\psi} w \\ &\text{s.t. } \textstyle\sum_{m=1}^{\bar{M}} w_m = 1, \; w_m \geq 0 \end{aligned} \tag{1.27}$$

Appendix B discusses the limiting distribution of $\bar{\mu}(\hat{w})$ and how it can be used for inference. The following section presents simulations and an application of this estimator to a portfolio choice problem.

1.5 Simulations

1.5.1 Optimal portfolio as a focus function

The general MR model selection and averaging theory developed in this paper was motivated by a portfolio choice problem considered in Garleanu and Pedersen (2013) (hereafter GP). GP assume asset returns follow the system:

$$\begin{aligned} r_{t+1} &= Bf_t + u_{t+1} \\ f_{t+1} &= (I - C) f_t + \varepsilon_{t+1} \end{aligned} \tag{1.28}$$

with

$$\begin{aligned} &E_t(u_{t+1}) = 0, \quad var_t(u_{t+1}) = \Omega_u, \quad cov_t(u_{t+1}, \varepsilon_{t+1}) = \Omega_{u\varepsilon} \\ &E_t(\varepsilon_{t+1}) = 0, \quad var_t(\varepsilon_{t+1}) = \Omega_\varepsilon \end{aligned} \tag{1.29}$$

where r_t is a $S_1 \times 1$ vector of asset returns and f_t is a $S_2 \times 1$ vector of driving factors. This is a restricted VAR(1) specification which fits the setup of Section 1.2. In this case, r_{t+1} takes the place of y_{t+1} in the current notation.

In this setup, x_t is the number of shares held of the corresponding assets in r_t. In addition, r_t is defined as price changes $r_t = p_t - p_{t-1}$. This is in contrast to much of the portfolio choice literature where x_t would be portfolio weights (i.e. proportions of wealth invested in each asset) and $r_t = (p_t - p_{t-1})/p_{t-1}$. This is needed so quadratic transaction costs can be used.

Given this specification, an investor decides which vector of asset shares x_t to hold this period. The investor chooses x_t to maximize a mean-variance portfolio problem which penalizes for risk and transaction costs (assumed to be quadratic i.e. $TC(\Delta x_t) = \frac{1}{2}\Delta x_t' \Lambda \Delta x_t$ where Λ is symmetric positive-definite). In particular, they maximize:

$$\max_{x_t} E_t\left[x_t' r_{t+1} - \frac{\gamma}{2} x_t' \Omega_u x_t - \frac{1}{2}\Delta x_t' \Lambda \Delta x_t\right] \tag{1.30}$$

where $\gamma > 0$ is a risk aversion coefficient. If transaction costs are proportional to the amount of risk ($\Lambda = \lambda\Omega_u$), GP show the solution to (1.30) is:

$$x_t = \left(1 - \tfrac{\gamma}{\gamma+\lambda}\right) x_{t-1} + \left(\tfrac{\gamma}{\gamma+\lambda}\right) Markowitz_t, \tag{1.31}$$

where

$$Markowitz_t = (\gamma \Omega_u)^{-1} B f_t. \tag{1.32}$$

The optimal policy x_t is characterized by trading partially toward the $Markowitz_t$ portfolio and holding some proportion of the current portfolio x_{t-1}. See GP Example 3 for more discussion.

The problem of estimating x_t can be naturally cast in the FIC framework developed in Section 2. In this case, the focus function μ is the vector of portfolio shares x_t and parameters θ and Θ correspond to B and Ω_u. Uncertainty about B corresponds to the situation where an investor wants to implement x_t, but is unsure which potential factors f_t to include. Uncertainty about Ω_u corresponds to the problem of estimating covariance between assets, possibly setting some terms to zero to avoid estimation error. The outlined regressions are usually estimated with relatively small sample sizes to deal with model instability. As a result, we have a standard bias-variance trade-off interpretation: adding more factors allows for more complicated dynamics, but small sample sizes mean this significantly increases estimator variance.

A number of studies have concluded that simple trading strategies can dominate more complicated ones. This is because of estimation error from trying to recover complicated asset dynamics from small samples (see DeMiguel, Nogales and Uppal (2016) and DeMiguel, Garlappi and Uppal (2009)). Ideally, an estimation method would explicitly account for increased variance in more complicated models and increased bias in simpler ones. The best model would be chosen by optimizing the trade-off between these competing errors. The trade-off should account for the final object of interest, the portfolio trading rule x_t. The focused information criterion and model averaging methods developed above were designed with this in mind. We now apply these methods to the portfolio choice situation in simulations and an empirical application.

1.5.2 Simulation specification and results

We illustrate the performance of our methods in simulations designed to mimic our application. A portfolio of five assets is considered. The portfolio size is restricted by computational cost. Solving the optimization problem for submodel weights in FIC model averaging is the main constraint. Our setup outlined below corresponds to 2048 submodels. Even this moderately

sized case represents a considerable computational burden when thousands of simulations are needed. Simulations took approximately a week on a 60 core computing cluster. Analyzing systems with a larger number of assets and factors is an important problem. Presently, we focus on a simple and realistic case to provide insight into the method's performance. Our simulations are described as follows:

1. Portfolio shares are estimated using different amounts of data (window size). The values $T = \{100, 150, 200, 250\}$ are considered. Small samples sizes such as these are standard in practice.

2. First, we simulate factors of sample size T. Each asset r_{it+1} is assumed to have two factors: f_{it}^s, $s = 1, 2$. These represent short-term (f_{it}^1) and long-term (f_{it}^2) information, short-term being less persistent. The factors are unique across assets and follow AR(1) processes:

$$f_{it+1}^s = (1 - c_i^s) f_{it}^s + \varepsilon_{it+1}^s.$$

   The errors follow $\varepsilon_{it+1}^s \sim N\left(0, var\left(\varepsilon_{it+1}^s\right)\right)$. Parameters c_i^1, c_i^2 and $var\left(\varepsilon_{it+1}^s\right)$ are assumed equal for all $i \in \{1, \ldots, 5\}$. The values of these parameters are reported in Table 1.1.

3. We use the simulated factors to obtain a random sample of returns that follow a specific version of (1.28):

$$r_{it+1} = b_{1i} f_{it}^1 + b_{2i} f_{it}^2 + u_{it+1}$$

   where $u_{it+1} \sim N\left(0, \Omega_{u,ii}\right)$ and the coefficients b_{1i}, b_{2i} are equal for all i. In addition, returns are assumed normalized by their unconditional standard deviation. This is why 0.9 is chosen for variance.

4. We consider 6 different scenarios for pairs b_{1i}, b_{2i} that are reported in Table 1.1. These correspond to different degrees of predictability of returns. All pairs are chosen so the predictive mean is significantly smaller than the error variance. The signal-to-noise ratio has an approximate maxima of 1/3 over all specifications. This is the empirically relevant situation.

5. We apply FIC and FIC model averaging (denoted FICav) to estimate x_t. This requires

estimating b_{1i}, b_{2i} and Ω_u. These parameters correspond to θ and Θ described in section 1.2.

6. To implement FIC, we must specify which variables are core and auxiliary. It is assumed all b_{1i}, b_{2i} and all covariance terms in Ω_u are auxiliary. For b_{1i}, b_{2i}, submodels are formed from all subsets of these values in each of the 5 return equations. For Ω_u, submodels are formed from two specifications: (1) Ω_u is unconstrained and (2) Ω_u is diagonal. More complicated submodels are not considered for computational reasons.

7. Estimation with FIC also requires derivatives of μ with respect to parameters. Using the closed-form expression for x_t (1.31) the required derivatives are:

$$\begin{aligned} \frac{\partial x_t}{\partial \Omega_{u,ij}} &= -\left(\frac{1}{\gamma+\lambda}\right)\left[\Omega_u^{-1}\frac{\partial \Omega_u}{\partial \Omega_{u,ij}}\Omega_u^{-1}\right]Bf_t \\ \frac{\partial x_t}{\partial b_{ji}} &= \left(\frac{1}{\gamma+\lambda}\right)\Omega_u^{-1}\frac{\partial B}{\partial b_{ji}}f_t \end{aligned} \tag{1.33}$$

   The factors f_t used in the focus are the final period values in the simulated data. Therefore, f_t is random across simulations. This better represents rolling window trading strategies implemented in applications.

8. Computing the focus function - portfolio shares (1.31) - requires hyper-parameters. Following GP, the absolute risk aversion is set to $\gamma = 10^{-9}$, which is considered reasonable for a large asset manager. The transaction cost parameter is $\lambda = 10^{-6}$, a number viewed as conservative.

9. A number of competitors are also computed for comparison: (1) full-model OLS (OLS); (2) AIC and BIC selected models (AIC and BIC); (3) to see if there is value added in optimizing submodel weights, final portfolios are estimated by equally weighting the portfolios from all submodels (EW); (4) portfolios are estimated with FIC but no model averaging (FIC).

Each simulation is repeated 500 times. The average SAMSE for the estimated portfolios relative to the optimal portfolios is our performance metric. The results are presented in Table 1.2.

The consistently best performers over all simulations are FICav and EW. Both methods have good results in all specifications. For smaller window sizes $T = 100, 150$, FICav clearly

In all scenarios	$c_i^1 = 0.5$	$c_i^2 = 0.1$
	$\Omega_{u,ii} = 0.9$	$\Omega_{u,ij} = 0.4,\ i \neq j$
	$var\left(\varepsilon_{it+1}^1\right) = 20$	$var\left(\varepsilon_{it+1}^2\right) = 5$
	$\Omega_{\varepsilon,ij} = 4,\ i \neq j$	$\Omega_{u\varepsilon,ij} = 0$
	b_{1i}	b_{2i}
Scenario 1	0.01	-0.01
Scenario 2	0.01	-0.005
Scenario 3	0.01	-0.001
Scenario 4	0	-0.01
Scenario 5	0	-0.005
Scenario 6	0	-0.001

Table 1.1 – Simulated Parameters

outperforms EW. For larger sizes $T = 200, 250$ the results are more mixed. EW tends to perform better when return coefficients are larger and predictability is higher. When predictability is lower, FICav is best. When predictability is lowest, FICav beats all methods by a significant margin.

These results suggest FICav does well when there is uncertainty about the relevance of factors given the amount of data available. With more data, there is less uncertainty with higher coefficients. When the coefficients are lower, the question is less resolved. In more complicated models with a larger number of covariates, more data will be needed to resolve this uncertainty.

EW is similar to equally weighted forecast combinations. This approach often does well in forecasting. See Timmermann (2006) for a survey and Claeskens, Magnus, Vasnev and Wang (2016) for a theoretical explanation. It is interesting we find strong performance for EW in our portfolio choice case.

In daily trading, a larger number of observations such as $T = 200, 250$ will usually be used. In these cases, we expect predictability to be low. With weekly or monthly data, predictability is higher, but our window size will often be lower. This would be in keeping with $T = 100, 150$. These two situations correspond to the cases where FICav performs best compared to competitors. Therefore, our simulations suggest that FICav is likely to perform well in many systematic trading contexts. This is what we find in our empirical application.

		FICav	EW	OLS	AIC	BIC	FIC
$T = 100$							
b_{1i}	b_{2i}						
0.01	−0.01	4.5302	6.3279	34.2575	17.4270	15.6629	24.5871
0.01	−0.005	3.6701	4.3433	36.9764	16.0701	14.7769	23.9436
0.01	−0.001	3.4690	4.7763	33.0678	17.6854	15.3047	21.0974
0	−0.01	1.0732	3.2116	34.2426	11.4254	10.4932	18.3754
0	−0.005	0.3010	2.8001	33.4551	12.8679	10.7594	18.1785
0	−0.001	0.0780	2.8566	36.3061	12.2589	10.7077	17.1469
$T = 150$							
b_{1i}	b_{2i}						
0.01	−0.01	4.3333	5.2271	21.4836	10.7591	9.9172	15.2787
0.01	−0.005	3.5131	3.4910	22.4661	11.8565	10.3562	16.3633
0.01	−0.001	3.5139	3.2722	21.0513	9.7811	8.7249	15.2467
0	−0.01	1.0393	2.2289	21.1134	8.8578	8.1197	13.1943
0	−0.005	0.2728	1.6981	21.0552	7.5322	7.0410	11.8398
0	−0.001	0.0530	1.7006	20.8802	8.4218	7.2766	12.5089
$T = 200$							
b_{1i}	b_{2i}						
0.01	−0.01	4.0787	4.0548	13.9658	8.8504	8.0542	12.5504
0.01	−0.005	3.6966	3.4619	12.9879	7.7561	6.9376	11.6288
0.01	−0.001	3.3242	2.6768	15.0313	7.9720	7.7051	11.1033
0	−0.01	1.0003	1.8802	12.6276	6.1816	5.4085	8.8025
0	−0.005	0.2666	1.3682	16.1418	7.1397	6.1273	9.4356
0	−0.001	0.0399	1.2818	15.3056	6.3143	5.0851	8.8245
$T = 250$							
b_{1i}	b_{2i}						
0.01	−0.01	4.2335	4.0230	11.3642	7.2625	6.8187	11.4659
0.01	−0.005	3.4736	2.9863	10.7567	6.9468	6.6358	8.7341
0.01	−0.001	3.3656	2.3416	13.1793	7.2718	6.5057	11.0019
0	−0.01	1.0147	1.6919	11.9479	6.6895	5.7949	9.1400
0	−0.005	0.2737	0.9149	11.9601	4.6237	3.9844	6.2438
0	−0.001	0.0289	0.8120	11.4313	4.2774	3.7218	7.1031

Table 1.2 – Simulated Average Portfolio Sum of Mean Squared Errors, 10^{10}

1.6 Commodity Futures Trading Application

In this section, FIC model averaging is applied to commodity futures trading. Fifteen highly liquid commodity futures contracts are considered: Soybean Oil (BO), Crude Oil (CL), Corn (C_), Gold (GC), Goldman Sachs Commodity Index (GI), Copper (HG), Orange Juice (JO),

Coffee (KC), Lumber (LB), Live Cattle (LC), Live Hogs (LH), Natural Gas (NG), Rough Rice (NR), Platinum (PL) and Silver (SI). Daily data on prices is taken from the Pinnacle continuous futures database. The dates considered are Jan 3, 2012 - July 1, 2016. All prices are in US dollars

All prices are first differenced then standardized by their sample standard deviation. This is standard in practice. Next, two predictive factors were constructed for each asset. These are the previous five day and twenty day averages of the standardized differenced prices. This setup is able to capture short term momentum in prices with a longer term reversal. These choices mimic our simulations and are a common approach in practice. Many other choices for factors are also possible, but are not explored here.

Twelve different portfolios were considered, each consisting of five assets drawn from the pool of fifteen described above. These twelve portfolios are presented in Table 1.3. Different portfolios are considered to explore the stability of the estimation approach and trading results. For each portfolio, rolling window estimation and trading was conducted over the sample period. The window size used was $T = 200$ observations. The estimated model was predictive regressions of the standardized price changes with the relevant factors. All hyper-parameter and submodel choices are the same as in simulations.

Transaction costs were assumed to have the quadratic form described in Section 1.5. As in GP, the covariance matrix used to compute transaction costs was the estimated residual covariance matrix for the entire observed sample. This was computed using non-normalized prices. In real trading, transaction costs would be observed, not computed as in this stylized setup.

Trading is as in GP. An agent starts with no shares in any of the assets. Each period, she trades to hold the estimated optimal portfolio. Profits are computed from price changes minus transaction costs. Sharpe ratios and total profits are then derived. These are our performance measures. Note that our trading is out-of-sample whereas GP considers in-sample results.

Total profits in millions and annualized Sharpe ratios from all twelve portfolios are presented in Table 1.3. Averages over all portfolios for each estimator are also reported. Estimation took approximately five days on a 60 core computing cluster. However, estimating a single portfolio position takes approximately 30 minutes. Implementing daily trading is clearly feasible.

In keeping with our simulations, the FICav and EW estimators are by far the best. All

other methods lose significant amounts of money and have negative Sharpe ratios on average. The FICav estimator has both a positive average profit (112.2) and average Sharpe ratio (0.2729). EW has a small average Sharpe ratio (0.0059) and significantly negative average profit (-575.3).

FICav takes significantly smaller positions than the other estimators. This can be seen from profits. Other estimators take large positions which yield high profits if they pay off. However, they frequently result in large losses. FICav takes smaller positions and its profits and losses are relatively small. To illustrate this, for Portfolio 1, Figures 1.7.1 and 1.7.2 show cumulative profits and Figures 1.7.3 and 1.7.4 portfolio positions for FICav and EW over the trading period. The positions of EW are much larger than FICav. Their respective largest values are approximately five million and one million shares. For this portfolio, EW results in larger profits but a much smaller Sharpe ratio. This is partially a consequence of larger volatility from larger positions and position changes.

FICav only loses money in three of the twelve portfolios and losses are small. In particular, they are small compared to profits from FICav in other portfolios. The three losses are -73.6, -16.8 and -1.1 whereas its three largest profits are 206.1, 286.8 and 457.5. EW has losses in half the portfolios and they can be large. The three largest losses from EW are -5,751.8, -5,472.5 and -1,197.8 whereas its three largest profits are 2,547.6, 1,540.4 and 991.8. On balance, FICav is the most robust procedure with the best results. It has the highest average Sharpe ratio by a significant margin. It is the only procedure that profits on average over all portfolios.

OLS, AIC, BIC and FIC all have very poor results on average. To explore the robustness of this finding, we also estimate trading strategies for OLS, AIC, BIC and EW with expanding windows. FIC and FICav procedures are designed for small samples sizes and therefore these estimates are not computed. The results for OLS, AIC and BIC are greatly improved. However, all still have significantly negative average profits and Sharpe ratios. EW improves as well. Its average Sharpe ratio (0.0782) is larger and average profits less negative (-124.1). These values are still significantly worse than FICav with rolling windows.

The results from expanding window estimation are not surprising. The additional data means more stable estimators, but they are not accurate enough to produce large profits or Sharpe ratios on average. The underlying asset dynamics in our sample are likely changing through time. If the regression coefficients and error covariance structure are changing, expanding window estimates will converge to an average of these parameters. Rolling windows

can potentially capture the changing structure and profit from it. Our results suggest FICav is able to do this.

We make no claim that our setup is optimal. However, the results show that FICav improves on standard estimation methods. Evaluation of different setups is beyond the scope of this paper.

1.7 Conclusion

In this paper, model averaging estimation of an MR using the FIC was considered. Regression coefficients and error covariance terms were assumed to be localized. This allowed for model selection and averaging focused on particular functions of the underlying parameters. The FIC approach chooses the submodel which optimizes bias-variance trade-off when estimating a focus function. Model averaging takes the FIC idea and averages over submodels in order to improve AMSE. This is a rigorous way of accounting for uncertainty about which aspects of the model should be included to minimize estimation error. The model averaging estimator was shown in simulations to outperform several competing methods. An application to futures data showed similar improvements.

		FICav	EW	OLS	AIC	BIC	FIC
Portfolio 1	SR	1.0342	0.5557	0.4402	0.0441	0.1115	0.7564
JO,GI,KC,LH,NG	P	206.1	621.3	3,754.8	186.7	501.1	2,135.8
Portfolio 2	SR	0.2936	0.9189	-0.2766	-0.2721	-0.2800	0.7296
GI,KC,LH,NG,NR	P	102.5	2,547.6	-2,630.9	-1,274.1	-1,268.8	3,232.8
Portfolio 3	SR	-0.1870	0.4816	-1.2718	-0.7025	-0.6611	-0.8151
KC,LH,NG,NR,PL	P	-73.6	991.8	-8,631.7	-2,662.2	-2,201.4	-3,065.7
Portfolio 4	SR	0.6677	0.2322	-2.7551	-3.7402	-3.9668	-3.6116
LH,NG,NR,PL,SI	P	204.6	1,540.4	-31,730.2	-22,222.0	-19,295.1	-20,856.1
Portfolio 5	SR	0.0704	-0.8377	-2.4394	-2.8137	-3.4324	-3.3537
NG,NR,PL,SI,GC	P	27.1	-5,472.5	-50,279.3	-34,459.9	-35,213.2	-36,227.2
Portfolio 6	SR	0.5541	-0.0584	-2.6477	-2.3153	-3.1962	-4.2987
NR,PL,SI,GC,C_	P	457.5	-203.9	-55,454.1	-39,242.2	-36,904.2	-34,437.7
Portfolio 7	SR	0.5217	-0.1010	-2.3199	-2.6520	-2.5933	-2.7356
PL,SI,GC,C_,BO	P	286.8	-429.2	-53,527.3	-38,604.2	-35,863.3	-33,880.3
Portfolio 8	SR	0.2552	-0.9061	-2.5898	-1.6496	-2.6866	-3.1087
SI,GC,C_,BO,HG	P	125.7	-5,751.8	-61,228.4	-20,815.2	-23,463.3	-38,867.9
Portfolio 9	SR	0.1273	-0.8251	-0.1509	-0.8352	-1.1462	-0.7846
GC,C_,BO,HG,CL	P	12.4	-1,197.9	-716.3	-1,683.0	-1,998.0	-2,529.9
Portfolio 10	SR	-0.3113	-0.0809	-0.1863	-0.6080	-1.0001	-0.5645
C_,BO,HG,CL,LC	P	-16.8	-99.2	-712.2	-984.9	-1,172.1	-1,923.0
Portfolio 11	SR	-0.0382	0.3402	-0.1308	-0.1076	-0.2884	0.7154
BO,HG,CL,LC,LB	P	-1.1	167.8	-204.1	-81.2	-223.9	710.6
Portfolio 12	SR	0.2873	0.3514	0.0030	-0.4915	-0.4410	0.4564
HG,CL,LC,LB,JO	P	14.8	382.2	5.1	-267.3	-311.2	516.4
Averages	SR	0.2729	0.0059	-1.1938	-1.3453	-1.6317	-1.3846
	P	112.2	-575.3	-21779.6	-13509.1	-13117.8	-13766.0

Table 1.3 – Commodity Futures Trading Results, Rolling Window; annualized Sharpe ratio (SR), cumulative profits in millions (P)

		EW	OLS	AIC	BIC
Portfolio 1	SR	0.3496	0.1283	0.1017	-0.6619
JO,GI,KC,LH,NG	P	515.0	400.2	188.5	-454.6
Portfolio 2	SR	0.1673	-0.4520	-0.7410	-0.7530
GI,KC,LH,NG,NR	P	121.1	-3,624.6	-2,740.1	-2,773.8
Portfolio 3	SR	0.3316	-0.5918	-0.3031	-0.3328
KC,LH,NG,NR,PL	P	336.4	-4,371.0	-333.1	-290.3
Portfolio 4	SR	0.2254	-0.7753	-0.6906	-0.8527
LH,NG,NR,PL,SI	P	584.8	-6,152.9	-5,351.5	-1,939.6
Portfolio 5	SR	-0.4323	-0.5088	0.0795	0.1605
NG,NR,PL,SI,GC	P	-1,340.7	-5,511.8	1,278.1	1,763.8
Portfolio 6	SR	-0.3085	-0.5881	0.2212	0.1535
NR,PL,SI,GC,C_	P	-1,235.6	-6,230.6	3,796.8	1,699.2
Portfolio 7	SR	0.3246	-0.1508	0.1340	0.1047
PL,SI,GC,C_,BO	P	279.1	-1,492.1	1,550.9	1,166.9
Portfolio 8	SR	-0.2931	-0.4872	-0.0431	-0.0682
SI,GC,C_,BO,HG	P	-530.4	-4,825.7	-480.6	-780.9
Portfolio 9	SR	-0.3962	0.0060	-0.0174	-0.0174
GC,C_,BO,HG,CL	P	-770.0	24.2	-64.0	-64.0
Portfolio 10	SR	0.2322	0.2950	-0.2853	-0.2005
C_,BO,HG,CL,LC	P	47.8	432.0	-102.1	-63.0
Portfolio 11	SR	0.1834	-0.3662	-0.0874	-0.2000
BO,HG,CL,LC,LB	P	52.0	-334.6	-29.1	-73.7
Portfolio 12	SR	0.5545	-0.5773	-0.6725	-0.7679
HG,CL,LC,LB,JO	P	451.9	-571.1	-218.4	-239.7
Averages	SR	0.0782	-0.3390	-0.1920	-0.2863
	P	-124.1	-2688.2	-208.7	-170.8

Table 1.4 – Commodity Futures Trading Results, Expanding Window; annualized Sharpe ratio (SR), cumulative profits in millions (P)

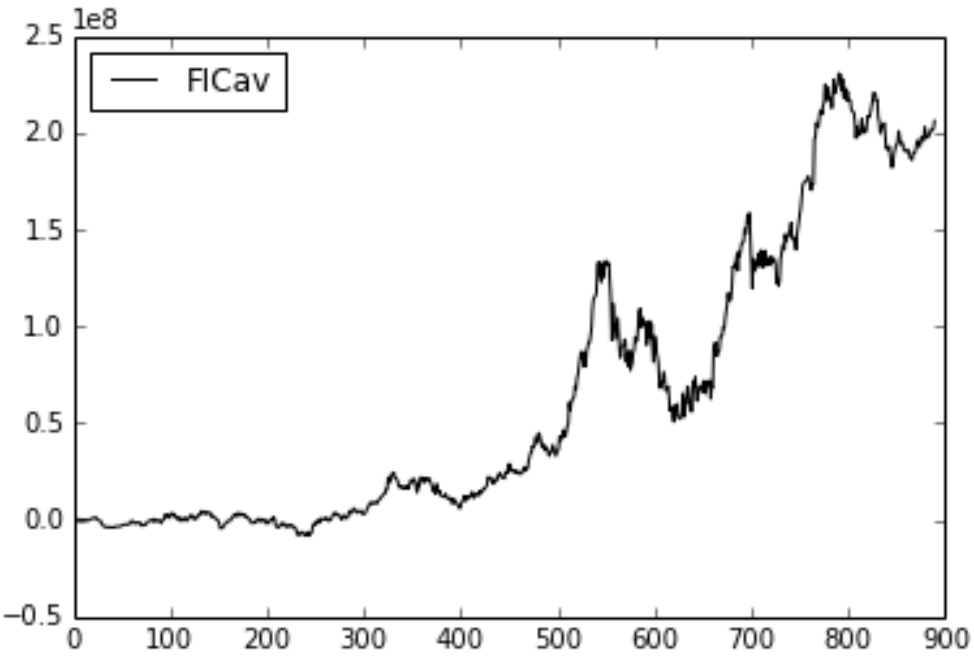

Figure 1.7.1 – FICav Account Curve, Portfolio 1

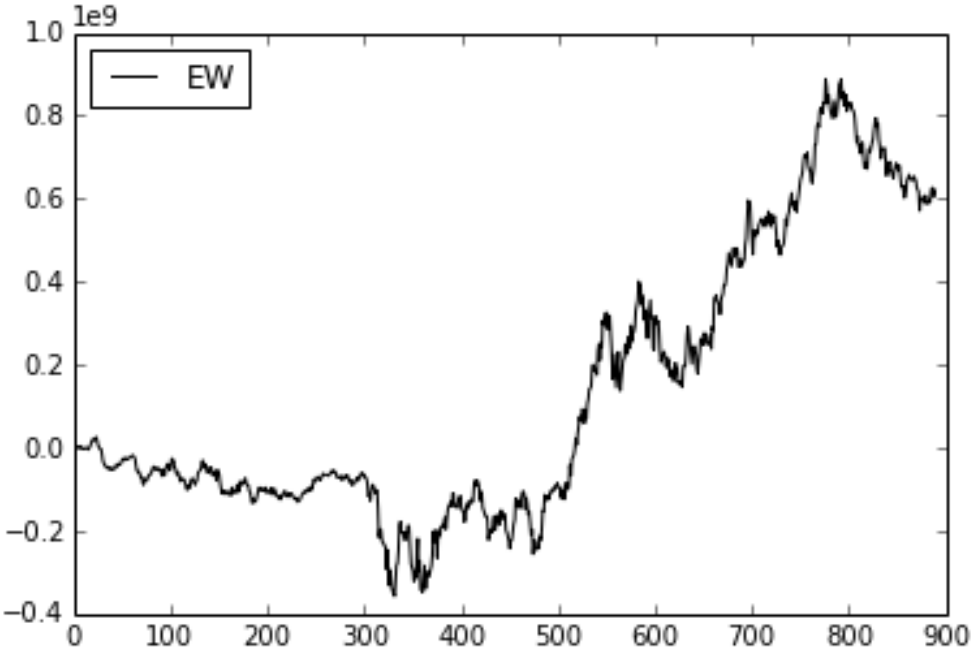

Figure 1.7.2 – EW Account Curve, Portfolio 1

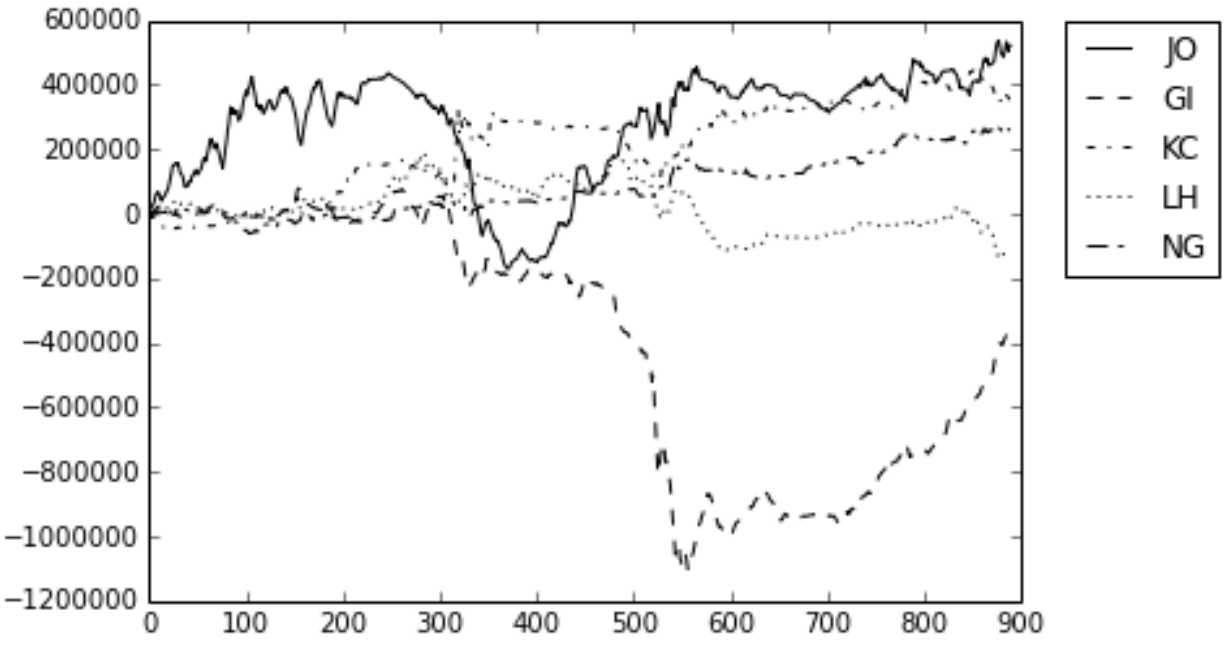

Figure 1.7.3 – FICav Portfolio Positions, Portfolio 1

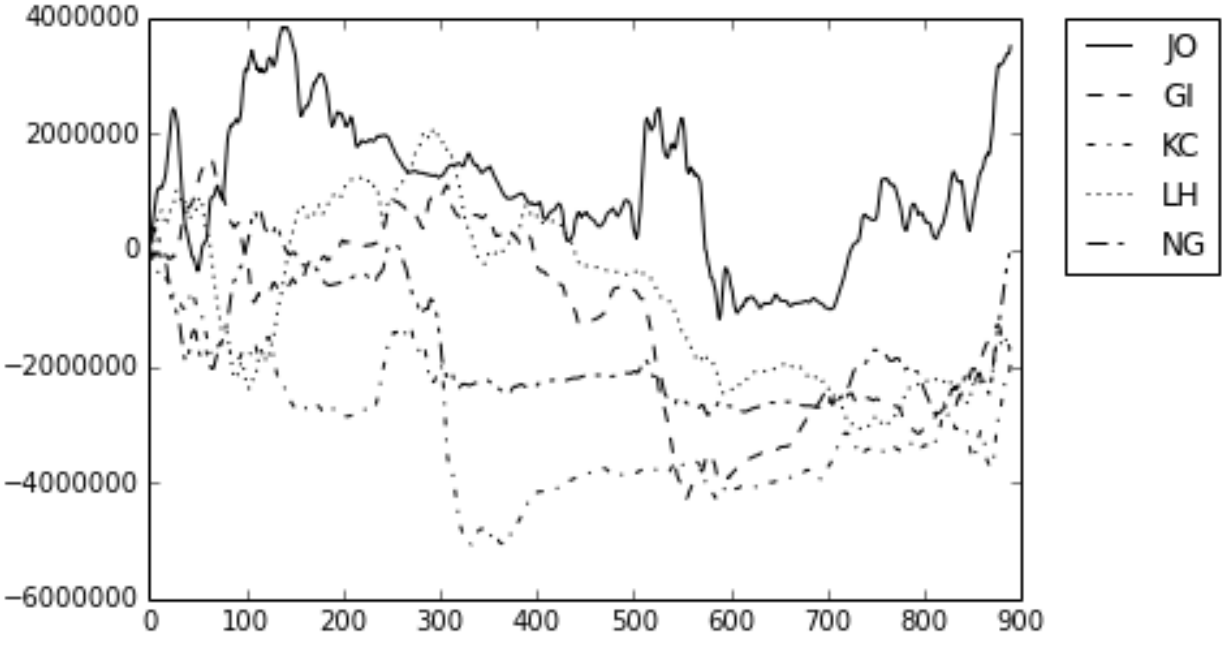

Figure 1.7.4 – EW Portfolio Positions, Portfolio 1

1.8 Appendix

Assumption 2. As $T \to \infty$, for $i, j \in 1, \ldots, M$:

$$T^{-1} H_i' H_j \xrightarrow{p} Q_{ij} = \begin{bmatrix} Q_{X_i' X_j} & Q_{X_i' Z_j} \\ Q_{Z_i' X_j} & Q_{Z_i' Z_j} \end{bmatrix}$$

and

$$\begin{bmatrix} \frac{1}{\sqrt{T}} H_i' \epsilon_j \\ \frac{1}{\sqrt{T}} J_T \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{T}} X_i' \epsilon_j \\ \frac{1}{\sqrt{T}} Z_i' \epsilon_j \\ \frac{1}{\sqrt{T}} J_T \end{bmatrix} \xrightarrow{d} \begin{bmatrix} R_{ij} \\ J \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{H_i' \epsilon_j \epsilon_j' H_i} & \Sigma_{H_i' \epsilon_j J} \\ \Sigma_{J \epsilon_j' H_i} & \Sigma_{22} \end{bmatrix} \right)$$

$$= N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{X_i' \epsilon_j \epsilon_j' X_i} & \Sigma_{X_i' \epsilon_j \epsilon_j' Z_i} & \Sigma_{X_i' \epsilon_j J} \\ \Sigma_{Z_i' \epsilon_j \epsilon_j' X_i} & \Sigma_{Z_i' \epsilon_j \epsilon_j' Z_i} & \Sigma_{Z_i' \epsilon_j J} \\ \Sigma_{J' \epsilon_j' X_i} & \Sigma_{J' \epsilon_j' Z_i} & \Sigma_{22} \end{bmatrix} \right)$$

where $\Sigma_{H_i' \epsilon_j \epsilon_j' H_i} = E\left(H_i' \epsilon_j \epsilon_j' H_i \right)$ and similarly for other objects in the covariance matrix.

Assumption 2 implies Condition 3 in the main text. Note that Assumption 2 is more general than that needed for the expressions of $T^{-1} H' H$ and

$$\begin{bmatrix} \frac{1}{\sqrt{T}} H' \epsilon \\ \frac{1}{\sqrt{T}} J_T \end{bmatrix}$$

which only contain elements of Q_{ii} and $\frac{1}{\sqrt{T}} H_i' \epsilon_i$. This generality is needed for deriving the limiting expressions for the covariances in Theorem A1.

Assumption 2 is weak. Convergence of $1/\sqrt{T} H_i' \epsilon_j$ could be shown with standard martingale methods in many situations (Hall and Heyde (1980)). Convergence of $1/\sqrt{T} J_T$ will follow from moment conditions on the error terms and a standard central limit theorem. Other primitive assumptions leading to these results are also possible.

1.9 Appendix A

The following theorem describes joint asymptotic distributions of the least squares estimators and error covariance estimators. The notation $\tilde{\gamma}_{m^c} = 0$ and $\tilde{\Upsilon}_{mm^c} = 0$ is used to represent parameters set to zero in submodel m. This is distinct from the true parameters γ_{m^c} and Υ_{mm^c} which are not zero. Although $\tilde{\gamma}_{m^c} = 0$ and $\tilde{\Upsilon}_{mm^c} = 0$ are not actually estimated from data, we use this notation to make this distinction.

Theorem (**A1**). *Suppose that Assumptions 1-2 hold. As $T \to \infty$, we have*

$$\begin{bmatrix} \sqrt{T}\left(\hat{\theta}_f - \theta\right) \\ \sqrt{T}\left(\hat{\Theta}_f - \Theta\right) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} Q^{-1}R \\ J \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q^{-1}\Sigma_{11}Q^{-1} & Q^{-1}\Sigma_{12} \\ \Sigma_{21}Q^{-1} & \Sigma_{22} \end{bmatrix}\right) \tag{1.34}$$

for the full model and

$$\begin{bmatrix} \sqrt{T}\left(\hat{\theta}_m - \theta_m\right) \\ \sqrt{T}\left(\hat{\Theta}_{mm} - \Theta_{mm}\right) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} A_m\delta + B_mR \\ F_m'J \end{bmatrix} \sim N\left(\begin{bmatrix} A_m\delta \\ 0 \end{bmatrix}, \begin{bmatrix} B_m\Sigma_{11}B_m' & B_m\Sigma_{12}F_m \\ F_m'\Sigma_{21}B_m' & F_m'\Sigma_{22}F_m \end{bmatrix}\right) \tag{1.35}$$

for a submodel m. R and J are normal random vectors as defined in Condition 3. A_m, B_m and F_m are matrices specific to a submodel m:

$$\begin{aligned} A_m &= Q_m^{-1}S_m'QS_0\left(I - \Pi_m'\Pi_m\right) \\ B_m &= Q_m^{-1}S_m' \end{aligned}$$

The coefficients for excluded variables $\tilde{\gamma}_{m^c}$ and covariances $\tilde{\Upsilon}_{mm^c}$ are set to zero and hence

$$\begin{bmatrix} \sqrt{T}\left(\tilde{\gamma}_{m^c} - \gamma_{m^c}\right) \\ \sqrt{T}\left(\tilde{\Upsilon}_{mm^c} - \Upsilon_{mm^c}\right) \end{bmatrix} = \begin{bmatrix} -\delta_{m^c} \\ -\Delta_{m^c} \end{bmatrix} \tag{1.36}$$

Theorem A1 implies that all estimators considered are consistent. In addition, the limiting distributions of all estimators are expressed in terms of the same normal random vectors R and J described in Condition 3. As can be seen above, the difference between distributions of competing submodel estimators is determined by Δ, δ, A_m, B_m and F_m. The estimator of the full model has no asymptotic bias. Submodel estimators have bias $A_m\delta$ in the estimated

parameters $\hat{\theta}_m$ resulting from excluding auxiliary variables (eq. 1.36). $\tilde{\gamma}_{m^c}$ and $\tilde{\Upsilon}_{mm^c}$ also result in bias as shown in (1.36). Interestingly, the term $\hat{\Theta}_{mm}$ results in no asymptotic bias. This is because the estimators $\hat{\Theta}_{mm}$ are not influenced by which parameters in Υ were set to zero. $\hat{\Theta}_{mm}$ is influenced by which parameters in γ were set to zero, but this influence disappears asymptotically because of the localization. The asymptotic covariance in our estimators differs across submodels as well. This can be seen in (1.34) and (1.35). Finally, note that $\tilde{\gamma}_{m^c}$ and $\tilde{\Upsilon}_{mm^c}$ have no asymptotic covariance because they are fixed at zero.

Define

$$\begin{aligned} \hat{\delta} &= \sqrt{T}\hat{\gamma}_f \\ \hat{\Delta} &= \sqrt{T}\hat{\Upsilon}_f \end{aligned} \tag{1.37}$$

where $\hat{\gamma}_f$ and $\hat{\Upsilon}_f$ are the estimators from the full model. Define $\hat{\lambda}_t$ as the empirical version of λ_t. This means true residuals are replaced by estimated residuals and true covariance terms by sample estimates. The following lemma describes construction of an asymptotically unbiased plug-in estimator of $SAMSE\left(\hat{\mu}_m\right)$ (i.e. FIC_m) by describing estimators of its components.

Lemma (A2). *The consistent estimators of D_θ, D_Θ, C_m, P_m, Σ_{11}, Σ_{12} and Σ_{22} required in the expression of AMSE are*

$$\begin{aligned} \hat{D}_\theta &= \frac{\partial \mu\left(\hat{\theta}_f, \hat{\Theta}_f\right)}{\partial \theta} \rightarrow D_\theta \\ \hat{D}_\Theta &= \frac{\partial \mu\left(\hat{\theta}_f, \hat{\Theta}_f\right)}{\partial \Theta} \rightarrow D_\Theta \\ \hat{P}_m &= S_m \hat{Q}_m^{-1} S_m' \rightarrow P_m \\ \hat{C}_m &= \left(\hat{P}_m \hat{Q} - I\right) S_0 \rightarrow C_m \end{aligned}$$

where $\hat{Q}$ is a consistent estimator of Q outlined in Assumption 2 and $\hat{Q}_m = S_m' \hat{Q} S_m \xrightarrow{p} Q_m$. With identically distributed and serially uncorrelated errors, consistent estimators of Σ_{11}, Σ_{12} and Σ_{22} are given by

$$\begin{aligned} \hat{\Sigma}_{11} &= \tfrac{1}{T}\textstyle\sum_{t=1}^T H_t \hat{\epsilon}_t \hat{\epsilon}_t' H_t' \rightarrow E\left(H_t \epsilon_t \epsilon_t' H_t'\right) \\ \hat{\Sigma}_{12} &= \tfrac{1}{T}\textstyle\sum_{t=1}^T H_t \hat{\epsilon}_t \hat{\lambda}_t' \rightarrow E\left(H_t \epsilon_t \lambda_t'\right) \\ \Sigma_{22} &= \tfrac{1}{T}\textstyle\sum_{t=1}^T \hat{\lambda}_t \hat{\lambda}_t' \rightarrow E\left(\lambda_t \lambda_t'\right) \end{aligned}$$

The limiting distributions of δ and Δ are

$$\begin{aligned}\hat{\delta} &\xrightarrow{d} R_\delta = \delta + S_0' Q^{-1} R \sim N\left(\delta, S_0' Q^{-1} \Sigma_{11} Q^{-1} S_0\right) \\ \hat{\Delta} &\xrightarrow{d} J_\Delta = \Delta + F_0' J \sim N\left(\Delta, F_0' \Sigma_{22} F_0\right)\end{aligned}$$

and hence unbiased estimators of $\delta\delta'$, $\delta\Delta'$ and $\Delta\Delta'$ required in the expression of AMSE are

$$\begin{aligned}\text{unbiased}\left(\delta\delta'\right) &= \hat{\delta}\hat{\delta}' - S_0' \hat{Q}^{-1} \hat{\Sigma}_{11} \hat{Q}^{-1} S_0 \\ \text{unbiased}\left(\delta\Delta'\right) &= \hat{\delta}\hat{\Delta}' - S_0' \hat{Q}^{-1} \hat{\Sigma}_{12} F_0 \\ \text{unbiased}\left(\Delta\Delta'\right) &= \hat{\Delta}\hat{\Delta}' - F_0' \hat{\Sigma}_{22} F_0\end{aligned}$$

Finally, U_m and L_m are non-random and known.

Simulations showed that the more complicated unbiased estimators in Lemma A2 don't produce better finite sample results than simply $\hat{\delta}\hat{\delta}'$, $\hat{\delta}\hat{\Delta}'$and $\hat{\Delta}\hat{\Delta}'$. These simpler estimators are used in our simulations and application in the main text.

The following theorem shows the asymptotic normality of the averaging estimator $\bar{\mu}(w)$ with fixed weights w.

Theorem (**A3**). *Suppose that Assumptions 1-2 hold. As $T \to \infty$, we have*

$$\sqrt{T}\left(\bar{\mu}(w) - \mu\right) \xrightarrow{d} N\left(D_\theta' C_w \delta + D_\Theta' L_w \Delta, V\right) \tag{1.38}$$

where $\bar{M}$ is the number of submodels,

$$\begin{aligned}C_w &= \textstyle\sum_{m=1}^{\bar{M}} w_m C_m \\ L_w &= \textstyle\sum_{m=1}^{\bar{M}} w_m L_m\end{aligned}$$

and

$$V = \textstyle\sum_{m=1}^{\bar{M}} w_m^2 Var(\Lambda_m) + \sum\sum_{m \neq l} w_m w_l Cov(\Lambda_m, \Lambda_l) + \sum\sum_{m \neq l} w_m w_l Cov(\Lambda_m, \Lambda_l)'$$

with

$$Cov\left(\Lambda_m, \Lambda_l\right) \;=\; \left[\begin{array}{cc} D_\theta^{'} & D_\Theta^{'} \end{array}\right] \left[\begin{array}{cc} P_m \Sigma_{11} P_l & P_m \Sigma_{12} U_l \\ U_m \Sigma_{21} P_l & U_m \Sigma_{22} U_l \end{array}\right] \left[\begin{array}{c} D_\theta \\ D_\Theta \end{array}\right]$$

The asymptotic bias and variance of the averaging estimator are $\left(D_\theta^{'} C_w \delta + D_\Theta^{'} L_w \Delta\right)$ and V respectively. These terms are weighted sums of bias and covariance terms derived for single submodel estimators $\hat{\mu}_m$ in Theorem 1 in the main text. Because changing the weights w changes the distribution in (1.38), it is possible that weighting across $\hat{\mu}_m$ improves estimator performance with respect to a given loss metric.

1.9.1 Proof of Theorem A1

We start by showing the limiting distribution of the parameters for full and restricted model.

Limiting distribution of the parameters in full and restricted model

By Assumption 2 and the application of the continuous mapping theorem, it follows that

$$\sqrt{T}\left(\hat{\theta}_f - \theta\right) \;=\; \left(\tfrac{1}{T} H^{'} H\right)^{-1} \left(\tfrac{1}{\sqrt{T}} H^{'} \epsilon\right) \;\xrightarrow{d}\; Q^{-1} R \;\sim\; N\left(0, Q^{-1} \Sigma_{11} Q^{-1}\right) \tag{1.39}$$

We next show the asymptotic distribution of the least squares estimator for each submodel. Note that, $H_m = \left[X, Z\Pi_m^{'}\right] = HS_m$ and $Z = HS_0$. By some algebra, it follows that

$$\begin{aligned}
\hat{\theta}_m \;&=\; \left(H_m^{'} H_m\right)^{-1} H_m^{'} y = \left(H_m^{'} H_m\right)^{-1} \left(H_m^{'} \left(X\beta + Z\gamma + \epsilon\right)\right) \\
&=\; \left(H_m^{'} H_m\right)^{-1} \left(H_m^{'} \left(X\beta + Z\left(I + \Pi_m^{'}\Pi_m - \Pi_m^{'}\Pi_m\right)\gamma + \epsilon\right)\right) \\
&=\; \left(H_m^{'} H_m\right)^{-1} \left(H_m^{'} \left(X\beta + Z\Pi_m^{'}\Pi_m\gamma + Z\left(I - \Pi_m^{'}\Pi_m\right)\gamma + \epsilon\right)\right) \\
&=\; \left(H_m^{'} H_m\right)^{-1} \left(H_m^{'} \left(\left[X, Z\Pi_m^{'}\right] \left[\begin{array}{c} \beta \\ \gamma_m \end{array}\right] + Z\left(I - \Pi_m^{'}\Pi_m\right)\gamma + \epsilon\right)\right) \\
&=\; \left(H_m^{'} H_m\right)^{-1} \left(H_m^{'} \left(H_m\theta_m + Z\left(I - \Pi_m^{'}\Pi_m\right)\gamma + \epsilon\right)\right) \\
&=\; \left(H_m^{'} H_m\right)^{-1} H_m^{'} H_m\theta_m + \left(H_m^{'} H_m\right)^{-1} H_m^{'} Z\left(I - \Pi_m^{'}\Pi_m\right)\gamma + \left(H_m^{'} H_m\right)^{-1} H_m^{'}\epsilon \\
&=\; \theta_m + \left(H_m^{'} H_m\right)^{-1} \left(HS_m\right)^{'} \left(HS_0\right)\left(I - \Pi_m^{'}\Pi_m\right)\gamma + \left(H_m^{'} H_m\right)^{-1} \left(HS_m\right)^{'}\epsilon \\
&=\; \theta_m + \left(H_m^{'} H_m\right)^{-1} S_m^{'} H^{'} HS_0\left(I - \Pi_m^{'}\Pi_m\right)\gamma + \left(H_m^{'} H_m\right)^{-1} S_m^{'} H^{'}\epsilon
\end{aligned} \tag{1.40}$$

Therefore by Assumptions 1-2 and the application of the continuous mapping theorem, we

have

$$\begin{aligned}\sqrt{T}\left(\hat{\theta}_m - \theta_m\right) &= \left(\tfrac{1}{T}H_m'H_m\right)^{-1}\left(\tfrac{1}{T}S_m'H'HS_0\right)\left(I - \Pi_m'\Pi_m\right)\sqrt{T}\gamma + \left(\tfrac{1}{T}H_m'H_m\right)^{-1}S_m'\left(\tfrac{1}{\sqrt{T}}H'\epsilon\right) \\ &\xrightarrow{d} Q_m^{-1}S_m'QS_0\left(I - \Pi_m'\Pi_m\right)\delta + Q_m^{-1}S_m'R \\ &= A_m\delta + B_mR \sim N\left(A_m\delta, Q_m^{-1}S_m'\Sigma_{11}S_mQ_m^{-1}\right)\end{aligned} \tag{1.41}$$

where $A_m = Q_m^{-1}S_m'QS_0\left(I - \Pi_m'\Pi_m\right)$ and $B_m = Q_m^{-1}S_m'$. This completes the proof.

Limiting distribution of the parameter and covariance in a full model

To describe covariance, we introduce some further notation. Let

$$\begin{aligned}\underset{T\times 1}{y_i} &= \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT}\end{bmatrix}, \quad \underset{T\times p_i}{X_i} = \begin{bmatrix} x_{i1}' \\ \vdots \\ x_{iT}'\end{bmatrix}, \quad \underset{T\times q_i}{Z_i} = \begin{bmatrix} z_{i1}' \\ \vdots \\ z_{iT}'\end{bmatrix} \\ \underset{T\times 1}{\epsilon_i} &= \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT}\end{bmatrix}, \quad \underset{(p_i+k_i)\times 1}{\theta_i} = \begin{bmatrix} \beta_i \\ \\ \gamma_i\end{bmatrix} \quad \underset{T\times(p_i+k_i)}{H_i} = [X_i\ Z_i]\end{aligned} \tag{1.42}$$

where y_i is $T \times 1$ vector of dependent variable i $(i = 1, \ldots, M)$ and ϵ_i is its $T \times 1$ vector of error terms; X_i and Z_i are $T \times p_i$ and $T \times k_i$ matrices of core and auxiliary regressors specific to the dependent variable y_i, with H_i matrix of $T \times (p_i + k_i)$ being their stacked version, and θ_i is the parameter vector containing core and auxiliary coefficients β_i and γ_i specific to y_i. Further define S_{im} to be an $(p_i + k_i) \times (p_i + k_{im})$ selection matrix that chooses the variables from the matrix of covariates H_i related to dependent variable y_i:

$$H_{im} = H_iS_{im}$$

Denote θ_{im} to be an $(p_i + k_{im}) \times 1$ parameter vector of a submodel m specific to the variables in equation for y_i such that we have $y_i = H_{im}\theta_{im} + \epsilon_{im}$.

Define the following:

$$\hat{\Theta}_f^* = \begin{bmatrix} \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_{it}\epsilon_{jt}\right)^{core} \\ \vdots \end{bmatrix} \\ G\times 1 \\ \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_{kt}\epsilon_{lt}\right)^{aux} \\ \vdots \end{bmatrix} \\ N\times 1 \end{bmatrix} = \frac{1}{T}\begin{bmatrix} \begin{bmatrix} \left(\epsilon_i'\epsilon_j\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\epsilon_k'\epsilon_l\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} \tag{1.43}$$

We first show that $\hat{\Theta}_f$ (sample error covariance matrix based on the OLS residuals) has the same asymptotic distribution as $\hat{\Theta}_f^*$. To see this, use the fact that

$$\begin{aligned} \epsilon_i &= y_i - H_i\theta_i \\ &= y_i - H_i\hat{\theta}_i + H_i\hat{\theta}_i - H_i\theta_i \\ &= \hat{\epsilon}_i + H_i\left(\hat{\theta}_i - \theta_i\right) \end{aligned} \tag{1.44}$$

to expand the expression for $\hat{\Theta}_f^*$ as follows:

$$\begin{aligned} \hat{\Theta}_f^* &= \frac{1}{T}\begin{bmatrix} \begin{bmatrix} \left(\epsilon_i'\epsilon_j\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\epsilon_k'\epsilon_l\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} = \frac{1}{T}\begin{bmatrix} \begin{bmatrix} \left(\hat{\epsilon}_i'\hat{\epsilon}_j\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\hat{\epsilon}_k'\hat{\epsilon}_l\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} + \frac{1}{T}\begin{bmatrix} \begin{bmatrix} \left(\hat{\epsilon}_i'H_j\left(\hat{\theta}_j - \theta_j\right)\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\hat{\epsilon}_k'H_l\left(\hat{\theta}_l - \theta_l\right)\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} \\ &+ \frac{1}{T}\begin{bmatrix} \begin{bmatrix} \left(\left(\hat{\theta}_i - \theta_i\right)' H_i'\hat{\epsilon}_j\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\left(\hat{\theta}_k - \theta_k\right)' H_k'\hat{\epsilon}_l\right)^{core} \\ \vdots \end{bmatrix} \end{bmatrix} + \frac{1}{T}\begin{bmatrix} \begin{bmatrix} \left(\left(\hat{\theta}_i - \theta_i\right)' H_i'H_j\left(\hat{\theta}_j - \theta_j\right)\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\left(\hat{\theta}_k - \theta_k\right)' H_k'H_l\left(\hat{\theta}_l - \theta_l\right)\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} \\ &= \hat{\Theta}_f + \frac{1}{T}R_{11} + \frac{1}{T}R_{12} + \frac{1}{T}R_3 \end{aligned} \tag{1.45}$$

Consequently

$$\sqrt{T}\left(\hat{\Theta}_f^* - \hat{\Theta}\right) = \frac{1}{\sqrt{T}}R_{11} + \frac{1}{\sqrt{T}}R_{12} + \frac{1}{\sqrt{T}}R_3 \tag{1.46}$$

Now a typical element of $\frac{1}{\sqrt{T}}R_{11}$ or $\frac{1}{\sqrt{T}}R_{12}$ is $\frac{1}{\sqrt{T}}\left(\hat{\theta}_i-\theta_i\right)' H_i'\hat{\epsilon}_j$ and $\frac{1}{\sqrt{T}}H_i'\hat{\epsilon}_j$ is an $(p_j+k_i)\times 1$ vector which can be written as follows

$$\begin{aligned}\tfrac{1}{\sqrt{T}}H_i'\hat{\epsilon}_j &= \tfrac{1}{\sqrt{T}}H_i'\left(y_j-H_j\hat{\theta}_j\right)\\ &= \tfrac{1}{\sqrt{T}}H_i'\left(\epsilon_j+H_j\theta_j-H_j\hat{\theta}_j\right)\\ &= \tfrac{1}{\sqrt{T}}H_i'\left(\epsilon_j-H_j\left(\hat{\theta}_j-\theta_j\right)\right)\\ &= \tfrac{1}{\sqrt{T}}H_i'\epsilon_j-\tfrac{H_i'H_j}{T}\sqrt{T}\left(\hat{\theta}_j-\theta_j\right)\end{aligned} \tag{1.47}$$

From Assumption 2 we have $\frac{1}{\sqrt{T}}H_i'\epsilon_j \xrightarrow{d} N\left(0,\Sigma_{H_i'\epsilon_j\epsilon_j'H_i}\right)$, $\frac{H_i'H_j}{T}\xrightarrow{p}Q_{ij}$, and as shown above, $\sqrt{T}\left(\hat{\theta}-\theta\right)\xrightarrow{d}N\left(0,Q^{-1}\Sigma_{11}Q^{-1}\right)$ and thus $\sqrt{T}\left(\hat{\theta}_j-\theta_j\right)\xrightarrow{d}N\left(0,\left(Q^{-1}\Sigma_{11}Q^{-1}\right)_{\theta_j}\right)$ where $\left(Q^{-1}\Sigma_{11}Q^{-1}\right)_{\theta_i}$ is the covariance matrix related to θ_j. Consequently, we have

$$\tfrac{1}{\sqrt{T}}H_i'\hat{\epsilon}_j \xrightarrow{d} N\left(0,\Sigma_{H_i'\epsilon_j\epsilon_j'H_i}\right)+Q_{ij}N\left(0,\left(Q^{-1}\Sigma_{11}Q^{-1}\right)_{\theta_j}\right) \tag{1.48}$$

Thus $\frac{1}{\sqrt{T}}H_i'\hat{\epsilon}_j$ converges to the sum of normal distributions and thus (by proposition 7.3, Hamilton, 1994):

$$\tfrac{1}{\sqrt{T}}\left(\hat{\theta}_i-\theta_i\right)' H_i'\hat{\epsilon}_j \xrightarrow{p} 0 \tag{1.49}$$

since $\left(\hat{\theta}_i-\theta_i\right)\xrightarrow{p}0$. Thus we have $\frac{1}{\sqrt{T}}R_{11}\xrightarrow{p}0$ and $\frac{1}{\sqrt{T}}R_{12}\xrightarrow{p}0$.

Next, inspecting $\frac{1}{\sqrt{T}}R_3$, one can see that its typical element converges to zero in probability:

$$\tfrac{1}{\sqrt{T}}\left(\hat{\theta}_i-\theta_i\right)' H_i'H_j\left(\hat{\theta}_j-\theta_j\right) = \left(\hat{\theta}_i-\theta_i\right)'\tfrac{H_i'H_j}{T}\sqrt{T}\left(\hat{\theta}_j-\theta_j\right)\xrightarrow{p}0 \tag{1.50}$$

which holds by proposition 7.3, Hamilton, 1994 since $\frac{H_i'H_j}{T}\xrightarrow{p}Q_{ij}$, $\sqrt{T}\left(\hat{\theta}_j-\theta_j\right)\xrightarrow{d}N\left(0,\left(Q^{-1}\Sigma_{11}Q^{-1}\right)_{\theta_j}\right)$ and $\left(\hat{\theta}_i-\theta_i\right)\xrightarrow{p}0$. This implies that $\frac{1}{\sqrt{T}}R_3\xrightarrow{p}0$.

The above results imply that

$$\sqrt{T}\left(\hat{\Theta}_f^*-\hat{\Theta}_f\right)\xrightarrow{p}0 \tag{1.51}$$

meaning that

$$\sqrt{T}\left(\hat{\Theta}_f^*-\Theta\right)\xrightarrow{p}\sqrt{T}\left(\hat{\Theta}_f-\Theta\right) \tag{1.52}$$

It also follows that

$$\begin{aligned}\sqrt{T}\left(\hat{\Phi}^* - \Phi\right) &\xrightarrow{p} \sqrt{T}\left(\hat{\Phi} - \Phi\right) \\ \sqrt{T}\left(\hat{\Upsilon}^* - \Upsilon\right) &\xrightarrow{p} \sqrt{T}\left(\hat{\Upsilon} - \Upsilon\right)\end{aligned} \tag{1.53}$$

This implies:

$$\begin{bmatrix} \sqrt{T}\left(\hat{\theta}_f - \theta\right) \\ \sqrt{T}\left(\hat{\Theta}_f - \Theta\right) \end{bmatrix} = \begin{bmatrix} Q^{-1}\left(\frac{1}{\sqrt{T}}H'\epsilon\right) \\ \frac{1}{\sqrt{T}}J_T \end{bmatrix} \xrightarrow{d} \begin{bmatrix} Q^{-1}R \\ J \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q^{-1}\Sigma_{11}Q^{-1} & Q^{-1}\Sigma_{12} \\ \Sigma_{21}Q^{-1} & \Sigma_{22} \end{bmatrix}\right) \tag{1.54}$$

Limiting distribution of the parameter and covariance in a submodel

We next show the joint asymptotic distribution of the least squares parameters and error variances for each submodel. Rearranging the expression for the estimated errors $\hat{\epsilon}_{im}$:

$$\begin{aligned}\hat{\epsilon}_{im} &= y_i - H_{im}\hat{\theta}_{im} \\ &= \left(y_i - H_i\hat{\theta}_i\right) + \left(H_i\hat{\theta}_i - H_{im}\hat{\theta}_{im}\right) \\ &= \left(y_i - H_i\hat{\theta}_i\right) + H_i\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right) \\ &= \hat{\epsilon}_i + H_i\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right)\end{aligned} \tag{1.55}$$

Using this, we can express the covariance matrix for a submodel m:

$$
\begin{aligned}
&\hat{\Theta}_{mm} \\
= \quad & F_m^{'} \hat{\Theta}_m \\
= \quad & F_m^{'} \tfrac{1}{T} \begin{bmatrix} \begin{bmatrix} \left(\hat{\epsilon}_{im}^{'} \hat{\epsilon}_{jm}\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\hat{\epsilon}_{km}^{'} \hat{\epsilon}_{lm}\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} = F_m^{'} \tfrac{1}{T} \begin{bmatrix} \begin{bmatrix} \left(\hat{\epsilon}_i^{'} \hat{\epsilon}_j\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\hat{\epsilon}_k^{'} \hat{\epsilon}_l\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} + F_m^{'} \tfrac{1}{T} \begin{bmatrix} \begin{bmatrix} \left(\hat{\epsilon}_i^{'} H_j \left(\hat{\theta}_j - S_{jm} \hat{\theta}_{jm}\right)\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\hat{\epsilon}_k^{'} H_l \left(\hat{\theta}_l - S_{lm} \hat{\theta}_{lm}\right)\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} \\
+ \quad & F_m^{'} \tfrac{1}{T} \begin{bmatrix} \begin{bmatrix} \left(\left(\hat{\theta}_i - S_{im} \hat{\theta}_{im}\right)^{'} H_i^{'} \hat{\epsilon}_j^{'}\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\left(\hat{\theta}_k - S_{km} \hat{\theta}_{km}\right)^{'} H_k^{'} \hat{\epsilon}_l^{'}\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} + F_m^{'} \tfrac{1}{T} \begin{bmatrix} \begin{bmatrix} \left(\left(\hat{\theta}_i - S_{im} \hat{\theta}_{im}\right)^{'} H_i^{'} H_j \left(\hat{\theta}_j - S_{jm} \hat{\theta}_{jm}\right)\right)^{core} \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \left(\left(\hat{\theta}_k - S_{km} \hat{\theta}_{km}\right)^{'} H_k^{'} H_l \left(\hat{\theta}_l - S_{lm} \hat{\theta}_{lm}\right)\right)^{aux} \\ \vdots \end{bmatrix} \end{bmatrix} \\
= \quad & F_m^{'} \hat{\Theta}_f + F_m^{'} \tfrac{1}{T} R_{11m} + F_m^{'} \tfrac{1}{T} R_{12m}^{'} + F_m^{'} \tfrac{1}{T} R_{3m}
\end{aligned}
\tag{1.56}
$$

Consequently

$$
\sqrt{T}\left(\hat{\Theta}_{mm} - \Theta_{mm}\right) = F_m^{'} \sqrt{T}\left(\hat{\Theta}_f - \Theta\right) + \sqrt{T}\left(F_m^{'} \tfrac{1}{T} R_{11m} + F_m^{'} \tfrac{1}{T} R_{12m}^{'} + F_m^{'} \tfrac{1}{T} R_{3m}\right) \tag{1.57}
$$

because $F_m^{'}\Theta = F_m^{'}\Theta_m = \Theta_{mm}$. This holds because $\Theta = \Theta_m$. Now a typical element of $\frac{1}{\sqrt{T}} R_{11m}$ or $\frac{1}{\sqrt{T}} R_{12m}$ is $\frac{1}{\sqrt{T}}\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right)^{'} H_i^{'} \hat{\epsilon}_j^{'}$ and as shown above for the case of the full model $\frac{1}{\sqrt{T}} H_i^{'} \hat{\epsilon}_j$ converges to the sum of normal distributions. We know $\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right) \xrightarrow{p} 0$ by the asymptotic distributions of $\hat{\theta}$ and $\hat{\theta}_m$ derived above which show they both converge to zero. This implies that

$$
\tfrac{1}{\sqrt{T}}\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right)^{'} H_i^{'} \hat{\epsilon}_j^{'} \xrightarrow{p} 0 \tag{1.58}
$$

and hence $\frac{1}{\sqrt{T}} R_{11m} \xrightarrow{p} 0$ and $\frac{1}{\sqrt{T}} R_{12m} \xrightarrow{p} 0$.

To proceed, we must show convergence in distribution of $\sqrt{T}\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right)$. The term $\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right)$ can be obtained from $\left(\hat{\theta} - S_m \hat{\theta}_m\right)$ by selecting only those coefficients that are related to the dependent variable y_i. Thus, denoting S^i is as a selection matrix that takes θ

and selects only the coefficients related to the variable y_i we have:

$$S^i\left(\hat{\theta} - S_m\hat{\theta}_m\right) = \hat{\theta}_i - S_{im}\hat{\theta}_{im} \tag{1.59}$$

By expanding and using the above asymptotic convergence results:

$$\begin{aligned}\sqrt{T}\left(S_m\hat{\theta}_m - \hat{\theta_f}\right) &= S_m\sqrt{T}\left(\hat{\theta}_m - \theta_m\right) + \sqrt{T}\left(S_m\theta_m - \theta\right) - \sqrt{T}\left(\hat{\theta}_f - \theta\right) \\ &\xrightarrow{d} S_m\left(A_m\delta + B_mR\right) + \sqrt{T}\left(S_m\theta_m - \theta\right) - Q^{-1}R\end{aligned} \tag{1.60}$$

Some matrix manipulation shows that $\sqrt{T}\left(S_m\theta_m - \theta\right) = S_0\left(\Pi_m'\Pi_m - I\right)\delta$. Therefore, $\sqrt{T}\left(S_m\hat{\theta}_m - \hat{\theta_f}\right)$ converges in distribution.

Next, inspecting R_{3m}, one can see their typical element is of the form

$$\tfrac{1}{\sqrt{T}}\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right)' H_i'H_j\left(\hat{\theta}_j - S_{jm}\hat{\theta}_{jm}\right) = \left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right)' \tfrac{H_i'H_j}{T}\sqrt{T}\left(\hat{\theta}_j - S_{jm}\hat{\theta}_{jm}\right) \xrightarrow{p} 0 \tag{1.61}$$

with the convergence in probability happening since $\sqrt{T}\left(\hat{\theta}_j - S_{jm}\hat{\theta}_{jm}\right)$ converges in distribution, $\frac{H_i'H_j}{T} \xrightarrow{p} Q_{ij}$ and $\left(\hat{\theta}_i - S_{im}\hat{\theta}_{im}\right) \xrightarrow{p} 0$ and applying the Proposition 7.3 from Hamilton (1994) yields the result. Thus finally we have

$$\begin{aligned}\sqrt{T}\left(\hat{\Theta}_{mm} - \Theta_{mm}\right) &= F_m'\sqrt{T}\left(\hat{\Theta}_f - \Theta\right) + \left(F_m'\tfrac{1}{\sqrt{T}}R_{11m} + F_m'\tfrac{1}{\sqrt{T}}R_{12m}' + F_m'\tfrac{1}{\sqrt{T}}R_{3m}\right) \\ &\to F_m'J\end{aligned} \tag{1.62}$$

Using this together with the result from the distribution of the parameters from a submodel and Assumption 2:

$$\begin{bmatrix}\sqrt{T}\left(\hat{\theta}_m - \theta_m\right) \\ \sqrt{T}\left(\hat{\Theta}_{mm} - \Theta_{mm}\right)\end{bmatrix} = \begin{bmatrix}A_m\delta + B_m\left(\frac{1}{\sqrt{T}}H'\epsilon\right) \\ F_m'\left(\frac{1}{\sqrt{T}}J_T\right)\end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix}A_m\delta \\ 0\end{bmatrix}, \begin{bmatrix}B_m\Sigma_{11}B_m' & B_m\Sigma_{12}F_m \\ F_m'\Sigma_{21}B_m' & F_m'\Sigma_{22}F_m\end{bmatrix}\right) \tag{1.63}$$

This completes the proof of Theorem A1.

1.9.2 Proof of Theorem 1

Define

$$\begin{aligned}
\gamma_{m^c} &= \{\gamma : \gamma_j \notin \gamma_m, \text{ for } \gamma_j \text{ in } \gamma\} \\
\Upsilon_{mm^c} &= \{\omega_{ij} : \omega_{ij} \notin \Upsilon_{mm}, \text{ for } \omega_{ij} \text{ in } \Upsilon\} \\
D_{\gamma_{m^c}} &= \frac{\partial \mu}{\partial \gamma_{m^c}} \\
D_{\Upsilon_{mm^c}} &= \frac{\partial \mu}{\partial \Upsilon_{mm^c}}
\end{aligned} \tag{1.64}$$

That is, γ_{m^c} and Υ_{mm^c} are vectors of parameters γ_j and ω_{ij} that are not included in submodel m. Hence, again since $\Phi_m = \Phi$ for core covariances we can write $\mu(\theta, \Theta)$ as

$$\mu(\beta, \gamma_m, \gamma_{m^c}, \Phi_m, \Upsilon_{mm}, \Upsilon_{mm^c}) = \mu(\beta, \gamma_m, \gamma_{m^c}, \Phi, \Upsilon_{mm}, \Upsilon_{mm^c}) \tag{1.65}$$

and

$$\mu(\theta_m, \Theta_{mm}) = \mu(\beta, \gamma_m, 0, \Phi, \Upsilon_{mm}, 0) \tag{1.66}$$

Note that $\gamma = O\left(T^{-1/2}\right)$ and $\Upsilon = O\left(T^{-1/2}\right)$ by Assumption 1. Then by a standard Taylor series expansion of $\mu(\theta, \Theta)$ about $\gamma_{m^c} = 0$ and $\Upsilon_{mm^c} = 0$, it follows that

$$\begin{aligned}
&\mu(\beta, \gamma_m, \gamma_{m^c}, \Phi, \Upsilon_{mm}, \Upsilon_{mm^c}) \\
&= \mu(\beta, \gamma_m, 0, \Phi, \Upsilon_{mm}, 0) + D'_{\gamma_{m^c}} \gamma_{m^c} + D'_{\Upsilon_{mm^c}} \Upsilon_{mm^c} + O\left(T^{-1}\right) \\
&= \mu(\beta, \gamma_m, 0, \Phi, \Upsilon_{mm}, 0) + D'_{\gamma} \left(I - \Pi'_m \Pi_m\right) \gamma + D'_{\Upsilon} \left(I - \Gamma'_m \Gamma_m\right) \Upsilon + O\left(T^{-1}\right)
\end{aligned} \tag{1.67}$$

where we used the fact that $\left(I - \Pi'_m \Pi_m\right) \gamma$ and $\left(I - \Gamma'_m \Gamma_m\right) \Upsilon$ include parameters and error covariances that are excluded from a submodels and set those which are included to zero. Thus

$$\mu(\theta, \Theta) - \mu(\theta_m, \Theta_{mm}) = D'_{\gamma} \left(I - \Pi'_m \Pi_m\right) \gamma + D'_{\Upsilon} \left(I - \Gamma'_m \Gamma_m\right) \Upsilon + O\left(T^{-1}\right) \tag{1.68}$$

By Assumptions 1-2 and the application of the delta method, we have

$$\begin{aligned}
&\sqrt{T}\left(\mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) - \mu\left(\theta, \Theta\right)\right) \\
&= \sqrt{T}\left(\mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) - \mu\left(\theta_m, \Theta_{mm}\right)\right) - \sqrt{T}\left(\mu\left(\theta, \Theta\right) - \mu\left(\theta_m, \Theta_{mm}\right)\right) \\
&\xrightarrow{d} D'_{\theta_m, \Theta_{mm}} \begin{bmatrix} A_m\delta + B_m R \\ F'_m J \end{bmatrix} - D'_\gamma\left(I - \Pi'_m\Pi_m\right)\delta - D'_\Upsilon\left(I - \Gamma'_m\Gamma_m\right)\Delta \\
&= \begin{bmatrix} D'_{\theta_m}, & D'_{\Theta_{mm}} \end{bmatrix} \begin{bmatrix} A_m\delta + B_m R \\ F'_m J \end{bmatrix} - D'_\gamma\left(I - \Pi'_m\Pi_m\right)\delta - D'_\Upsilon\left(I - \Gamma'_m\Gamma_m\right)\Delta \\
&= D'_{\theta_m}\left(A_m\delta + B_m R\right) - D'_\gamma\left(I - \Pi'_m\Pi_m\right)\delta + D'_{\Theta_{mm}}\left(F'_m J\right) - D'_\Upsilon\left(I - \Gamma'_m\Gamma_m\right)\Delta
\end{aligned} \tag{1.69}$$

Now let $P_m = S_m\left(S'_m Q S_m\right)^{-1} S'_m$. Focusing on the first two terms and using the facts that $D'_{\theta_m} = D'_\theta S_m$, $D'_\gamma = D'_\theta S_0$:

$$\begin{aligned}
&D'_{\theta_m}\left(A_m\delta + B_m R\right) - D'_\gamma\left(I - \Pi'_m\Pi_m\right)\delta \\
&= D'_{\theta_m} A_m\delta - D'_\gamma\left(I - \Pi'_m\Pi_m\right)\delta + D'_{\theta_m} B_m R \\
&= \left(D'_\theta S_m Q_m^{-1} S'_m Q S_0 - D'_\theta S_0\right)\left(I - \Pi'_m\Pi_m\right)\delta + D'_\theta S_m Q_m^{-1} S'_m R \\
&= \left(D'_\theta S_m\left(S'_m Q S_m\right)^{-1} S'_m Q S_0 - D'_\theta S_0\right)\delta - D'_\theta\left(S_m\left(S'_m Q S_m\right)^{-1} S'_m Q S_0 - S_0\right)\Pi'_m\Pi_m\delta + D'_\theta P_m R
\end{aligned} \tag{1.70}$$

Now inspecting the second term in the above equation and making use of the fact that

$$S_0\Pi'_m = S_m \begin{bmatrix} 0_{P\times K_m} \\ I_{K_m} \end{bmatrix}$$

we have

$$\begin{aligned}
&-S_m Q_m^{-1} S'_m Q S_0\Pi'_m\Pi_m + S_0\Pi'_m\Pi_m \\
&= -S_m Q_m^{-1} S'_m Q S_m \begin{bmatrix} 0_{P\times K_m} \\ I_{K_m} \end{bmatrix}\Pi_m + S_m \begin{bmatrix} 0_{P\times K_m} \\ I_{K_m} \end{bmatrix}\Pi_m \\
&= -S_m\left(S'_m Q S_m\right)^{-1} S'_m Q S_m \begin{bmatrix} 0_{P\times K_m} \\ I_{K_m} \end{bmatrix}\Pi_m + S_m \begin{bmatrix} 0_{P\times K_m} \\ I_{K_m} \end{bmatrix}\Pi_m \\
&= 0
\end{aligned} \tag{1.71}$$

Therefore, the second term disappears. Hence, the above expression becomes

$$\begin{aligned}
&D_{\theta_m}^{'}\left(A_m\delta + B_m R\right) - D_{\gamma}^{'}\left(I - \Pi_m^{'}\Pi_m\right)\delta \\
&= \left(D_{\theta}^{'}S_m\left(S_m^{'}QS_m\right)^{-1}S_m^{'}QS_0 - D_{\theta}^{'}S_0\right)\delta + D_{\theta}^{'}P_m R \\
&= D_{\theta}^{'}\left(S_m\left(S_m^{'}QS_m\right)^{-1}S_m^{'}QS_0 - S_0\right)\delta + D_{\theta}^{'}P_m R \\
&= D_{\theta}^{'}\left(P_m Q - I\right)S_0\delta + D_{\theta}^{'}P_m R \\
&= D_{\theta}^{'}C_m\delta + D_{\theta}^{'}P_m R
\end{aligned} \tag{1.72}$$

Let $L_m = -F_0\left(I - \Gamma_m^{'}\Gamma_m\right)$ and $U_m = F_m F_m^{'}$. Next focusing on the last two terms and using the facts that $D_{\Theta_{mm}}^{'} = D_{\Theta_m}^{'}F_m = D_{\Theta}^{'}F_m$ and $D_{\Upsilon}^{'} = D_{\Theta}^{'}F_0$:

$$\begin{aligned}
&D_{\Theta_{mm}}^{'}\left(F_m^{'}J\right) - D_{\Upsilon}^{'}\left(I - \Gamma_m^{'}\Gamma_m\right)\Delta \\
&= D_{\Theta}^{'}F_m F_m^{'}J - D_{\Theta}^{'}F_0\left(I - \Gamma_m^{'}\Gamma_m\right)\Delta \\
&= D_{\Theta}^{'}L_m\Delta + D_{\Theta}^{'}U_m J
\end{aligned} \tag{1.73}$$

Now returning to the expression characterizing the distribution and replacing its terms with the expressions above we have:

$$\begin{aligned}
\sqrt{T}\left(\mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) - \mu\left(\theta, \Theta\right)\right) &\xrightarrow{d} D_{\theta}^{'}C_m\delta + D_{\theta}^{'}P_m R + D_{\Theta}^{'}L_m\Delta + D_{\Theta}^{'}U_m J \\
&= \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}^{'}\begin{bmatrix} C_m\delta + P_m R \\ L_m\Delta + U_m J \end{bmatrix}
\end{aligned} \tag{1.74}$$

As a consequence of Assumption 2 we know that:

$$\begin{bmatrix} R \\ J \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \tag{1.75}$$

This together with the fact that P_m and U_m are symmetric means

$$\begin{bmatrix} C_m\delta + P_m R \\ L_m\Delta + U_m J \end{bmatrix} \sim N\left(\begin{bmatrix} C_m\delta \\ L_m\Delta \end{bmatrix}, \begin{bmatrix} P_m\Sigma_{11}P_m & P_m\Sigma_{12}U_m \\ U_m\Sigma_{21}P_m & U_m\Sigma_{22}U_m \end{bmatrix}\right) \tag{1.76}$$

and finally implies that

$$\begin{aligned}
&\sqrt{T}\left(\mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) - \mu\left(\theta, \Theta\right)\right) \\
&\xrightarrow{d} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' \begin{bmatrix} C_m\delta + P_m R \\ L_m\Delta + U_m J \end{bmatrix} \\
&\sim \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' N\left(\begin{bmatrix} C_m\delta \\ L_m\Delta \end{bmatrix}, \begin{bmatrix} P_m\Sigma_{11}P_m & P_m\Sigma_{12}U_m \\ U_m\Sigma_{21}P_m & U_m\Sigma_{22}U_m \end{bmatrix}\right)
\end{aligned} \tag{1.77}$$

1.9.3 Proof of Lemma A2

To use the expression for $AMSE$ for model selection, we need to estimate the unknown parameters D_θ, D_Θ, C_m, P_m, Σ_{11},Σ_{12} and Σ_{22}. Since $\hat{\theta}_f$ and $\hat{\Theta}_f$ are consistent estimators of θ and Θ, it follows that consistent estimators of D_θ and D_Θ are based on the full model estimates:

$$\begin{aligned}
\hat{D}_\theta &= \frac{\partial\mu(\hat{\theta}_f, \hat{\Theta}_f)}{\partial\theta} \rightarrow D_\theta \\
\hat{D}_\Theta &= \frac{\partial\mu(\hat{\theta}_f, \hat{\Theta}_f)}{\partial\Theta} \rightarrow D_\Theta
\end{aligned} \tag{1.78}$$

Next note that both C_m and P_m are functions of Q and selection matrices, which can be consistently estimated by the sample analog. Let $\hat{Q} = \frac{1}{T}\sum_{t=1}^{T} H_t H_t'$ and then $\hat{Q} \xrightarrow{p} Q$ as a consequence of Assumption 2. Consistent estimators of Σ_{11}, Σ_{12} and Σ_{22} are also available and have the following expressions under the assumption that error terms are uncorrelated and identically distributed:

$$\begin{aligned}
\hat{\Sigma}_{11} &= \frac{1}{T}\sum_{t=1}^{T} H_t\hat{\epsilon}_t\hat{\epsilon}_t' H_t' \rightarrow E\left(H_t\epsilon_t\epsilon_t' H_t'\right) \\
\hat{\Sigma}_{12} &= \frac{1}{T}\sum_{t=1}^{T} H_t\hat{\epsilon}_t\hat{\lambda}_t' \rightarrow E\left(H_t\epsilon_t\lambda_t'\right) \\
\Sigma_{22} &= \frac{1}{T}\sum_{t=1}^{T} \hat{\lambda}_t\hat{\lambda}_t' \rightarrow E\left(\lambda_t\lambda_t'\right)
\end{aligned} \tag{1.79}$$

Finally, U_m and L_m are non-random and known. We now consider the estimator for the local parameters δ and Δ. From Theorem A1, we know that

$$\begin{aligned}
&\sqrt{T}\left(\hat{\theta}_f - \theta\right) \xrightarrow{d} Q^{-1}R \sim N\left(0, Q^{-1}\Sigma_{11}Q^{-1}\right) \\
&\sqrt{T}\left(S_0'\hat{\theta}_f - S_0'\theta\right) \xrightarrow{d} S_0'Q^{-1}R \sim N\left(0, S_0'Q^{-1}\Sigma_{11}Q^{-1}S_0\right) \\
&\sqrt{T}\hat{\gamma}_f \xrightarrow{d} \sqrt{T}\gamma + S_0'Q^{-1}R \sim N\left(\sqrt{T}\gamma, S_0'Q^{-1}\Sigma_{11}Q^{-1}S_0\right) \\
&\hat{\delta} \xrightarrow{d} R_\delta = \delta + S_0'Q^{-1}R \sim N\left(\delta, S_0'Q^{-1}\Sigma_{11}Q^{-1}S_0\right)
\end{aligned} \tag{1.80}$$

As shown above, $\hat{\delta}$ is an asymptotically unbiased estimator for δ and converges in distribution to a linear function of the normal random vector R. Since the mean of $R_\delta R_\delta^{'}$ is

$$E\left(\left(\delta + S_0'Q^{-1}R\right)\left(\delta + S_0'Q^{-1}R\right)^{'}\right) \;=\; \delta\delta^{'} + S_0'Q^{-1}\Sigma_{11}Q^{-1}S_0 \tag{1.81}$$

then $\hat{\delta}\hat{\delta}^{'} - S_0'\hat{Q}^{-1}\hat{\Sigma}_{11}\hat{Q}^{-1}S_0$ provides an asymptotically unbiased estimator of $\delta\delta^{'}$. Similarly, we can also construct an asymptotically unbiased estimator of Δ by using the estimator from the full model. From Theorem A1, we know that

$$\begin{aligned}
&\sqrt{T}\left(\hat{\Theta}_f - \Theta\right) \xrightarrow{d} J \sim N\left(0, \Sigma_{22}\right) \\
&\sqrt{T}\left(F_0'\hat{\Theta}_f - F_0'\Theta\right) \xrightarrow{d} F_0'J \sim N\left(0, F_0'\Sigma_{22}F_0\right) \\
&\sqrt{T}\hat{\Upsilon} \xrightarrow{d} \sqrt{T}\Upsilon + F_0'J \sim N\left(\sqrt{T}\Upsilon, F_0'\Sigma_{22}F_0\right) \\
&\hat{\Delta} \xrightarrow{d} R_\Delta = \Delta + F_0'J \sim N\left(\Delta, F_0'\Sigma_{22}F_0\right)
\end{aligned} \tag{1.82}$$

Now the mean of $R_\delta J_\Delta^{'}$ is

$$\begin{aligned}
E\left(\left(\delta + S_0'Q^{-1}R\right)\left(\Delta + F_0'J\right)^{'}\right) \;&=\; \delta\Delta^{'} + \delta E\left[J^{'}\right]F_0 + E\left[S_0'Q^{-1}R\Delta^{'}\right] + S_0'Q^{-1}E\left[RJ^{'}\right]F_0 \\
&=\; \delta\Delta^{'} + S_0'Q^{-1}\Sigma_{12}F_0
\end{aligned} \tag{1.83}$$

Analogously, the mean of $J_\Delta R_\delta^{'}$ is $\Delta\delta^{'} + F_0'\Sigma_{12}'\left(Q^{-1}\right)^{'}S_0$. Consequently, the unbiased estimators of $\delta\Delta^{'}$ and $\Delta\delta^{'}$ are $\hat{\delta}\hat{\Delta}^{'} - S_0'\hat{Q}^{-1}\hat{\Sigma}_{12}F_0$ and $\hat{\Delta}\hat{\delta}^{'} - F_0'\hat{\Sigma}_{12}'\left(\hat{Q}^{-1}\right)^{'}S_0$ correspondingly.

Finally, the mean of $J_\Delta J_\Delta^{'}$is

$$\begin{aligned} E\left(\left(\Delta + F_0^{'}J\right)\left(\Delta + F_0^{'}J\right)^{'}\right) &= \Delta\Delta^{'} + \Delta E\left[J^{'}\right]F_0 + E\left[F_0^{'}J\Delta^{'}\right] + F_0^{'}E\left[JJ^{'}\right]F_0 \\ &= \Delta\Delta^{'} + F_0^{'}\Sigma_{22}F_0 \end{aligned} \tag{1.84}$$

and thus the unbiased estimator of $\Delta\Delta^{'}$ is $\hat{\Delta}\hat{\Delta}^{'} - F_0^{'}\hat{\Sigma}_{22}F_0$.

1.9.4 Proof of Theorem A3

From Theorem 1 in the main text, there is a joint convergence in distribution of all $\sqrt{T}\left(\mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) - \mu\left(\theta, \Theta\right)\right)$ to Λ_m since all of Λ_m can be expressed in terms of R and J. Since the weights are non-random, it follows that

$$\sqrt{T}\left(\bar{\mu}\left(w\right) - \mu\right) = \sum_{m=1}^{\bar{M}} w_m \sqrt{T}\left(\hat{\mu}_m - \mu\right) \xrightarrow{d} \sum_{m=1}^{\bar{M}} w_m \Lambda_m = \Lambda \tag{1.85}$$

Therefore, the asymptotic distribution of the averaging estimator is a weighted average of the normal distribution, which is also a normal distribution. By Theorem 1 and standard algebra, we can show the mean of Λ is

$$\begin{aligned} E\left(\sum_{m=1}^{\bar{M}} w_m \Lambda_m\right) &= \sum_{m=1}^{\bar{M}} w_m E\left(\Lambda_m\right) \\ &= \sum_{m=1}^{\bar{M}} w_m \left[D_\theta^{'} C_m \delta + D_\Theta^{'} L_m \Delta\right] \\ &= D_\theta^{'}\left(\sum_{m=1}^{\bar{M}} w_m C_m\right)\delta + D_\Theta^{'}\left(\sum_{m=1}^{\bar{M}} w_m L_m\right)\Delta \\ &= D_\theta^{'} C_w \delta + D_\Theta^{'} L_w \Delta \end{aligned} \tag{1.86}$$

where $C_w = \sum_{m=1}^{\bar{M}} w_m C_m$ and $L_w = \sum_{m=1}^{\bar{M}} w_m L_m$. Next, we show the variance of Λ. For any two submodels, we have

$$\begin{aligned}
Cov\left(\Lambda_m, \Lambda_l\right) &= E\left\{ \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' \left(\begin{bmatrix} C_m\delta + P_m R \\ L_m\Delta + U_m J \end{bmatrix} - E\begin{bmatrix} C_m\delta + P_m R \\ L_m\Delta + U_m J \end{bmatrix} \right) \right\} \\
&\times \left\{ \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' \left(\begin{bmatrix} C_l\delta + P_l R \\ L_l\Delta + U_l J \end{bmatrix} - E\begin{bmatrix} C_l\delta + P_m R \\ L_l\Delta + U_l J \end{bmatrix} \right) \right\}' \\
&= E\left\{ \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' \begin{bmatrix} P_m R \\ U_m J \end{bmatrix} \right\} \times \left\{ \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' \begin{bmatrix} P_l R \\ U_l J \end{bmatrix} \right\}' \\
&= \begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} E \begin{bmatrix} P_m R R' P_l & P_m R J' U_l \\ U_m J R' U_l & U_m J J' U_l \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \\
&= \begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} E \begin{bmatrix} P_m \Sigma_{11} P_l & P_m \Sigma_{12} U_l \\ U_m \Sigma_{21} P_l & U_m \Sigma_{22} U_l \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}
\end{aligned} \tag{1.87}$$

where the second equality holds buy the fact that D_θ, D_Θ, C_m, L_m, P_m, δ and Δ are constant vectors and Assumption 2. Therefore, the variance of Λ is

$$\begin{aligned}
&var\left(\textstyle\sum_{m=1}^{\bar{M}} w_m \Lambda_m\right) \\
&= \textstyle\sum_{m=1}^{\bar{M}} w_m^2 var\left(\Lambda_m\right) + \sum \sum_{m \neq l} w_m w_l Cov\left(\Lambda_m, \Lambda_l\right) + \sum \sum_{m \neq l} w_m w_l Cov\left(\Lambda_m, \Lambda_l\right)' = V
\end{aligned} \tag{1.88}$$

This completes the proof.

1.10 Appendix B

The following section characterizes the limiting distributions of the outlined plug-in-averaging estimator with data-dependent weights. It also describes the construction of a valid confidence interval for the estimated parameter of interest.

1.10.1 Asymptotic distributions of FIC and plug-in averaging estimator

As mentioned before, the FIC model selection estimator is a special case of the model averaging estimator. Model selection puts all weight on the model with the smallest value of the information criterion and gives other models zero weight. The weight function of the FIC estimator is thus

$$\hat{w}_m \;=\; 1\left\{FIC_{min} = \min\left(FIC_1, FIC_2, ..., FIC_{\bar{M}}\right)\right\} \tag{1.89}$$

where $1\{\cdot\}$ is an indicator function which takes value 1 if $FIC_{min} = \min\left(FIC_1, FIC_2, ..., FIC_{\bar{M}}\right)$ and 0 otherwise. From Lemma A2, we have consistent estimators of $\hat{D}_\theta$, $\hat{D}_\Theta$, $\hat{C}_m$, $\hat{P}_m$, $\hat{\Sigma}_{11}$, $\hat{\Sigma}_{12}$ and $\hat{\Sigma}_{22}$. Also $\hat{\delta} \xrightarrow{d} R_\delta$ and $\hat{\Delta} \xrightarrow{d} J_\Delta$, and thus we can show that $\hat{\delta}\hat{\delta}' \xrightarrow{d} R_\delta R_\delta'$, $\hat{\delta}\hat{\Delta}' \xrightarrow{d} R_\delta J_\Delta'$, $\hat{\Delta}\hat{\delta}' \xrightarrow{d} J_\Delta R_\delta'$, and $\hat{\Delta}\hat{\Delta}' \xrightarrow{d} J_\Delta J_\Delta'$, and hence for $m \in \left\{1, ..., \bar{M}\right\}$

$$\begin{aligned} FIC_m \;\xrightarrow{d}\; & \textstyle\sum_{i=1}^{N_{FIC}} S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} C_m R_\delta R_\delta' C_m' & C_m R_\delta J_\Delta' L_m' \\ L_m J_\Delta R_\delta' C_m' & \hat{L}_m J_\Delta J_\Delta' L_m' \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'} \\ + \; & \textstyle\sum_{i=1}^{N_{FIC}} S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} P_m \Sigma_{11} P_m & P_m \Sigma_{12} U_m \\ U_m \Sigma_{21} P_m & U_m \Sigma_{22} U_m \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'} \end{aligned} \tag{1.90}$$

This result implies that FIC_m has a non-standard limiting distribution. Note that this estimator is not corrected for bias as shown in Lemma A2. Similar result holds for the bias corrected version of FIC_m. The following theorem presents the asymptotic distribution of the plug-in averaging estimator defined above.

Theorem (B1). *Let $\hat{w} = \arg\min w'\hat{\psi}w$ (with $\sum_{m=1}^{\bar{M}} w_m = 1$, $w_m \geq 0$) be the plug-in weights. Suppose that Assumptions 1-2 hold and $\hat{\Sigma}_{11}$, $\hat{\Sigma}_{12}$ and $\hat{\Sigma}_{22}$ converge in probability. As $T \to \infty$, we have*

$$w'\hat{\psi}w \;\xrightarrow{d}\; w'\psi^* w$$

where ψ^{*i} *is and* $\bar{M} \times \bar{M}$ *matrix with* (m,l)*element*

$$\begin{aligned}\psi^{*i}_{m,l} &= S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} C_m R_\delta R_\delta' C_l' & C_m R_\delta J_\Delta' L_l' \\ L_m J_\Delta R_\delta' C_l' & \hat{L}_m J_\Delta J_\Delta' L_l' \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'} \\ &+ S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} P_m \Sigma_{11} P_l & P_m \Sigma_{12} U_l \\ U_m \Sigma_{21} P_l & U_m \Sigma_{22} U_l \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'}\end{aligned}$$

and $\psi^* = \sum_{i=1}^{N_{FIC}} \psi^{*i}$*. Also, we have*

$$\begin{aligned}\hat{w} \xrightarrow{d} w^* &= \arg\min\ w' \psi w \\ &\quad s.t.\ \textstyle\sum_{m=1}^{\bar{M}} w_m = 1, \quad w_m \geq 0\end{aligned}$$

and

$$\sqrt{T}\left(\bar{\mu}\left(\hat{w}\right) - \mu\right) \xrightarrow{d} \textstyle\sum_{m=1}^{\bar{M}} w_m^* \Lambda_m$$

where Λ_m *is defined in Theorem 1.*

Similar to the previous result, Theorem B1 says the limiting distribution of the plug-in estimator is non-standard with asymptotically random weights. This is the direct consequence of the fact that the local parameters δ and covariances Δ cannot be consistently estimated, and their estimates $\hat{\delta}$ and $\hat{\Delta}$ are random in the limit. These results illustrate the cost of not knowing the correct model: one can only obtain distributional convergence of the model weights. Importantly, the parameter of interest estimated using plug-in-averaging $\bar{\mu}\left(\hat{w}\right)$ is still consistent. The key to this result is the joint convergence of all submodel estimators μ_m and estimated weights $\hat{w}$. This comes from the fact that both can be expressed in terms of the same normal random vectors R and J.

1.10.2 Valid confidence interval

The above results allows us to construct valid inference confidence intervals on the parameters of interests. Let $w\left(m \middle| \hat{\delta}, \hat{\Delta}\right)$ denote a data-dependent weight function for the mth model.

Consider an averaging estimator of the focus parameter μ as

$$\begin{array}{rcl} \hat{\mu} & = & \sum_{m=1}^{\bar{M}} w\left(m \left| \hat{\delta}, \hat{\Delta} \right.\right) \hat{\mu}_m \\ & \text{s.t.} & \sum_{m=1}^{\bar{M}} w\left(m \left| \hat{\delta}, \hat{\Delta} \right.\right) = 1, \quad w\left(m \left| \hat{\delta}, \hat{\Delta} \right.\right) \geq 0 \end{array} \tag{1.91}$$

The following theorem describes the limiting distribution of the averaging estimator with data-dependent weights.

Theorem (**B2**). *Assume* $w\left(m \left| \hat{\delta}, \hat{\Delta} \right.\right) \xrightarrow{d} w\left(m \left| R_\delta, J_\Delta \right.\right)$. *Suppose that Assumptions 1-2 hold. As* $T \to \infty$, *we have*

$$\sqrt{T}\left(\bar{\mu} - \mu\right) \xrightarrow{d} \left[\begin{array}{cc} D_\theta' & D_\Theta' \end{array} \right] \left[\begin{array}{c} Q^{-1}R + \left(\sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) C_m\right) R_\delta \\ J + \left(\sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) L_m\right) R_\Delta \end{array} \right]$$

Theorem B2 shows that the limiting distribution of the averaging estimator with data-dependent weights is nonstandard in general since the estimated weights are asymptotically random. A direct construction of of Wald-statistic is not valid since the limiting distribution of $\sqrt{T}\left(\bar{\mu} - \mu\right)$ is a nonlinear function of the normal random vectors R and J and the local parameters δ and Δ. The following lemma describes the valid test statistics for testing the hypothesis on $\bar{\mu}$.

Lemma (**B3**). *Let* $\hat{\Xi}$ *be a consistent estimator of*

$$Var\left(\left[\begin{array}{cc} D_\theta' & D_\Phi' \end{array} \right] \left[\begin{array}{c} Q^{-1}R \\ J \end{array} \right] \right) = \left[\begin{array}{cc} D_\theta' & D_\Phi' \end{array} \right] \left[\begin{array}{cc} Q^{-1}\Sigma_{11}Q^{-1} & Q^{-1}\Sigma_{12} \\ \Sigma_{21}Q^{-1} & \Sigma_{22} \end{array} \right] \left[\begin{array}{c} D_\theta \\ D_\Phi \end{array} \right]$$

Suppose we are interested in testing the hypothesis $H_0 : R_\chi \mu = r$, *where* R_χ *is* $(N_\chi \times N_{FIC})$.

Then the relevant feasible Wald statistic is

$$\begin{aligned}
\hat{W} &= \left(\sqrt{T}\left(R_\chi\bar{\mu}-R_\chi\mu\right)-R_\chi\begin{bmatrix}D_\theta\\ D_\Phi\end{bmatrix}'\begin{bmatrix}\left(\sum_{m=1}^{\bar{M}}w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{C}_m\right)\hat{\delta}\\ \left(\sum_{m=1}^{\bar{M}}w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)L_m\right)\hat{\Delta}\end{bmatrix}\right)'\left(R_\chi\hat{\Xi}R_\chi'\right)^{-1}\\
&\times\left(\sqrt{T}\left(R_\chi\bar{\mu}-R_\chi\mu\right)-R_\chi\begin{bmatrix}D_\theta\\ D_\Phi\end{bmatrix}'\begin{bmatrix}\left(\sum_{m=1}^{\bar{M}}w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{C}_m\right)\hat{\delta}\\ \left(\sum_{m=1}^{\bar{M}}w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)L_m\right)\hat{\Delta}\end{bmatrix}\right)\\
&\sim\chi^2_{N_\chi}
\end{aligned}$$

This result can now be used in conducting the hypothesis on our multivariate parameter of interest. In the context of our portfolio choice application, we are practically interested whether the estimated portfolio vector x_1 is statistically different from the previous vector x_0 i.e. if there is a need to rebalance the portfolio.

1.10.3 Proof of Theorem B1

We first show that the limiting distribution of $\hat{\psi}_{m,l}$. By Lemma A1, we have $\hat{\theta}_f \xrightarrow{p} \theta$ and $\hat{\Theta}_f \xrightarrow{p} \Theta$, which implies that $\hat{D}_\theta \xrightarrow{p} D_\theta$ and $\hat{D}_\Theta \xrightarrow{p} D_\Theta$. Since $\hat{D}_\theta$, $\hat{D}_\Theta$, $\hat{C}_m$, $\hat{P}_m$, $\hat{\Sigma}_{11}$, $\hat{\Sigma}_{12}$ and $\hat{\Sigma}_{22}$ are consistent estimators of D_θ, D_Θ, C_m, P_m, Σ_{11}, Σ_{12} and Σ_{22}, we have

$$\begin{aligned}
&\left(\begin{bmatrix}\hat{D}_\theta' & \hat{D}_\Theta'\end{bmatrix}\begin{bmatrix}\hat{P}_m\hat{\Sigma}_{11}\hat{P}_l & \hat{P}_m\hat{\Sigma}_{12}U_l\\ U_m\hat{\Sigma}_{21}\hat{P}_l & U_m\hat{\Sigma}_{22}U_l\end{bmatrix}\begin{bmatrix}\hat{D}_\theta\\ \hat{D}_\Theta\end{bmatrix}\right)\\
&\xrightarrow{d}\left(\begin{bmatrix}D_\theta' & D_\Theta'\end{bmatrix}\begin{bmatrix}P_m\Sigma_{11}P_l & P_m\Sigma_{12}U_l\\ U_m\Sigma_{21}P_l & U_m\Sigma_{22}U_l\end{bmatrix}\begin{bmatrix}D_\theta\\ D_\Theta\end{bmatrix}\right)
\end{aligned}\tag{1.92}$$

by the continuous mapping theorem. Recall from Lemma A1 that $\hat{\delta} \xrightarrow{d} R_\delta$ and $\hat{\Delta} \xrightarrow{d} J_\delta$. Then by the application of Slutsky's theorem, we have

$$\begin{aligned}
\hat{\psi}_{m,l} \;=\; & \underbrace{\sum_{i=1}^{N_{FIC}} S^i \left(\begin{bmatrix} \hat{D}_\theta' & \hat{D}_\Theta' \end{bmatrix} \begin{bmatrix} \hat{C}_m \hat{\delta} \hat{\delta}' \hat{C}_l' & \hat{C}_m \hat{\delta} \hat{\Delta}' \hat{L}_l' \\ L_m \hat{\Delta} \hat{\delta}' \hat{C}_l' & \hat{L}_m \hat{\Delta} \hat{\Delta}' \hat{L}_l' \end{bmatrix} \begin{bmatrix} \hat{D}_\theta \\ \hat{D}_\Theta \end{bmatrix} \right) S^{i'}}_{\text{cross-bias}} \\
+ \; & \underbrace{\sum_{i=1}^{N_{FIC}} S^i \left(\begin{bmatrix} \hat{D}_\theta' & \hat{D}_\Theta' \end{bmatrix} \begin{bmatrix} \hat{P}_m \hat{\Sigma}_{11} \hat{P}_l & \hat{P}_m \hat{\Sigma}_{12} U_l \\ U_m \hat{\Sigma}_{21} \hat{P}_l & U_m \hat{\Sigma}_{22} U_l \end{bmatrix} \begin{bmatrix} \hat{D}_\theta \\ \hat{D}_\Theta \end{bmatrix} \right) S^{i'}}_{\text{co-variance}} \\
\xrightarrow{d} \; & \underbrace{\sum_{i=1}^{N_{FIC}} S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} C_m R_\delta R_\delta' C_l' & C_m R_\delta R_\Delta' L_l' \\ L_m R_\Delta R_\delta' C_l' & \hat{L}_m R_\Delta R_\Delta' L_l' \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'}}_{\text{cross-bias}} \\
+ \; & \underbrace{\sum_{i=1}^{N_{FIC}} S^i \left(\begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} P_m \Sigma_{11} P_l & P_m \Sigma_{12} U_l \\ U_m \Sigma_{21} P_l & U_m \Sigma_{22} U_l \end{bmatrix} \begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix} \right) S^{i'}}_{\text{co-variance}} = \psi_{m,l}^*
\end{aligned} \tag{1.93}$$

Now since all $\psi_{m,l}^*$ can be expressed in terms of the normal random vectors R and J, there is joint convergence in distribution of all $\hat{\psi}_{m,l}$ to $\psi_{m,l}^*$. Hence, it follows, that $w' \hat{\psi} w \xrightarrow{d} w' \psi^* w$.

We next show the limiting distribution of $\hat{w}$. Note that $w' \psi^* w$ is a convex minimization problem since $w' \psi^* w$ is quadratic and ψ^* is positive definite. Hence, the limiting process $w' \psi^* w$ is continuous in w and has a unique minimum. Also note that $\hat{w} = O_p(1)$ by the fact that the simplex $\left\{ w \in [0,1]^{\bar{M}} : \sum_{m=1}^{\bar{M}} w_m = 1 \right\}$ is convex. Therefore, by Theorem 2.7 of Kim and Pollard (1990), the minimizer $\hat{w}$ converges in distribution to the minimizer of $w' \psi^* w$, which is w^*.

Finally, we show the asymptotic distribution of the plug-in-averaging estimator. Since both Λ_m and w_m^* can be expressed in terms of the same normal random vectors R and J, there is a joint convergence in distribution of all $\hat{\mu}_m$ and $\hat{w}_m$. By Theorem 1, it follows that

$$\sqrt{T} \left(\bar{\mu}(\hat{w}) - \mu \right) \;=\; \sum_{m=1}^{\bar{M}} \hat{w}_m \sqrt{T} \left(\hat{\mu}_m - \mu \right) \;\xrightarrow{d}\; \sum_{m=1}^{\bar{M}} w_m^* \Lambda_m \tag{1.94}$$

This completes the proof.

1.10.4 Proof of Theorem B2

From Theorem 1, there is a joint convergence in distribution of all $\sqrt{T}\left(\mu\left(\hat{\theta}_m, \hat{\Theta}_{mm}\right) - \mu\left(\theta, \Theta\right)\right)$ to Λ_m since all of Λ_m can be expressed in terms of R and J. Also, $w\left(m \left| \hat{\delta}, \hat{\Delta} \right.\right) \xrightarrow{d} w\left(m \left| R_\delta, J_\Delta \right.\right)$ where $w\left(m \left| R_\delta, J_\Delta \right.\right)$ is a function of random vectors R and J. Therefore,

$$
\begin{aligned}
\sqrt{T}\left(\bar{\mu} - \mu\right) &= \textstyle\sum_{m=1}^{\bar{M}} w\left(m \left| \hat{\delta}, \hat{\Delta} \right.\right) \sqrt{T}\left(\hat{\mu}_m - \mu\right) \\
&\xrightarrow{d} \textstyle\sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) \left(\begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' \begin{bmatrix} C_m\delta + P_m R \\ L_m\Delta + U_m J \end{bmatrix} \right) \\
&= \textstyle\sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) \left(\begin{bmatrix} D_\theta \\ D_\Theta \end{bmatrix}' \begin{bmatrix} C_m\left(R_\delta - S_0' Q^{-1} R\right) + P_m R \\ L_m\left(R_\Delta - F_0' J\right) + U_m J \end{bmatrix} \right) \\
&= \begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} \sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) \left[C_m\left(R_\delta - S_0' Q^{-1} R\right) + P_m R\right] \\ \sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) \left[L_m\left(J_\Delta - F_0' J\right) + U_m J\right] \end{bmatrix} \\
&= \begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} \sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) \left[\left(P_m Q - C_m S_0'\right) Q^{-1} R + C_m R_\delta\right] \\ \sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) \left[\left(-L_m F_0' + U_m\right) J + L_m J_\Delta\right] \end{bmatrix} \\
&= \begin{bmatrix} D_\theta' & D_\Theta' \end{bmatrix} \begin{bmatrix} Q^{-1} R + \left(\sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) C_m\right) R_\delta \\ J + \left(\sum_{m=1}^{\bar{M}} w\left(m \left| R_\delta, J_\Delta \right.\right) L_m\right) J_\Delta \end{bmatrix}
\end{aligned}
\tag{1.95}
$$

where the last equality holds by the fact that

$$
\begin{aligned}
P_m Q - C_m S_0' &= P_m Q - (P_m Q - I) S_0 S_0' \\
&= P_m Q - P_m Q S_0 S_0' + I S_0 S_0' \\
&= P_m Q - P_m Q \begin{bmatrix} 0_{P\times K} \\ I_K \end{bmatrix} \begin{bmatrix} 0_{K\times P} & I_K \end{bmatrix} + I \begin{bmatrix} 0_{P\times K} \\ I_K \end{bmatrix} \begin{bmatrix} 0_{K\times P} & I_K \end{bmatrix} \\
&= P_m Q \begin{bmatrix} I_P & 0_{P\times K} \\ 0_{K\times P} & 0_{K\times K} \end{bmatrix} + \begin{bmatrix} 0_{P\times P} & 0_{P\times K} \\ 0_{K\times P} & I_K \end{bmatrix} \\
&= S_m \left(S_m' Q S_m \right)^{-1} S_m' Q \left(\begin{bmatrix} I_K & 0_{P\times K_m} \\ 0_{K\times P} & \Pi_m' \end{bmatrix} \begin{bmatrix} I_P & 0_{P\times K} \\ 0_{K_m\times P} & 0_{K_m\times K} \end{bmatrix} \right) + \begin{bmatrix} 0_{P\times P} & 0_{P\times K} \\ 0_{K\times P} & I_K \end{bmatrix} \\
&= S_m \left(S_m' Q S_m \right)^{-1} S_m' Q \left(S_m \begin{bmatrix} I_P & 0_{P\times K} \\ 0_{K_m\times P} & 0_{K_m\times K} \end{bmatrix} \right) + \begin{bmatrix} 0_{P\times P} & 0_{P\times K} \\ 0_{K\times P} & I_K \end{bmatrix} \\
&= S_m \begin{bmatrix} I_P & 0_{P\times K} \\ 0_{K_m\times P} & 0_{K_m\times K} \end{bmatrix} + \begin{bmatrix} 0_{P\times P} & 0_{P\times K} \\ 0_{K\times P} & I_K \end{bmatrix} \\
&= \begin{bmatrix} I_P & 0_{P\times K} \\ 0_{K\times P} & 0_{K\times K} \end{bmatrix} + \begin{bmatrix} 0_{P\times P} & 0_{P\times K} \\ 0_{K\times P} & I_K \end{bmatrix} \\
&= I_{P+K}
\end{aligned}
\tag{1.96}
$$

and also the fact that

$$\begin{aligned}
-L_m F_0^{'} + U_m &= -\left(-F_0\left(I - \Gamma_m^{'}\Gamma_m\right)\right)F_0^{'} + F_m F_m^{'} \\
&= \begin{bmatrix} 0_{G\times N} \\ I_N \end{bmatrix}\left(I - \Gamma_m^{'}\Gamma_m\right)\begin{bmatrix} 0_{N\times G} & I_N \end{bmatrix} \\
&+ \begin{bmatrix} I_G & 0_{G\times N_m} \\ 0_{N\times G} & \Gamma_m^{'} \end{bmatrix}\begin{bmatrix} I_G & 0_{G\times N} \\ 0_{N_m\times G} & \Gamma_m \end{bmatrix} \\
&= \begin{bmatrix} 0_{G\times N} \\ \left(I - \Gamma_m^{'}\Gamma_m\right) \end{bmatrix}\begin{bmatrix} 0_{N\times G} & I_N \end{bmatrix} + \begin{bmatrix} I_G & 0_{G\times N} \\ 0_{N\times G} & \Gamma_m^{'}\Gamma_m \end{bmatrix} \\
&= \begin{bmatrix} 0_{G\times G} & 0_{G\times N} \\ 0_{N\times G} & \left(I - \Gamma_m^{'}\Gamma_m\right) \end{bmatrix} + \begin{bmatrix} I_G & 0_{G\times N} \\ 0_{N\times G} & \Gamma_m^{'}\Gamma_m \end{bmatrix} \\
&= I_{G+N}
\end{aligned} \tag{1.97}$$

This completes the proof.

1.10.5 Proof of Lemma B3

Since there is a simultaneous convergence in distribution (as shown in Theorem 1), it follows that

$$\begin{aligned}
&\sqrt{T}\left(\bar{\mu} - \mu\right) - \begin{bmatrix} \hat{D}_\theta^{'} & \hat{D}_\Theta^{'} \end{bmatrix}\begin{bmatrix} \left(\sum_{m=1}^{\bar{M}} w\left(m\middle|\hat{\delta}\right)\hat{C}_m\right)\hat{\delta} \\ \left(\sum_{m=1}^{\bar{M}} w\left(m\middle|\hat{\delta}\right)\hat{L}_m\right)\hat{\Delta} \end{bmatrix} \\
&\xrightarrow{d} \begin{bmatrix} D_\theta^{'} & D_\Theta^{'} \end{bmatrix}\begin{bmatrix} Q^{-1}R \\ J \end{bmatrix} \sim N\left(0, \begin{bmatrix} D_\theta^{'} & D_\Theta^{'} \end{bmatrix}\begin{bmatrix} Q^{-1}\Sigma_{11}Q^{-1} & Q^{-1}\Sigma_{12} \\ \Sigma_{21}Q^{-1} & \Sigma_{22} \end{bmatrix}\begin{bmatrix} D_\theta \\ D_\Phi \end{bmatrix}\right) \\
&\sim N\left(0, \Xi\right)
\end{aligned} \tag{1.98}$$

And thus

$$\left(\hat{\Xi}\right)^{-1/2}\left[\left[\sqrt{T}\left(\bar{\mu}-\mu\right)\right]-\left[\begin{array}{cc}\hat{D}_{\theta}^{\prime} & \hat{D}_{\Theta}^{\prime}\end{array}\right]\left[\begin{array}{c}\left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{C}_{m}\right)\hat{\delta}\\ \left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{L}_{m}\right)\hat{\Delta}\end{array}\right]\right] \overset{d}{\rightarrow} N_{N_{FIC}}\left(0,I\right) \tag{1.99}$$

which is an N_{FIC}-variate normal distribution. If the null hypothesis, $H_0 : R\mu = r$ is true, it follows directly that

$$\sqrt{T}\left(R_{\chi}\bar{\mu}-R_{\chi}\mu\right)-R\left[\begin{array}{cc}\hat{D}_{\theta}^{\prime} & \hat{D}_{\Theta}^{\prime}\end{array}\right]\left[\begin{array}{c}\left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{C}_{m}\right)\hat{\delta}\\ \left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{L}_{m}\right)\hat{\Delta}\end{array}\right] \overset{d}{\rightarrow} N\left(0,R_{\chi}\Xi R_{\chi}^{\prime}\right) \tag{1.100}$$

and hence

$$\left(R_{\chi}\Xi R_{\chi}^{\prime}\right)^{-1/2}\left(\sqrt{T}\left(R_{\chi}\bar{\mu}-R_{\chi}\mu\right)-R_{\chi}\left[\begin{array}{cc}\hat{D}_{\theta}^{\prime} & \hat{D}_{\Theta}^{\prime}\end{array}\right]\left[\begin{array}{c}\left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{C}_{m}\right)\hat{\delta}\\ \left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{L}_{m}\right)\hat{\Delta}\end{array}\right]\right) \overset{d}{\rightarrow} N_{N_{\chi}}\left(0,I\right) \tag{1.101}$$

This allows a test statistic to be formed

$$\begin{aligned} W &= \left(\sqrt{T}\left(R_{\chi}\bar{\mu}-R_{\chi}\mu\right)-R_{\chi}\left[\begin{array}{cc}D_{\theta}^{\prime} & D_{\Theta}^{\prime}\end{array}\right]\left[\begin{array}{c}\left(\sum_{m=1}^{\bar{M}} w\left(m\left|R_{\delta},J_{\Delta}\right.\right)C_{m}\right)\delta\\ \left(\sum_{m=1}^{\bar{M}} w\left(m\left|R_{\delta},J_{\Delta}\right.\right)L_{m}\right)\Delta\end{array}\right]\right)^{\prime}\left(R_{\chi}\Xi R_{\chi}^{\prime}\right)^{-1} \\ &\times\left(\sqrt{T}\left(R_{\chi}\bar{\mu}-R_{\chi}\mu\right)-R_{\chi}\left[\begin{array}{cc}D_{\theta}^{\prime} & D_{\Theta}^{\prime}\end{array}\right]\left[\begin{array}{c}\left(\sum_{m=1}^{\bar{M}} w\left(m\left|R_{\delta},J_{\Delta}\right.\right)C_{m}\right)\delta\\ \left(\sum_{m=1}^{\bar{M}} w\left(m\left|R_{\delta},J_{\Delta}\right.\right)L_{m}\right)\Delta\end{array}\right]\right) \\ &\sim \chi_{N_{\chi}}^{2} \end{aligned} \tag{1.102}$$

which is the sum of the squares of N_{χ} random variables (i.e. of N_{FIC} random variables if $N_{\chi} = I$), each asymptotically uncorrelated standard normal and so W is asymptotically $\chi_{N_{\chi}}^{2}$ distributed. A hypothesis test of size α can be conducted by comparing W against $C_{\alpha} = F^{-1}\left(1-\alpha\right)$ where $F\left(\cdot\right)$ is the cdf of a $\chi_{N_{\chi}}^{2}$. Since W is infeasible, we replace the unknown elements of the covariance matrix with consistent estimate to compute a feasible

Wald statistic,

$$\begin{aligned}\hat{W} &= \left(\sqrt{T}\left(R_\chi\bar{\mu}-R_\chi\mu\right)-R_\chi\left[\begin{array}{cc}\hat{D}_\theta' & \hat{D}_\Theta'\end{array}\right]\left[\begin{array}{c}\left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{C}_m\right)\hat{\delta}\\ \left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{L}_m\right)\hat{\Delta}\end{array}\right]\right)' \left(R_\chi\hat{\Xi}R_\chi'\right)^{-1} \\ &\times \left(\sqrt{T}\left(R_\chi\bar{\mu}-R_\chi\mu\right)-R_\chi\left[\begin{array}{cc}\hat{D}_\theta' & \hat{D}_\Theta'\end{array}\right]\left[\begin{array}{c}\left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{C}_m\right)\hat{\delta}\\ \left(\sum_{m=1}^{\bar{M}} w\left(m\left|\hat{\delta},\hat{\Delta}\right.\right)\hat{L}_m\right)\hat{\Delta}\end{array}\right]\right) \\ &\sim \chi^2_{N_\chi}\end{aligned} \tag{1.103}$$

which has the same asymptotic distribution as the infeasible Wald.

Chapter 2

Focused shrinkage with an application to portfolio choice

Abstract. We propose a shrinkage estimator for parameters θ which improves the mean squared error of functions $x(\theta)$ over standard choices. When the restricted model estimator is in the class of minimum distance estimators, we project onto the restricted parameter space using a matrix based on the derivatives of $x(\theta)$. The proposed matrix is shown to minimize the bias among estimators in this class. This choice is then incorporated into a shrinkage procedure. We derive a risk bound for shrinkage estimators which allows for arbitrary projection matrices. Using this result, it is shown the proposed projection matrix can lead to substantially lower risk. The improvement is largest when the restricted model has nontrivial bias. This is shown with our motivating example: implementing a mean-variance portfolio choice rule. Our method is applied by shrinking toward a model with restricted covariance matrix. Extensive simulations demonstrate improved risk in this case. The estimator is also implemented in a portfolio choice application to futures data. Our approach outperforms standard procedures here as well.

JEL: C31, C52, C53, C58.

2.1 Introduction

In estimation, we are often unsure if the parameters of interest satisfy a given restriction. If the restriction holds, estimating the tighter specification leads to improved performance. If the restriction fails, the full model could be the better choice. The basic idea of shrinkage estimation is to weight between restricted and full model estimates. This weighting is done using a test of the restriction. If the test strongly rejects, most of the weight is put on the full model. If the opposite is true, more weight is put on the restricted model. In this way,

shrinkage trades-off between the full and restricted models in a formal way. The trade-off is done as a smooth function of the test, in contrast to the hard-thresholding in pretest estimators. This leads to well documented improvements in risk properties. See Saleh (2006) for an excellent exposition of this testing perspective and a review of the literature.

In much of the shrinkage literature, the full model is parameterized by θ and the restricted model is a fixed point θ_0. For example, James and Stein (1961) estimate the mean of a multivariate normal by shrinking toward zero. When the restricted model is a single point θ_0 there is no estimation under the restriction. In other cases, the parameters are not restricted to a single point but to a linear subspace $\Theta_0 = \{\theta |\, R'\theta = a\}$ where R is a $m \times p$ matrix, a is a $p \times 1$ vector and θ is a $m \times 1$ vector. Testing is then of $\theta \in \Theta_0$. Early papers considering this include Stein (1966), Sclove (1968) and Oman (1982,a,b) among others. See Saleh (2006) for more details.

When the subspace Θ_0 is not a singleton, estimation of the restricted model is no longer trivial. In general, there will be several methods which satisfy the constraint. When $\widehat{\theta}$ is the unrestricted estimator, a standard approach is the following minimum distance estimator:

$$\widetilde{\theta}^V = \arg\min_{\theta \in \Theta_0} \left(\widehat{\theta} - \theta\right)' V^{-1} \left(\widehat{\theta} - \theta\right). \tag{2.1}$$

Here, V is an $m \times m$ invertible positive semidefinite (PSD) matrix. The restricted estimator $\widetilde{\theta}^V$ is a projection from $\widehat{\theta}$ to the closest point in the subspace Θ_0 where distance is measured using V^{-1}. When the subspace is defined with linear restrictions as in Θ_0, $\widetilde{\theta}^V$ has the following form:

$$\widetilde{\theta}^V = \widehat{\theta} - VR\left(R'VR\right)^{-1}\left(R'\widehat{\theta} - a\right).$$

This is the solution to (2.1) for any invertible PSD matrix V. When considering shrinkage, which V should we choose when estimating the restricted model? Assume the full model estimator satisfies $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \Rightarrow N\left(0, \Omega\right)$. Previous literature has chosen either $V = I$ or $V = \Omega$.[1] Are these the best choices? A related issue is how to set up the shrinkage weights optimally for a given V. Finally, how can we choose V and the weighting scheme jointly such that risk is minimal?

In this paper, we study the risk properties of shrinkage estimation when the restricted

[1] See Oman (1982a,b) for $V = I$ and Hansen (2016) for $V = \Omega$.

model is estimated using arbitrary V. The goal is to find a choice of V and a shrinkage weighting scheme which outperforms standard choices. We focus on the risk properties of estimating an $S \times 1$ vector of functions $x(\theta)$ with argument θ. This is related to the weighted mean squared error (MSE) risk criterion considered in Bhattacharya (1966), Sclove (1968), Berger (1976a,b, 1982), Berger, Bock, Brown, Casella and Gleser (1977), Saleh (2006) and Hansen (2016). Note that the identity function is a special case simplifying the problem. In the sequel, we show that choosing V using the relevant functions $x(\theta)$ can significantly improve the MSE of shrinkage estimators compared with standard choices. This idea is related to the Focused Information Criteria (see Claeskens and Hjort (2003)) because how we shrink depends on the functions of interest $x(\theta)$. We call our estimation approach focused shrinkage.

The methods developed below are general. However, our results were motivated by the problem of estimating a portfolio selection rule. In this situation, the functions $x(\theta)$ are portfolio positions of an investor. We use the notation $\mu(\theta)$ for this specific case. The underlying parameters θ describe how asset prices randomly propagate. Throughout the sequel, we consider portfolio functions $\mu(\theta)$ from Garleanu and Pedersen (2013) to motivate our methods. This is a generalization of the classic Markowitz (1952) results.

Many shrinkage results assume the data follows a normal distribution. In economic and financial applications such as our motivating example, this normality assumption is often questionable. As a result, we instead consider shrinkage estimators using a localization framework to derive asymptotic results. This avoids strong distributional assumptions. This approach is taken in Saleh (2006) and Hansen (2016) and our setup follows Hansen (2016). Parameters are assumed to be localized toward a fixed point: $\theta = \theta_0 + n^{-1/2}\Delta$. It is only known that $\theta_0 \in \Theta_0$, which does not necessarily pin down a single value. The term $n^{-1/2}\Delta$ formalizes our uncertainty that the restriction is correct. The $n^{-1/2}$ rate of convergence allows for meaningful asymptotic results. In general, we do not know Δ. The case $\Delta = 0$ can be thought of as the researcher being certain the restricted model holds. This assumption removes the need for shrinkage because there is no uncertainty about the restriction. Non-zero values of Δ allow scope for asymptotic improvements.

When the restrictions imposed are true ($\Delta = 0$), it is well known that projecting with $V = \Omega$ leads to the minimal asymptotic variance among estimators $\widetilde{\theta}^V$. This minimum variance property is then transferred to plug-in estimators of $x(\theta)$ through the delta method. Hansen (2016) argues that $V = \Omega$ should be used because of this optimality. In the sequel, we show

that $V = \Omega$ need not be optimal when $\Delta \neq 0$. In this case, the restriction $R'\theta = a$ is not true. The result is that projections imposing this restriction incur a bias term. This bias depends on Δ and is only zero when $\Delta = 0$. We show the bias term is minimized by choosing V using the form of $x(\theta)$, while the variance term is minimized by $V = \Omega$. The overall asymptotic MSE of $\widetilde{\theta}^V$ is not minimized when $V = \Omega$. Therefore, it is unclear which V should be used when estimating the restricted model in the $\Delta \neq 0$ case. Unless we are absolutely sure the restriction holds, it is possible other Vs preform better in finite samples. In shrinkage contexts, being unsure about the restriction is a fundamental assumption. This motivates considering other choices when applying shrinkage estimation.

Estimation of the restricted model is only one component of shrinkage. We next consider how to optimally weight between full and restricted estimators when V is arbitrary. Hansen (2016) derives a risk bound for a standard class of shrinkage estimators in our localization framework under the assumption $V = \Omega$. He also derives the optimal weighting scheme which minimizes this bound. Our second contribution is to generalize the risk bound derived in Hansen (2016) by allowing for arbitrary V. We find that different choices of V lead to different optimal weighting schemes. The general form of the bound suggests a choice of V (and corresponding weighting scheme) which depends on the derivatives of $x(\theta)$. Motivated by this, we propose a shrinkage procedure which utilizes the functions of interest $x(\theta)$.

Motivated by an empirical application to trading futures contracts, we design a simulation study which realistically represents price dynamics in this example. The goal is to estimate parameters which describe prices θ and incorporate this into a trading rule $\mu(\theta)$. The final trading rule $\mu(\theta)$ is the object of interest. In our simulations and empirical application, we propose shrinking toward subspaces commonly considered in portfolio choice contexts. Using a trading rule from Garleanu and Pedersen (2013), our shrinkage estimator is shown to substantially reduce MSE compared with standard procedures. The MSE is reduced for almost all parameters considered. While these results are specific to our motivating example, they show the possibility of similar improvements in other situations.

Finally, we apply our procedure to an empirical application in trading futures contracts. We consider trading in a diversified set of eight highly liquid futures over the period 2007-2016. Estimated trading rules are compared using Sharpe ratios, by far the most common measure used. We find that focused shrinkage outperforms estimation with standard methods by a wide margin.

The use of shrinkage methods in portfolio choice is widespread. We only mention a few representative papers here. Most methods focus on covariance matrix estimation. Ledoit and Wolf (2003, 2004) propose shrinking an unrestricted covariance estimator toward a target with more structure. The target is often a factor model. Another approach restricts portfolio weights directly. Jagannathan and Ma (2003) and DeMiguel, Garlappi, Nogales and Uppal (2009) use this idea when implementing the global minimum variance (GMV) portfolio. As shown in DeMiguel et al. (2009), restricting GMV portfolio weights is equivalent to shrinking the estimated covariance matrix. Both these approaches have been extended to large dimensional cases. See Fan, Fan and Lv (2008) and Fan, Zhang and Yu (2012). Most methods for estimating large dimensional covariance matrices can be thought of as shrinkage. See Fan, Liao and Liu (2016) for a general overview.

Our methods differ from previous literature such as Ledoit and Wolf (2003, 2004) in that we focus on the portfolio trading rule. Our weighting between the full and restricted model optimizes the AMSE of the estimated portfolio weights, not the covariance matrix itself. As the final portfolio weights are the objects of interest, this directly addresses the problem. Ledoit and Wolf (2003, 2004) focus on measures of distance between the estimated and true covariance matrix such as the trace or Frobenius norm. These measures ignore the intention of plugging the estimate into a trading rule. The localization of parameters in our setup is another point of departure. This gives a formal bias-variance trade-off and finite sample interpretation. These are both absent in previous work.

Our approach allows for arbitrary trading rules, including the regression based Markowitz rule considered in the sequel. Estimation and shrinkage accounts for both regression coefficients and the residual covariance matrix simultaneously. Because of this joint treatment, the procedure can minimize the AMSE of the estimated trading rule. Methods such as Ledoit and Wolf (2003, 2004) focus solely on the covariance matrix, ignoring the mean. The level of shrinkage does not depend on the estimated mean or the resulting trading rule. Our methods allow for general restrictions of regression coefficients and covariance matrices. This makes more general shrinkage targets possible.

Estimates which constrain portfolio weights such as Jagannathan and Ma (2003) also utilize the final trading rule. However, results in this area depend on the GMV portfolio's form. In particular, GMV does not use the mean of returns. Only a covariance estimate is needed. This rules out application to the Markowitz rule considered below. The approach

would have to be modified for application to mean-variance portfolios. This literature also does not consider localization of parameters or the resulting bias-variance trade-off. See Chapter 3 for a full comparison between our methods and previous results in the GMV case.

The remainder of the paper is organized as follows. Section 2 derives risk properties of minimum distance estimators under localization. Section 3 derives risk bounds for shrinkage estimators under arbitrary projection matrices V. Our focused shrinkage procedure is proposed here. Section 4 describes the portfolio choice model which motivated our theoretical results. Simulations are presented showing the superior risk properties of focused shrinkage compared with standard alternatives. Section 5 applies focused shrinkage to portfolio choice in futures contracts. Again, our method is shown to outperform alternatives. Section 6 concludes. All proofs are presented in the appendix. Additional simulation results are presented there as well.

2.2 Risk of Minimum Distance Estimators

In this section, we consider the risk properties of minimum distance estimators under localization. As outline above, the statistical model has a parameterization $\theta \in \Theta$ where Θ is an open set in euclidean space. The restricted model subspace $\Theta_0 = \{\theta | \, R'\theta = a\}$ is a subset of Θ. The localization scheme assumes parameters follow an array structure. With n observations, the data is distributed according to $\theta_n = \theta_0 + n^{-1/2}\Delta$. The limit point θ_0 is in Θ_0. This setup has been widely used in the literature. See van der Vaart (1998) for more discussion.

Our basic assumption is that $\widehat{\theta}$ is a full model estimator such that:

$$\sqrt{n}\left(\widehat{\theta} - \theta_n\right) \Rightarrow Z \sim N\left(0, \Omega\right). \tag{2.2}$$

This is true in a wide variety of situation, making our results widely applicable. The corresponding restricted model estimators $\widetilde{\theta}^V$ are defined as in the previous section. Estimators in this class are in the subspace Θ_0 for any invertible PSD matrix V. Here and in the sequel we ignore the possibility that $\widetilde{\theta}^V \notin \Theta$ in finite samples. This will not be the case asymptotically.

The risk criterion considered in this paper is asymptotic MSE of the plug-in estimator $x\left(T_n\right)$ where T_n is an arbitrary estimator of θ_n:

$$\lim_{n\to\infty} E\left[n\left(x\left(T_n\right) - x\left(\theta_n\right)\right)'\left(x\left(T_n\right) - x\left(\theta_n\right)\right)\right]. \tag{2.3}$$

Because of the delta method, this is equivalent to

$$\rho\left(W, T_n\right) = \lim_{n\to\infty} E\left[n(T_n - \theta_n)' W (T_n - \theta_n)\right], \tag{2.4}$$

where $D_\theta = \frac{\partial}{\partial\theta} x\left(\theta_0\right)$ and

$$W = D_\theta' D_\theta.$$

The situation is equivalent to considering the weighted asymptotic MSE with weight matrix W. In our context, W represents the functions we are interested in estimating $x\left(\theta\right)$. It requires no additional work to consider the risk (2.4) with arbitrary PSD matrices W. Throughout the sequel, we consider (2.4) with general W. The notation $\rho\left(W, T_n\right)$ is used to represent (2.4) with arbitrary W and T_n.

Hansen (2016) derives the result

$$\sqrt{n}\left(\widetilde{\theta}^V - \theta_n\right) \Rightarrow Z - VR\left(R'VR\right)^{-1} R'\left(Z + \Delta\right),$$

where Z is the asymptotic distribution from (2.2). Defining $P = VR\left(R'VR\right)^{-1} R'$, the limiting distribution can be written as

$$\left(I - P\right) Z - P\Delta,$$

which is

$$N\left(-P\Delta, \left(I - P\right)\Omega\left(I - P\right)'\right). \tag{2.5}$$

When $\Delta = 0$, (2.5) is the classical limiting distribution of minimum distance estimators under the restriction $R'\theta = a$. It is well known that the covariance matrix $\left(I - P\right)\Omega\left(I - P\right)'$ is minimized (in the sense of a PSD matrix) when $V = \Omega$. The result is that, in the case where the restriction exactly holds ($\Delta = 0$), the optimal projection is $V = \Omega$. However, when $\Delta \neq 0$ the distribution incurs a bias term $-P\Delta$. This is because $\widetilde{\theta}^V$ is in Θ_0, but the true parameters θ_n are not. This additional bias term changes the risk properties of $\widetilde{\theta}^V$. The risk has the

following form:

$$\begin{aligned}\rho\left(W,\widetilde{\theta}^{V}\right) &= E\left[\left(\left(I-P\right)Z-P\Delta\right)'W\left(\left(I-P\right)Z-P\Delta\right)\right],\\ &= tr\left(W\cdot\left(I-P\right)\Omega\left(I-P\right)'\right)+\Delta'P'WP\Delta,\\ &= var\left(W,\widetilde{\theta}^{V}\right)+bias^{2}\left(W,\widetilde{\theta}^{V}\right)\end{aligned}$$

Theorem 1. *Let W be an arbitrary PSD matrix and Ω be invertible. The variance term is minimized by choosing $V=\Omega$. If in addition W is invertible, the bias term is minimized by choosing $V=W^{-1}$.*

The variance term is still minimized by projection with $V=\Omega$. The weight matrix W does not change the optimality of this choice. On the other hand, the bias term is minimized by projecting with $V=W^{-1}$. In the case where $W=D_{\theta}'D_{\theta}$ corresponding to estimating $x\left(\theta\right)$, this involves using the form of $x\left(\theta\right)$ when deciding on V. As there is no reason for $\Omega=W^{-1}$, when $\Delta\neq0$ it is no longer the case that $V=\Omega$ is the optimal choice. The optimal choice will depend on the relative weights of bias and variance components. When Δ is small, $\rho\left(W,\widetilde{\theta}^{V}\right)$ will be minimized with V close to Ω. When Δ is larger, the bias term becomes increasingly important, improving the performance of W^{-1}.

The invertibility of the matrix W is important for these results. In many situations, when the risk weighting W is motivated by functions of interest $x\left(\theta\right)$, W will not be invertible. This is because, if the dimension of $x\left(\theta\right)$ is smaller than the dimension of θ, the resulting matrix $W=D_{\theta}'D_{\theta}$ must be rank deficient. While we do not have a specific matrix which minimizes the bias term in this case, it is clear that Ω will not minimize it in general. As a result, the same criticism of $V=\Omega$ is valid. In cases like this we consider regularization below.

The optimal choice of V for bias-variance trade-off is unclear. It is unlikely this term can be chosen entirely from data in practice. The optimal V depends on Δ, Ω and W in general. Each of these components must be estimated. In addition, the term Δ is not consistently estimable because of the localization structure. Finally, V can have large dimension. In our empirical application, V is a 52-dimensional symmetric matrix. There is little chance this many tuning parameters can be well estimated. Because of these issues, the paper focuses on $V=\Omega$ and $V=W^{-1}$.[2]

[2]When $x\left(\theta\right)=\theta$ as in classical shrinkage, $W=D_{\theta}'D_{\theta}$ is I. It follows that the optimal projection for the bias term is $V=I$. Because of this, we consider the choice $V=I$ a special case of $V=W^{-1}$.

In shrinkage applications, $\Delta \neq 0$ is the empirically relevant case with Δ possibly having large values. As a result, choosing $V = W^{-1}$ can have smaller risk and be more accurate in finite samples. This is especially true when $\widetilde{\theta}^V$ is incorporated into a shrinkage estimator. Weighting between $\widehat{\theta}$ and $\widetilde{\theta}^{W^{-1}}$ can significantly outperform standard choices. This is shown with our motivating example in Section 2.4. The proposed shrinkage estimator is presented in the following section.

2.3 Focused Shrinkage

In this section, we first define the class of shrinkage estimators considered. This follows the framework of Hansen (2016), but with slightly changed notation. Shrinkage estimation requires a test that the restriction $R'\theta = a$. In the case where the functions of interest are $x(\theta)$, the test used is

$$\overline{D}_n = \left[\sqrt{n}\left(x(\widehat{\theta}) - x(\widetilde{\theta}^V)\right)\right]' \left[\sqrt{n}\left(x(\widehat{\theta}) - x(\widetilde{\theta}^V)\right)\right].$$

This is asymptotically equivalent to the following by the delta method:

$$\begin{aligned} \widetilde{D}_n &= \left[D_\theta\sqrt{n}\left(\widehat{\theta} - \widetilde{\theta}^V\right)\right]' \left[D_\theta\sqrt{n}\left(\widehat{\theta} - \widetilde{\theta}^V\right)\right], \\ &= n\left(\widehat{\theta} - \widetilde{\theta}^V\right)' D_\theta' D_\theta \left(\widehat{\theta} - \widetilde{\theta}^V\right). \end{aligned}$$

Notice that the matrix $D_\theta' D_\theta$ is present in this test. More generally, when considering an arbitrary PSD matrix W in $\rho(W, T_n)$, the test will be

$$D_n = n\left(\widehat{\theta} - \widetilde{\theta}^V\right)' W \left(\widehat{\theta} - \widetilde{\theta}^V\right). \tag{2.6}$$

The weighting between full and restricted models has a positive-rule form

$$\widehat{w}_n = \left(1 - \frac{\tau}{D_n}\right)_+, \tag{2.7}$$

where $\tau > 0$ is chosen by the researcher. This is standard and chosen for simplicity. The final shrinkage estimator is

$$\widehat{\theta}^* = \widehat{w}_n\widehat{\theta} + (1 - \widehat{w}_n)\widetilde{\theta}^V. \tag{2.8}$$

Hansen (2016) derives an asymptotic risk bound for $\widehat{\theta}^*$ under the assumption $V = \Omega$. He notes that this is optimal when $\Delta = 0$ and does not consider other choices. However, as we have shown in the previous section, this is not the case when $\Delta \neq 0$. Motivated by this, we generalize Hansen's bound allowing for arbitrary projection matrices V.[3]

The notation $\lambda_{\max}(M)$ is used for the maximum eigenvalue of a symmetric matrix M.

Theorem 2. *Let V be an invertible PSD matrix and W a PSD matrix. Define the following matrices:*

$$\begin{aligned} B &= R\left(R'VR\right)^{-1}R'VWVR\left(R'VR\right)^{-1}R', \\ \overline{A} &= \left(R'VR\right)^{-1}R'VW\Omega R, \\ A^* &= W^{1/2}VR\left(R'VR\right)^{-1}R'VW^{1/2}, \\ K &= W^{1/2}\Omega R\left(R'VR\right)^{-1}R'\Omega W^{1/2}. \end{aligned}$$

The full model risk is:

$$\rho\left(W,\widehat{\theta}\right) = tr\left(W\Omega\right).$$

The shrinkage estimator $\widehat{\theta}^*$ which uses V for the restricted model estimator $\widetilde{\theta}^V$ has an upper risk bound of:

$$\begin{aligned} \rho\left(W,\widehat{\theta}^*,\tau\right) &\leq tr\left(W\Omega\right) - \tau E\left\{\frac{2\left[tr\left(\overline{A}\right) - 2\lambda_{\max}^{1/2}\left(A^*\right)\lambda_{\max}^{1/2}\left(K\right)\right] - \tau}{\left(Z+\Delta\right)'B\left(Z+\Delta\right)}\right\} \\ &\equiv \overline{\rho}\left(W,\widehat{\theta}^*,\tau\right). \end{aligned} \tag{2.9}$$

[3]Hansen (2016) presents his analogous risk bound in a slightly different way. He considers a uniform bound over a ball of Δ values defined as

$$\mathbf{H}(c) = \left\{\Delta \middle| \Delta'B\Delta \leq tr\left(A^*\right)c\right\}$$

for any $c > 0$ and with $V = \Omega$. His bound can be stated similarly to ours by omitting the final step in his proof, in particular taking the supremum over $\Delta \in \mathbf{H}(c)$. Hansen (2016) derives several results for a risk bound which is uniform over $\mathbf{H}(c)$. Our bound is pointwise in Δ and our comparison with Hansen's results are of his analogous pointwise bound. This is true for the remainder of the paper and we make no further mention of it.

When $V = \Omega$, the equality $A^* = K$ holds and these matrices are equivalent to the corresponding A matrix in Hansen (2016). In addition, because of the properties of the trace, $tr(A) = tr\left(\overline{A}\right) = tr\left(A^*\right) = tr(K)$. Finally, the matrix B is equivalent to the corresponding matrix in Hansen (2016). The result is that, when $V = \Omega$, the terms $tr\left(\overline{A}\right) - 2\lambda_{\max}^{1/2}\left(A^*\right)\lambda_{\max}^{1/2}(K)$ and $\left(Z+\Delta\right)'B\left(Z+\Delta\right)$ reduce to the corresponding terms in Hansen (2016). Therefore, the risk bounds (2.9) and (2.11) are equivalent to those in Hansen (2016). The same τ^* presented above is optimal in this case as well. The conclusion is that our result subsumes the corresponding result in Hansen (2016), but allows for consideration of different V.

This only improves the risk if the following condition holds:

$$0 < \tau \leq 2\left[tr\left(\overline{A}\right) - 2\lambda_{\max}^{1/2}\left(A^*\right)\lambda_{\max}^{1/2}\left(K\right)\right]. \tag{2.10}$$

If (2.10) holds, the optimal choice of τ for a given V is:

$$\tau^* = \left[tr\left(\overline{A}\right) - 2\lambda_{\max}^{1/2}\left(A^*\right)\lambda_{\max}^{1/2}\left(K\right)\right],$$

resulting in the risk bound:

$$\begin{aligned}\rho\left(W, \widehat{\theta}^*, \tau^*\right) &\leq tr\left(W\Omega\right) - E\left\{\frac{\left[tr\left(\overline{A}\right) - 2\lambda_{\max}^{1/2}\left(A^*\right)\lambda_{\max}^{1/2}\left(K\right)\right]^2}{\left(Z+\Delta\right)'B\left(Z+\Delta\right)}\right\} \\ &= \overline{\rho}\left(W, \widehat{\theta}^*, \tau^*\right).\end{aligned} \tag{2.11}$$

The condition (2.10) must hold for (2.9) to show any improvement over full model estimation. This is related to other eigenvalue conditions required for risk improvements or minimax estimators in the shrinkage literature. See for example Bock (1975), Berger (1976b), Berger, Bock, Brown, Casella and Gleser (1977), Judge and Bock (1978), Fomby and Hill (1979), Oman (1982a) among others. These are generalizations of the classical requirement $\tau = p - 2 > 0$.[4]

While (2.10) is complicated, it can be seen as a restriction on the unbalancedness of the eigenvalues in the matrices $\overline{A}$, A^* and K. This is clearer in the $V = \Omega$ case which reduces to $2\left[tr\left(A^*\right) - 2\lambda_{\max}\left(A^*\right)\right]$. Here, the sum of the eigenvalues of A^* must be greater than $2\lambda_{\max}\left(A^*\right)$. This fails if $\lambda_{\max}\left(A^*\right)$ is much larger than all other eigenvalues. As the matrices $\overline{A}$, A^* and K depend on Ω, V and W, (2.10) is also related to how balanced these primitive matrices are. The following theorem makes this clear in a specific case.

The notation $\kappa\left(M\right)$ is used for the condition number of the symmetric matrix M (the largest eigenvalue divided by the smallest eigenvalue).

Theorem 3. *Assume $R'R = I_p$ and $a = 0$. Let W and Ω be invertible PSD matrices and*

[4](2.10) reduces to the analogous condition in Hansen (2016) when $V = \Omega$.

$V = W^{-1}$. *Then*

$$\overline{\rho}\left(W, \widehat{\theta}^*, \tau\right) \leq tr\left(W\Omega\right) - \tau E\left\{\frac{2\left[\lambda_{\min}\left(\Omega\right)\lambda_{\min}\left(W\right)\left\{p - 2\kappa\left(\Omega\right)\kappa^2\left(W\right)\right\}\right] - \tau}{\left(Z+\Delta\right)' B\left(Z+\Delta\right)}\right\},$$

and it is required that

$$0 \leq \left\{p - 2\kappa\left(\Omega\right)\kappa^2\left(W\right)\right\}$$

for the bound to show any risk improvement for $\widehat{\theta}^*$ *over* $\widehat{\theta}$. *If instead* $V = \Omega$, *then*

$$\overline{\rho}\left(W, \widehat{\theta}^*, \tau\right) \leq tr\left(W\Omega\right) - \tau E\frac{2\left[\frac{\lambda_{\min}(\Omega)\lambda_{\min}(W)}{\kappa(\Omega)}\left\{p - 2\kappa^3\left(\Omega\right)\kappa\left(W\right)\right\}\right] - \tau}{\left(Z+\Delta\right)' B\left(Z+\Delta\right)},$$

and it is required that

$$0 \leq \left\{p - 2\kappa^3\left(\Omega\right)\kappa\left(W\right)\right\}$$

for the bound to show any risk improvement for $\widehat{\theta}^*$ *over* $\widehat{\theta}$.

The results in Theorem 3 weaken the bounds from Theorem 2, but make it clear how eigenvalues of Ω and W influence risk. Improvements are only shown if $\kappa\left(\Omega\right)$ and $\kappa\left(W\right)$ are not too large. This corresponds to the eigenvalues of the matrices Ω and W being balanced. This is also the case for the tighter bound (2.9).

2.3.1 The Choice of V

Assuming (2.10) is satisfied, the value of τ minimizing the risk bound is $\tau^* = tr\left(\overline{A}\right) - 2\lambda_{\max}^{1/2}\left(A^*\right)\lambda_{\max}^{1/2}\left(K\right)$. Because the matrix B is PSD, $\widehat{\theta}^*$ dominates $\widehat{\theta}$. With τ^*, finding the optimal bound reduces to choosing V which minimizes (2.11). In the sequel, when implementing estimators, $\tau = \tau^*$ is always used. Some parameters in τ^* are unknown and plug-in estimators are used.

Ideally, V would be chosen to minimize (2.11) over all possible V. Given the bound's complicated form, it is unclear what the optimal choice of V is in general. This will depend on the terms

$$\left[tr\left(\overline{A}\right) - 2\lambda_{\max}^{1/2}\left(A^*\right)\lambda_{\max}^{1/2}\left(K\right)\right]^2 \tag{2.12}$$

and

$$E\left[\frac{1}{(Z+\Delta)' B (Z+\Delta)}\right]. \tag{2.13}$$

Both of these terms depend on the choice of V and its interaction with W, Ω and Δ. The restriction (2.10) is also required, so not all choices of V are available in every situation.

Despite this complexity, it is possible to choose V which optimizes over (2.13). To this end, write $B(M)$ for the matrix B constructed using $V = M$. Notice that $B(V)$ has a similar form to the asymptotic covariance of a minimum distance estimator (see Newey and McFadden (1994)). Using this form and arguments similar to those used to minimize the covariance of minimum distance estimators, if the matrix W is invertible, standard arguments show that $B\left(W^{-1}\right) \leq B(V)$ for any other V (see Newey and McFadden (1994) Section 5.2). This implies that, for any realization of Z and any value of Δ,

$$(Z+\Delta)' B\left(W^{-1}\right)(Z+\Delta) \leq (Z+\Delta)' B(V)(Z+\Delta),$$

for any V. A consequence of this is that (2.13) is maximized with $V = W^{-1}$.

The $bias^2\left(W, \widetilde{\theta}^V\right)$ term discussed in Section 2.2 is equivalent to $\Delta' B(V) \Delta$. This influences the risk bound through (2.13). When Δ is varied, (2.12) does not change, but larger values of Δ make (2.13) smaller. As $V = W^{-1}$ maximizes (2.13), this choice maximizes the risk improvement from (2.13). This result and the analysis in Section 2.2 suggest that, when bias is a serious consideration for Θ_0, $V = W^{-1}$ will preform well. Of course, differences in (2.12) can overturn the improvement. We find ample evidence in Section 2.4 of cases where $V = W^{-1}$ substantially outperforms $V = \Omega$.

There is no reason why $V = W^{-1}$ or $V = \Omega$ is optimal. A more sophisticated choice of optimal V using (2.11) is of interest. This idea is similar to choosing V using the bias-variance trade-off from Section 2.2. Direct estimation of optimal V is difficult for the same reasons given in Section 2.2. Restricting V to a convex combination between W^{-1} and Ω is a potentially fruitful approach. Simulations using a simple version of this idea did not preform well. Developing methods for choosing optimal V is an important topic. We leave this to future research.

Estimation Error

Implementing the shrinkage procedure requires estimating Ω and D_θ. In general, Ω is estimated more poorly than D_θ. The derivatives D_θ only require $\widehat{\theta}$ whereas Ω is the asymptotic covariance of $\widehat{\theta}$. In addition, Ω has significantly more parameters.

As a consequence of this difference in estimation error, $\widetilde{\theta}^V$ is often more accurate with $V = W^{-1}$. The error in τ^* is also reduced with this choice. This is partially because Ω is used less frequently. Another issue is the form of τ^*. The matrices $\overline{A}$, A^* and K are complex and τ^* requires their eigenvalues. A closer look at these matrices shows that $V = W^{-1}$ yields some simplifications. Because of these simplifications and the greater accuracy of D_θ, the error of τ^* is much smaller with $V = W^{-1}$.

We will show in Section 2.4 that, in the large majority of cases, simulations using $V = \Omega$ preform substantially worse than full model estimation. This happens despite the fact that (10) is satisfied. Under this restriction, risk improvements occur asymptotically. In contrast, $V = W^{-1}$ shows large risk improvements for almost all specifications. This stark difference is because of reduced estimation error. This improvement is a strong reason for using $V = W^{-1}$.

The poor performance of $V = \Omega$ in finite samples suggests other implications. While the optimal matrix V for the risk bound (2.11) is of theoretical interest, it is unclear it will result in finite sample improvements. Estimation error could nullify any benefits, even to the point where full model estimation is superior. Further theory and simulations are needed to determine the presence and extent of improved risk in this case.

2.3.2 The Choice of Θ_0

It has been suggested to us that, when bias is an issue, a better solution is to choose a more accurate restriction Θ_0. If Θ_0 closely represented the underlying model, bias will be of second order. In this case, $V = \Omega$ could be the better choice. This is easier said than done. The point of shrinkage is that we don't know if Θ_0 is correct. This includes the possibility that Δ is far enough from zero that $V = \Omega$ is not the best choice. In most cases, an *a priori* restriction Θ_0 where bias is of little importance is not obvious. If we can effectively pick a true restricted model, why do shrinkage at all?

The issue of bias becomes more important as the dimension of θ increases. With a large number of parameters, choosing Θ_0 with trivial bias is very unlikely. In the portfolio choice example considered below, our baseline case has 52 parameters and 28 restrictions. This is

for eight assets and these numbers increase rapidly with dimension. We propose two simple choices of Θ_0 used frequently in the literature. These are described in the sequel. Other Θ_0 which are *a priori* less biased are not obvious. Moreover, because of the instability of financial data, any more sophisticated Θ_0 may not be persistently accurate.

Another important issue is overfitting. Overfitting is a major cause of poor performance for statistical trading strategies. Because the output of any strategy is portfolio returns, it is very tempting to fit different models on the same data to improve results. Unsurprisingly, when this is taken out-of-sample the outcome is often poor. In the portfolio choice situation, a requirement that Θ_0 be picked with negligible bias is an invitation to overfitting. Trying different restrictions looking for a better fit is no guarantee of better results. As it is unlikely Θ_0 can be chosen *a priori* with negligible bias, $V = W^{-1}$ is often the best choice.

When implementing portfolio choice rules, simple restrictions which are somewhat plausible are standard in practice. More complicated restrictions are seen as implausible given the instability of financial data and likely to exacerbate overfitting. Section 2.4 fully discusses the portfolio choice problem and the chosen restrictions.

2.3.3 Focused Shrinkage with Unbalanced W

As noted above, in many cases where W follows from functions of interest $x(\theta)$, W will not be invertible. This is always the case when estimating optimal portfolio rules with the classical form of Markowitz (1952). Among other things, rules in the spirit of Markowitz require estimation of the mean of asset returns, their variance and covariances between them. This implies that the number of parameters will always be larger than the number of assets. The functions $\mu(\theta)$ from Garleanu and Pedersen (2013) used as our motivating example below are an extension of the basic Markowitz (1952) idea. As a result, the W corresponding to $\mu(\theta)$ will not be invertible. In these cases, we suggest the choices $W = D_\theta' D_\theta$ and $V = (D_\theta' D_\theta + \alpha I)^{-1}$ where α controls the level of regularization of $D_\theta' D_\theta$. Projecting using this regularized matrix allows for similar properties of the resulting risk bound. We call shrinkage estimators using this projection matrix focused shrinkage.

Even with the choices $W = D_\theta' D_\theta$ and $V = (D_\theta' D_\theta + \alpha I)^{-1}$, condition (2.10) can fail. Condition (2.10) is not trivial. We found with our motivating example that in most cases it does not hold. The main cause of this failure is the unbalancedness of $W = D_\theta' D_\theta$ derived from $\mu(\theta)$. This is caused by $W = D_\theta' D_\theta$ independently of the chosen V. In particular, $V = \Omega$

and $V = (D_\theta' D_\theta + \alpha I)^{-1}$ often both violate (2.10). It can be difficult to find V where (2.10) is satisfied. These findings are in line with Theorem 3. The main take-away is that some functions $x(\theta)$ create problems for shrinkage regardless of the chosen V. This is a result of the chosen risk criterion $\rho(W, T_n)$ when W is unbalanced. In the classical case $x(\theta) = \theta$, $W = I$ and this problem does not arise.

To address this issue, we consider the choices $W = (D_\theta' D_\theta + \alpha I)$ and $V = (D_\theta' D_\theta + \alpha I)^{-1}$. This changes our notion of risk by changing W . The interpretation of risk as a measure of how close an estimator $x(T_n)$ is to the true values $x(\theta_n)$ becomes perturbed. Risk in this case is:

$$\begin{aligned} n(T_n - \theta_n)' (D_\theta' D_\theta + \alpha I)(T_n - \theta_n) \;=\; & n(T_n - \theta_n)' (D_\theta' D_\theta)(T_n - \theta_n) \\ & + n(T_n - \theta_n)' (\alpha I)(T_n - \theta_n). \end{aligned} \tag{2.14}$$

The first term in (2.14) is the risk without regularization ($\alpha = 0$). The second term is standard MSE scaled by α. Choosing $W = (D_\theta' D_\theta + \alpha I)$ balances the goals of minimizing both these terms, where α determines their relative importance. In the sequel, α is chosen so both are of the same order. Our simulations and application show this approach preforms well compared with alternatives. The extra stability from regularization is worth the cost.

2.4 Risk and Portfolio Choice

2.4.1 Asset Dynamics and a Trading Rule

In classical portfolio choice theory, the market is assumed to have a vector of S random asset returns $\overline{r}_t = p_t - p_{t-1}$.[5] An optimal portfolio for an investor depends on the mean and covariance of the returns $\overline{r}_t$ along with some notion of risk preferences. This is the basic idea of Markowitz (1952) and the large literature that has followed. In this section, we describe a model of asset dynamics and a trading rule in this spirit. The setup is then used to analyze our estimation method.

[5]Note that we consider returns as price differences $p_t - p_{t-1}$. This is in contrast to the commonly used $(p_t - p_{t-1})/p_{t-1}$. Price differences are considered because Garleanu and Pedersen (2013) use them when deriving the trading rule presented below.

The dynamics of asset returns are assumed to follow:

$$\overline{r}_{t+1} = \overline{B} f_t + \overline{u}_{t+1}, \tag{2.15}$$

where

$$\begin{aligned} \overline{r}_t &\quad (S \times 1), \\ f_t &\quad (K \times 1), \\ \overline{u}_t &\quad (S \times 1), \\ \overline{B} &\quad (S \times K), \end{aligned}$$

and

$$\begin{aligned} E_t\left(u_{t-1}\right) &= 0, \\ var_t\left(u_{t-1}\right) &= \overline{\Sigma}. \end{aligned}$$

Here, f_t is a vector of factors driving $\bar{r}_t$. The system $(\bar{r}_t', f_t')$ is assumed to be stationary.

The analysis of risk that follows is motivated by our empirical application to trading futures contracts. Different futures contracts can have large differences in prices and price changes. This makes them of unequal scale and creates difficulty when comparing across assets. Because of this, we consider the return equation (2.15) with each asset $s \in \{1, \cdots, S\}$ normalized by its unconditional standard deviation σ^s. Normalized return equations have the form

$$r_{t+1}^s = \frac{\left(p_t^s - p_{t-1}^s\right)}{\sigma^s} = \frac{\overline{B}^s}{\sigma^s} f_t + \frac{\overline{u}_{t+1}^s}{\sigma^s}. \tag{2.16}$$

We write r_t for the vector of normalized returns. The terms B and u_{t+1} are the corresponding normalized coefficients and errors . Returns r_t have the following dynamics

$$r_{t+1} = B f_t + u_{t+1}. \tag{2.17}$$

As the assets' standard deviations σ^s are not observed, they must be estimated from data when implementing results. This is done with sample standard deviations $\widehat{\sigma}_T^s$. As $\widehat{\sigma}_T^s$ is consistent, asymptotic results are not impacted.

Returns are normalized as in (2.16) to make price changes comparable across assets. This

also makes the covariances in u_{t+1} of the same scale. This makes common subspace restrictions on asset returns more realistic. See below for more discussion. Normalization by standard deviations is common in industry practice and has been considered previously in the literature. For example, Moskowitz, Ooi and Pedersen (2012) normalize returns in an analogous way.

For each time period, μ_t is the vector of asset shares held by the investor. The shares in μ_t correspond to the return vector r_t. The notation $\Delta\mu_t$ is used for changes in asset holdings across periods. It is assumed that changing μ_t incurs the following transaction cost:

$$TC\left(\Delta\mu_t\right) = \frac{\lambda}{2}\Delta\mu_t'\overline{\Sigma}\Delta\mu_t$$

This depends on the non-normalized covariance matrix $\overline{\Sigma}$ and the parameter λ. The transaction cost $TC\left(\Delta\mu_t\right)$ and the following trading rule are essentially those proposed in Garleanu and Pedersen (2013):

$$\mu_t = \left(1 - \frac{\gamma}{\gamma+\lambda}\right)\mu_{t-1} + \left(\frac{\gamma}{\gamma+\lambda}\right)aim_t, \tag{2.18}$$

where

$$aim_t = \gamma^{-1}\Sigma^{-1}Bf_t.$$

An investor with current asset shares μ_{t-1} rebalances their portfolio to μ_t using this rule. The parameter γ represents investor risk aversion. Garleanu and Pedersen (2013) suggest the values $\gamma = 10^{-9}$ and $\lambda = 10^{-6}$ for a large asset manager and we adopt these in the sequel. The trading rule (2.18) is a function of the underlying parameters (B, Σ) which describe the statistical model (2.15). If our goal is to estimate portfolios μ_t, we are in the situation considered in previous sections.

The derivatives of the function $\mu\left(B, \Sigma\right)$ can easily be derived:

$$\begin{aligned}
\frac{\partial\mu\left(B,\Sigma\right)}{\partial b_{ij}} &= \left(\frac{\gamma}{\gamma+\lambda}\right)\frac{1}{\gamma}\Sigma^{-1}\frac{\partial B}{\partial b_{ij}}f_t, \\
\frac{\partial\mu\left(B,\Sigma\right)}{\partial \sigma_{ij}} &= -\left(\frac{\gamma}{\gamma+\lambda}\right)\frac{1}{\gamma}\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_{ij}}\Sigma^{-1}Bf_t.
\end{aligned}$$

The difference between derivatives of B and Σ is the reason $W = D_\theta' D_\theta$ is unbalanced. For realistic values of (B, Σ), changes in B change the optimal portfolio weights $\mu\left(B, \Sigma\right)$ far more

than Σ. This creates large derivatives in B compared with Σ. The result is an unbalanced matrix $D_\theta' D_\theta$. This issue will appear in most trading rules with Markowitz form.

2.4.2 Localization and Full Model Estimation

In order to put the asset dynamics (2.15) into the framework of this paper, the parameters (B, Σ) must be localized. These are in matrix form and need to be vectorized. We do this by defining vectors θ_B and θ_Σ. The parameters θ_B stack the coefficients from each return equation on top of each other. For the covariance matrix we define

$$\theta_\Sigma = \begin{bmatrix} \sigma_1^2 \\ \vdots \\ \sigma_S^2 \\ \sigma_{1,2} \\ \vdots \\ \sigma_{S-1,S} \end{bmatrix}.$$

The parameters θ_Σ stack the elements of the covariance matrix with the variance terms first. The covariance terms are second and their order is not particularly important. Stacking each of these components gives a vector of all parameters:

$$\theta = \begin{bmatrix} \theta_B \\ \theta_\Sigma \end{bmatrix}.$$

This convention will be used throughout the sequel. The parameters θ are localized $\theta_n = \theta_0 + n^{-1/2}\Delta$ where θ_0 is contained in a subspace Θ_0. Specifics about subspaces considered in this example are presented below.

Defining the $KS \times S$ matrix

$$F_t = \begin{bmatrix} f_t & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & f_t \end{bmatrix},$$

the equation $r_{t+1} = Bf_t + u_{t+1}$ can be written as:

$$r_{t+1} = F_t'\theta_B + u_{t+1}.$$

We assume data is observed over T time periods. This data is represented as follows:

$$\mathbf{r} = \begin{bmatrix} r_1 \\ \vdots \\ r_T \end{bmatrix}, \mathbf{F} = \begin{bmatrix} F_0' \\ \vdots \\ F_{T-1}' \end{bmatrix}.$$

The least squares estimator of θ_B is:

$$\widehat{\theta}_B = \left(\mathbf{F}'\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{r}.$$

Writing the stacked vector of residuals for equations $i = 1, \ldots, S$ as $\widehat{u}_i$, an estimator of θ_Σ can be defined as the average of the corresponding residuals:

$$\widehat{\theta}_\Sigma = \begin{bmatrix} \widehat{\sigma}_1^2 \\ \vdots \\ \widehat{\sigma}_S^2 \\ \widehat{\sigma}_{1,2} \\ \vdots \\ \widehat{\sigma}_{S-1.S} \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\widehat{u}_1'\widehat{u}_1 \\ \vdots \\ \frac{1}{T}\widehat{u}_S'\widehat{u}_S \\ \frac{1}{T}\widehat{u}_1'\widehat{u}_2 \\ \vdots \\ \frac{1}{T}\widehat{u}_{S-1}'\widehat{u}_S \end{bmatrix}.$$

Under weak restrictions:

$$\sqrt{n}\left(\begin{bmatrix} \widehat{\theta}_B \\ \widehat{\theta}_\Sigma \end{bmatrix} - \begin{bmatrix} \theta_{Bn} \\ \theta_{\Sigma n} \end{bmatrix}\right) \Rightarrow N\left(\mathbf{0}, \Omega\right). \tag{2.19}$$

More specifics about these asymptotic results can be found in Hamilton (1994). This estimator satisfies the assumption needed to apply the shrinkage methods described in previous sections.[6]

[6]The estimator $\left(\widehat{\theta}_B', \widehat{\theta}_\Sigma'\right)$ assumes that all parameters in (B, Σ) are estimated. In many situations, it is

2.4.3 Parameter Subspaces

There are several relevant subspaces for the portfolio choice example presented above. In the sequel we focus on two:

$$\begin{aligned}\Theta_0^1 &= \{\theta | \text{ The covariance terms in } \Sigma \text{ are all zero.}\}, \\ \Theta_0^2 &= \{\theta | \text{ The covariance terms in } \Sigma \text{ all have the same value.}\}.\end{aligned}$$

We call these restrictions exclusion and pooling respectively. Both these restrictions reduce the set of potential parameters Θ to a subspace Θ_0 which is not a single point. This implies that the choice of projection matrix V is relevant as described in previous sections. Estimation error in the covariance matrix of returns is a well known problem in portfolio choice. Applying our method with these subspaces helps address this issue.

Given our choices of subspace, the normalization of returns by standard deviations is important. For example, the subspace Θ_0^2 where all covariances are equivalent is unlikely to hold if all assets have large differences in scale. The test in the shrinkage estimator would strongly reject and there would be no risk improvement. When assets are normalized, it is reasonable to think this subspace is approximately true. In this case, Θ_0^2 is an equicorrelation model. As we show below, shrinking toward either of these subspaces can lead to improvements in both simulations and applications.

Many previous papers consider shrinking covariance terms when estimating a covariance matrix. See the introduction for citations and discussion. Our approach differs in that we choose a restricted estimator and weighting scheme which focuses on the function of interest $\mu_t(B, \Sigma)$. The approach also allows us to consider shrinkage toward arbitrary linear subspaces. As a consequence, we can consider shrinking covariance terms toward common values. Finally, our method allows us to consider subspaces which restrict non-covariance terms. Any linear subspace of the parameters $(\theta_B', \theta_\Sigma')$ can be used.

When estimating asset dynamics to apply a trading rule, often simple models are used with relatively small amounts of data. This is because financial data is noisy and unstable.

of interest to *a priori* restrict some parameters to zero. For Σ, this is simple because any of the terms in $\widehat{\theta}_\Sigma$ can be ignored. There is an issue when estimating θ_B because excluding variables impacts estimation of other variables. In addition, the residuals $\widehat{u}_i$ depend on first stage estimation of B. This is changed when excluding parameters *a priori*. Fortunately, it is easy to get a similar vector of estimators for B in this case. This is done by excluding components of F_t which correspond to parameters restricted to zero. Similar results as (2.19) hold in this case. We use such a restricted model in our results below.

Observations from many periods back may not accurately describe current asset dynamics. One approach is to make the model more complicated and consider a larger window of observation. This creates its own problems as a model can become more complicated in a large number of ways. Relatively simple models (such as the one presented above) and restricted parameters Θ_0 tend to preform better in general. But it is usually unclear if the imposed restriction $R'\theta = a$ is correct. In reality, restrictions are only approximately true, if true at all.

This is exactly the situation our shrinkage estimator is expected to preform well in. We do not believe the restrictions Θ_0^1 and Θ_0^2 are exactly true. But it is unclear which other Θ_0 is obviously better. This is certainly true before analysis of the data. The choices Θ_0^1 and Θ_0^2 can be approximately true, but with moderate to substantial bias. As the results that follow show, the proposed shrinkage estimator performs better than standard choices in this situation. A more restricted Θ_0 with smaller bias is not obvious *a priori* and unlikely to be stable. Searching over different Θ_0 for a better restriction leads to overfitting.

2.4.4 Baseline Specification

In the remainder of this section, we present simulations verifying that focused shrinkage outperforms other competitors. This is done using parameters which mimic the portfolio choice problem in our empirical application. We first describe the baseline parameters used in simulations. In doing this, we describe the factors f_t used in our application below. The baseline parameter values are set using initial estimates from futures data. These estimates come from a large number of futures contracts and are representative of the asset class in general. In our application, monthly observations and trading are considered. The system is described with this in mind.

Each asset return r_t^s is assumed to depend on two corresponding factors f_{1t}^s and f_{2t}^s. With this notation, the return equations have the following form:

$$\begin{aligned} r_{t+1}^1 &= \beta_1 f_{1t}^1 + \beta_2 f_{2t}^1 + u_{1t+1}, \\ r_{t+1}^2 &= \beta_3 f_{1t}^2 + \beta_4 f_{2t}^2 + u_{2t+1}, \\ &\vdots \\ r_{t+1}^S &= \beta_{2S-1} f_{1t}^S + \beta_{2S} f_{2t}^S + u_{1t+1}. \end{aligned} \tag{2.20}$$

The corresponding B and f_t are:

$$B = \begin{bmatrix} \beta_1 & \beta_2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \beta_3 & \beta_4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \beta_{2S-1} & \beta_{2S} \end{bmatrix}, \; f_t = \begin{bmatrix} f_{1t}^1 \\ f_{2t}^1 \\ \vdots \\ f_{1t}^S \\ f_{2t}^S \end{bmatrix}.$$

In order to simplify simulations, an autoregressive structure is assumed on f_t. This is a reasonable approximation for the factors used in the sequel. Specifically, we assume for $s \in \{1, \ldots, S\}$ and $j \in \{1, 2\}$:

$$f_{j(t+1)}^s = \left(1 - c_j^s\right) f_{jt}^s + \varepsilon_{j(t+1)}^s, \tag{2.21}$$

where $\left(1 - c_j^s\right)$ captures the persistence of factor j for security s. In general, the additional structure on the factors (2.21) is not needed for any of our theoretical results to hold. In particular, we do not use this structure in our empirical application.

In our application, the factors f_{1t}^s and f_{2t}^2 are taken to be returns over different time periods normalized by σ^s. These types of factors are frequently used in practice. The parameters for our baseline specification come from initial estimates using the following choices. For each r_t^s, f_{1t}^s and f_{2t}^2 are the previous 4-month and 12-month returns for asset s normalized by σ^s:

$$\begin{aligned} f_{1t}^s &= \frac{p_t^s - p_{t-4}^s}{\sigma^s}, \\ f_{2t}^s &= \frac{p_t^s - p_{t-12}^s}{\sigma^s}. \end{aligned}$$

The unobserved value σ^s is again replaced with the estimate $\widehat{\sigma}_T^s$. In our application, we also consider f_{1t}^s as the 1-month lagged return $\left(p_t^s - p_{t-1}^s\right)/\sigma^s$ instead of the 4-month lagged return. This is a good choice as f_{1t}^s and f_{2t}^s are meant to capture short and long run influences on returns respectively. The form (2.21) is why simulations are not based on $\left(p_t^s - p_{t-1}^s\right)/\sigma^s$. In this case, equations (2.20) and (2.21) contradict each other. In our application, results are presented for both cases.

We now outline the baseline choice of the parameters in our example. The coefficients on

f_{1t}^s are set to 0.1 and those on f_{2t}^s to -0.01. Errors for returns are assumed to have a normal distribution for simplicity:

$$u_t \sim N(0, \Sigma).$$

Here, variances are 0.9. Covariances are set in the following way. First, all terms are set to 0.2. Then, to every second covariance term we add $\overline{\Delta} = 0.1$ and from every third we subtract $\overline{\Delta} = 0.1$. For 3 securities the matrix is:

$$\Sigma = \begin{bmatrix} 0.9 & 0.2 & 0.3 \\ 0.2 & 0.9 & 0.1 \\ 0.3 & 0.1 & 0.9 \end{bmatrix}.$$

This makes covariance terms dispersed, where $\overline{\Delta}$ is the degree of dispersion. For the pooling subspace Θ_0^2, $\overline{\Delta}$ describes how much the model deviates from the subspace restriction.

The values $c_1^s = 0.5$ and $c_2^s = 0.1$ describe the degree of persistence of f_{1t}^s and f_{2t}^s respectively. The errors for factors are assumed to follow a normal distribution:

$$\varepsilon_t \sim N(0, \Phi).$$

The parameters in Φ are set to have $var(\varepsilon_{1t}^s) = 5$, $var(\varepsilon_{2t}^s) = 20$. All covariances are set to 4 in line with empirical results. We assume no covariance between u_t and ε_t.

The number of observations in all simulations is $T = 100$. This again is motivated by our application. The baseline number of assets is $S = 8$ and the corresponding number of factors is $K = S \times 2 = 16$. A larger S is not considered in order to keep the number of estimated parameters away from $T = 100$. This is because our asymptotic results are not derived for a diverging number of parameters. The initial holdings of the investor were set to $\mu_{t-1} = 0$.

2.4.5 Simulated Risk for the Trading Rule

With these baseline choices for parameters, the simulated risk of our estimator was explored over a range of values. This was done by varying specific parameters from the baseline specification. Simulated risk results are presented in this subsection. Corresponding risk bounds were also derived for comparison. These are presented and discussed in the appendix.

We considered the following scenarios.

Scenario 1 (Varying the variance of factors) Different values of $var\left(\varepsilon_{1t}^{s}\right)$ were considered on an equally spaced grid of 11 points ranging from 3 to 8. The corresponding $var\left(\varepsilon_{2t}^{s}\right)$ was set to $var\left(\varepsilon_{2t}^{s}\right) = var\left(\varepsilon_{1t}^{s}\right) \times 4$ for all s.

Scenario 2 (Varying the persistence of factors) Different values of c_1^s were considered on an equally spaced grid of 11 points ranging from 0.4 to 1. The corresponding c_2^s was set to $c_2^s = c_1^s/5$ for all s.

Scenario 3 (Varying the signal strength) Different values of $\beta\left(f_{1t}^{s}\right)$ were considered on an equally-spaced grid of 11 points ranging from 0.01 to 0.7. The corresponding $\beta\left(f_{2t}^{s}\right)$ was set to $\beta\left(f_{2t}^{s}\right) = \beta\left(f_{1t}^{s}\right)/\left(-10\right)$ for all s.

Scenario 4 (Varying the variance of returns) Different values of $var\left(u_t^s\right)$ were considered on an equally-spaced grid of 11 points ranging from 0.8 to 1.

Scenario 5 (Varying the covariance of returns) Different values of $cov\left(u_t^s, u_t^{\overline{s}}\right)$ for $s \neq \overline{s} \in \{1, \ldots, S\}$ were considered on an equally-spaced grid of 11 points ranging from 0.001 to 0.3.

Scenario 6 (Varying the dispersion of covariances) Different values of $\overline{\Delta}$ were considered on an equally-spaced grid of 11 points ranging from 0.001 to 0.45.

Scenario 7 (Varying the size of the system) Different numbers of assets S were considered ranging from 3 to 10. The corresponding K was set to $K = 2 \times S$ so there are 2 factors for each asset.

In each of these scenarios, 1000 simulations were conducted with $T = 100$ observations. Average risk is computed across these simulations. Simulated risk is computed with $\overline{W}$ while estimators are constructed assuming W. These values are distinct and can be different. The final vector f_t in the $T = 100$ observations was used for f_t in $\mu\left(\theta\right)$. Therefore, $\mu\left(\theta\right)$ is random across simulations and our risk measure is an average. This more closely represents the rolling window estimation used in our application below.

When computing shrinkage estimators, an estimated value of $D_\theta' D_\theta$ is needed. The pilot $\widetilde{\theta}^I$ was used and $D_{\widetilde{\theta}^I}' D_{\widetilde{\theta}^I}$ was plugged-in where needed. The covariance Ω must also be estimated.

This was done using its sample analog. In particular, when estimating Ω, we assumed it was not known f_t satisfies the structure (2.21).

The level of regularization considered was $\alpha = \delta \times \max(|D_\theta' D_\theta|)$ where $\max(|D_\theta' D_\theta|)$ is the largest absolute value in $D_\theta' D_\theta$. Scaling by $\max(|D_\theta' D_\theta|)$ makes αI and $D_\theta' D_\theta$ of the same order. Simulations showed results are not sensitive to the choice of δ (as long as it is not very small). The value $\delta = 1$ was used as this gave stable results.

For each realization, four estimators were computed. First, full model estimation (FM). Second, the shrinkage estimator with the choices $W = (D_\theta' D_\theta + \alpha I)$, $V = \Omega$. We call this estimator focused Hansen (FH). Third, our focused shrinkage (FS) estimator with the choices $W = (D_\theta' D_\theta + \alpha I)$, $V = (D_\theta' D_\theta + \alpha I)^{-1}$. Finally, the shrinkage estimator under the assumption $W = \Omega^{-1}$, $V = \Omega$. The different choice of $W = \Omega^{-1}$ changes the weighting scheme between restricted and full model estimators. We call this estimator unfocused Hansen (UH) because the weighting does not account for the functions of interest $x(\theta)$. For all estimators and all parameter values simulated, the condition (2.10) holds.

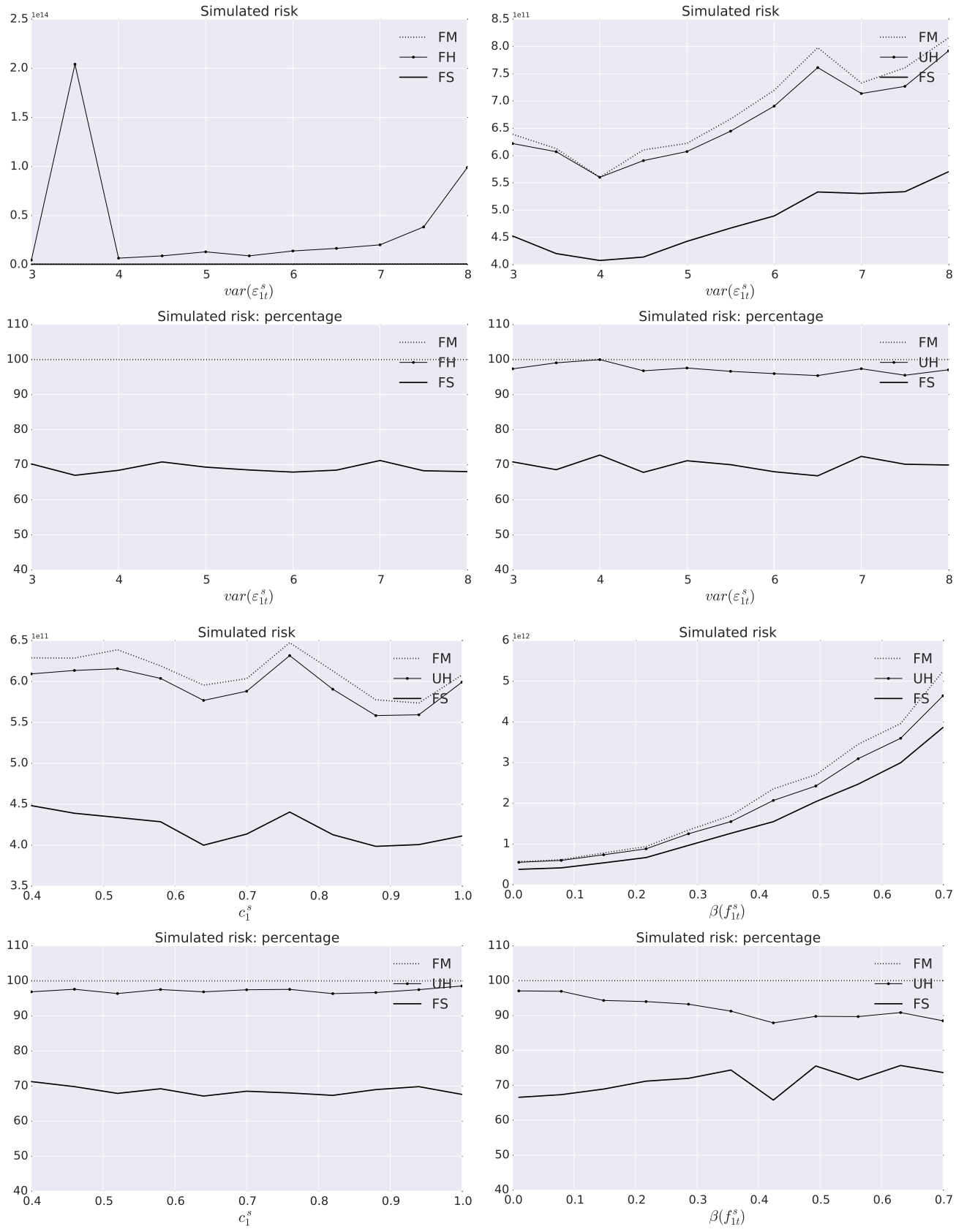

Figure 2.4.1 – Simulated risk, Θ_0^2 and $\overline{W} = D_\theta^{'} D_\theta$.

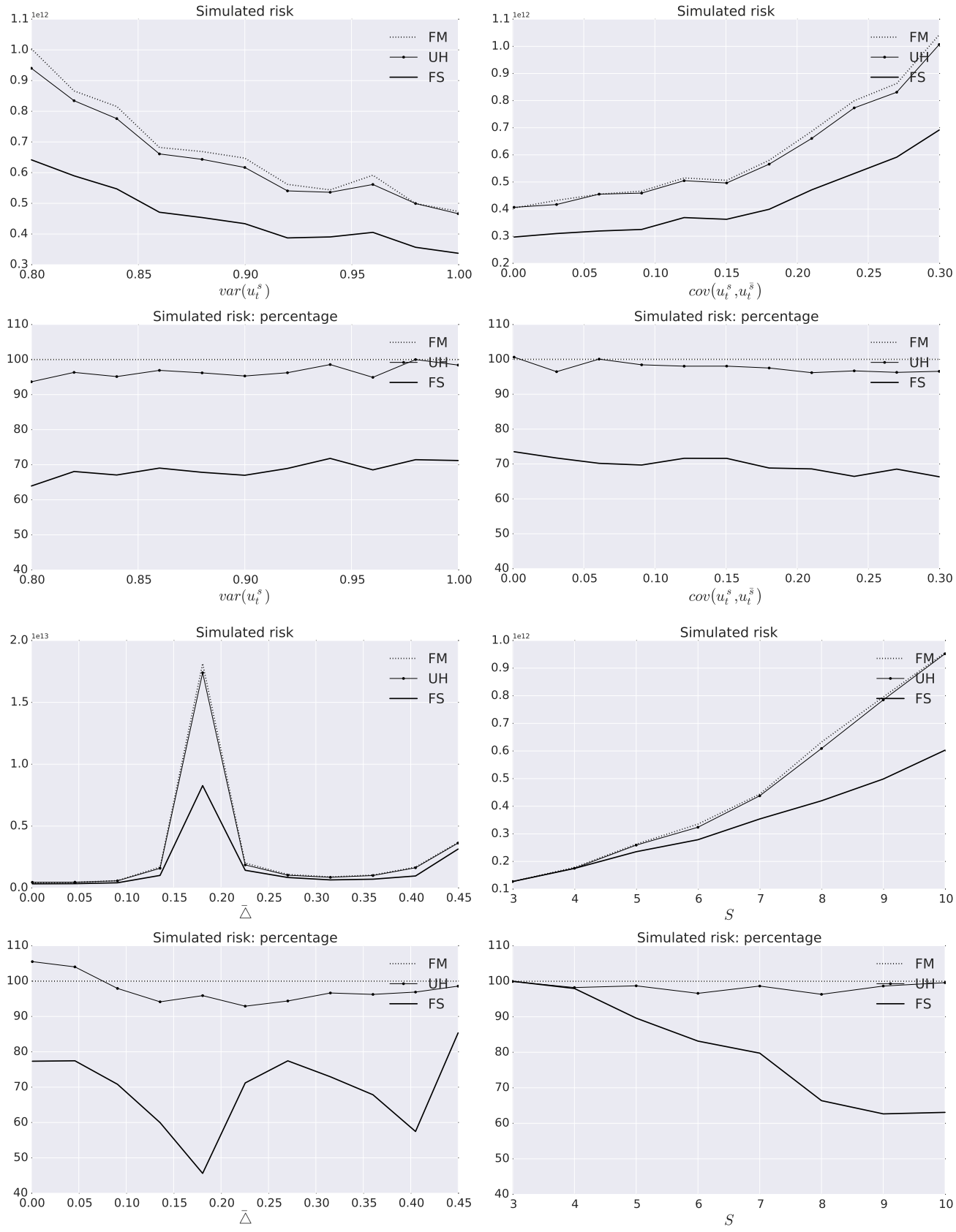

Figure 2.4.2 – Simulated risk, Θ_0^2 and $\overline{W} = D_\theta^{'} D_\theta$.

When simulating risk, both $\overline{W}_1 = D_\theta' D_\theta$ and $\overline{W}_2 = (D_\theta' D_\theta + \alpha I)$ were considered. The choice $\overline{W}_2$ is in line with $W = (D_\theta' D_\theta + \alpha I)$ used to compute estimators. As we will see, there is strong performance under $\overline{W}_1$ despite the regularization. Results for $\overline{W}_1$ are presented above. Results for $\overline{W}_2$ were similar. These are presented and discussed in the appendix.

Estimation with FH produced very poor results. Its simulated risk was frequently an order of magnitude larger than the other estimators. In Figure 1, for the pooling subspace Θ_0^2 and Scenario 1, we present the simulated risk of FH, FM and FS. The results are in the upper left-hand panel. The values from FH are so large, it is difficult to see FM and FS. Risk for other scenarios was similar. Because of this, other scenarios are not reported. Instead, we present risk comparisons of UH, FM and FS.

One reason for the poor results of FH is mismatch between $W = (D_\theta' D_\theta + \alpha I)$ and $\overline{W}_1 = D_\theta' D_\theta$. However, in the corresponding $\overline{W}_2 = (D_\theta' D_\theta + \alpha I)$ case the results are similar (see appendix). Again, FH has risk substantially above FM for a large majority of simulated parameters. This happens despite the fact that (2.10) is satisfied in all cases. In addition, even if bias $\overline{\Delta}$ is reduced to zero, the poor performance can persist. These problems are due to estimation error in primitive matrices. As the results show, risk can depart substantially from asymptotic theory in finite samples. In contrast, for both choices of $\overline{W}$, FS has almost uniformly lower risk than FM. These simulations confirm our claim that projecting with $D_\theta' D_\theta$ reduces estimation error and results in superior performance.

The remaining panels in Figures 2.4.1 and 2.4.2 show simulations where UH and FS were computed using Θ_0^2. The panels correspond to Scenarios 1-7. Simulated risk is presented in the top graph of each panel. Percentage risk compared with FM is presented in the bottom graph. The exclusion subspace Θ_0^1 produced similar results. These are presented in the appendix.

The simulated risk shows a large difference between shrinkage procedures. Risk for FS is universally below UH. In most cases it is substantially lower. We conclude that, in realistic scenarios for futures data, FS outperforms standard shrinkage choices by a wide margin. The simulations also show that UH and FS both outperform FM almost universally. UH does not suffer from the same estimation error issue as FH because its V and W result in simplifications of τ^* similar to those with FS.

2.5 Portfolio Choice for Futures Contracts

We now present an application of focused shrinkage to futures portfolio choice. Data was taken from the Pinnacle Continuous Futures Database. Trading with eight futures contracts was considered: Gold (GC), Euros (FX), British Pound (BN), Natural Gas (NG), Crude Oil (CL), 10 Year US Treasury Notes (TY), S&P 500 Mini (ES) and Goldman Sachs Commodity Index (GI). All of these contracts are traded in dollars. These contracts represent a wide variety of sectors. This allows for benefits from diversification which systematic trading rules are designed to capture. The data consists of monthly prices over the period 1997-2016. This window was used because the S&P 500 Mini contract started trading in 1997. Monthly data was used because of the well known problem of very low signal-to-noise ratio in financial data at higher frequencies.

Using this data, rolling window estimation with $T = 100$ observations was conducted. At each time period, the past $T = 100$ observations were used to estimate parameters. These estimates were then used to rebalance the portfolio using the trading rule $\mu_t(B, \Sigma)$. After rebalancing, we rolled one period forward and considered a new $T = 100$ window when rebalancing the next period. In order to begin this procedure, 100 previous observations were needed. In addition, as one factor is the previous 12-month return, an additional 12 periods were needed to start estimation. Therefore, despite the fact that our data set starts in late 1997, trading begins in 2007.

In our implementation of this rolling window estimation and trading rule, we considered the four estimators analyzed in the simulations. Estimation was conducted with the factors f_t chosen as described above. Both 1-month and 4-month lagged returns were considered for the first factor. It was assumed the investor starts with no position in any of the assets $\mu_{-1} = 0$. Rolling window estimation and portfolio rebalancing was conducted as described above. When changing asset positions across time periods, transaction costs $TC(\Delta\mu_t)$ were subtracted. The $\overline{\Sigma}$ matrix was required to compute these transaction costs. This matrix was estimated using the full model estimator over the whole sample period. The value $\gamma = 10^{-9}$ was used as explained above. The transaction cost parameter λ suggested by Garleanu and Pedersen (2013) is $\lambda = 10^{-6}$. This choice resulted in transaction costs that were unreasonably low. Because of this, $\lambda = 25 \times 10^{-6}$ was used to give more reasonable values.

The most common metric for describing the performance of a trading strategy is the

Sharpe ratio. This is the average return divided by its standard deviation over a trading period. The trading rule $\mu_t(B, \Sigma)$ was derived by Garleanu and Pedersen (2013) to result in good Sharpe ratios (while accounting for realistic transaction costs). Therefore, better estimation of $\mu_t(B, \Sigma)$ should result in a higher Sharpe ratio. For each estimator, the sample returns from implementing the trading rule were used to estimate Sharpe ratios over the trading period. These values were annualized (multiplied by $\sqrt{12}$) so they approximate yearly returns. Sharpe ratios are reported both with and without transaction costs. The turnover for each procedure is also reported. Turnover is the sum of absolute changes in all shares over the trading period. This gives a rough measure of transaction costs because changing positions incurs these costs. As transaction costs are often the difference between a successful and unsuccessful trading strategy, having low turnover is a desirable feature.

Table 1 reports our results. For the 4-month and 12-month factors and the exclusion subspace Θ_0^1, with no transaction costs, FS has a significantly larger Sharpe ratio compared to all competitors. The FM estimator ranks second. FH and UH preform poorly, with FH having negative average returns. The same patter appears when transaction costs are subtracted from returns. Notice that the turnover from FH is much larger than the other cases. This causes large transaction costs and poor performance. The same pattern holds with the pooled subspace Θ_0^2. These results are in line with our simulations.

For the 1-month and 12-month factors a similar pattern arises. For the exclusion subspace Θ_0^1, with no transaction costs, FH has the highest Sharpe ratio. However, FH again has high turnover. When transaction costs are subtracted, FS has the highest Sharpe ratio. The pooling subspace Θ_0^2 has similar results, with FS winning by a larger margin.

		1-month, 12-month Factors		4-month, 12-month Factors	
Subspace		Θ_0^1	Θ_0^2	Θ_0^1	Θ_0^2
Sharpe Ratio	FM	0.415	0.415	0.186	0.186
With Transaction Costs	UH	0.446	0.444	-0.030	0.017
	FH	-1.045	-0.334	-0.669	-1.061
	FS	0.495	0.525	0.424	0.451
Sharpe Ratio	FM	0.513	0.513	0.332	0.332
Without Transaction Costs	UH	0.537	0.535	0.112	0.160
	FH	0.732	0.122	-0.096	0.365
	FS	0.567	0.586	0.534	0.535
Turnover (in 10^8)	FM	10.017	10.017	9.043	9.043
	UH	9.571	9.621	8.782	8.854
	FH	127.685	922.934	184.978	111.174
	FS	7.483	5.935	7.835	6.006

Table 2.1 – Annualized Sharpe ratios and turnover for the FM, UH, FH and FS estimators.

2.6 Conclusion

In this paper, we proposed a focused shrinkage estimator which utilizes the functions of interest $x(\theta)$ when estimating the parameters θ. Risk properties for this procedure were derived and discussed. For our motivating example of portfolio choice, focused shrinkage was shown to have superior risk compared with standard alternatives. When applied to trading futures contracts, the procedure outperforms other methods as well. Most of these results are based on the choices $V = \Omega$ and $V = W^{-1}$. As the risk bounds are derived for arbitrary V, in future work it would be interesting to determine the optimal choice of V. An approximation to this could involve weighting between Ω and W^{-1}.

2.7 Appendix

2.7.1 Risk Simulations and Bounds

We first present simulated risk for shrinkage using the exclusion subspace Θ_0^1. This was done with $\overline{W} = D_\theta' D_\theta$ as presented in the main paper. The results are in figures 2.7.1-2.7.2. The same patter holds as before. The FH estimator again preforms very poorly and we only report the results for Scenario 1. The remaining panels in the Figures are simulation results

comparing FS, UH and FM. FS almost universally has lower risk than UH and FM. The difference is substantial for most parameters.

We next present simulated risk for $\overline{W} = (D_\theta' D_\theta + \alpha I)$. This is done for the FH estimator first. The results for the pooling subspace Θ_0^2 are presented in Figures 2.7.3-2.7.4. The results for the exclusion subspace Θ_0^1 are presented in Figures 2.7.5-2.7.6. In the pooling subspace, FS almost universally has better risk than FH. Furthermore, FH frequently has risk significantly above the FM estimator. For the exclusion subspace the results are similar. FH outperforms FS in some places, but for the large majority of parameters FS has lower risk. Again, FH has many parameters where it has much higher risk than FM. In contrast, for all parameters FS has lower risk than FM. We conclude that FS has superior risk properties for $\overline{W} = (D_\theta' D_\theta + \alpha I)$ as well as for $\overline{W} = D_\theta' D_\theta$.

Simulated risk comparisons for $\overline{W} = (D_\theta' D_\theta + \alpha I)$ with UH are now presented. The results for the pooling subspace Θ_0^2 are presented in Figures 2.7.7-2.7.8. The results for the exclusion subspace Θ_0^1 are presented in Figures 2.7.9-2.7.10. The picture is mixed. In the pooling subspace, FS has lower risk than UH for almost all parameters. For the exclusion subspace this is switched. Both shrinkage estimators have lower risk than the FM estimator for all parameters. As UH is not targeting the functions of interest, it is difficult to predict in which scenarios it will outperform focused estimation. As FS almost universally outperforms UH for the $\overline{W} = D_\theta' D_\theta$ case, and both estimators have similar risk properties in the $\overline{W} = (D_\theta' D_\theta + \alpha I)$ case, we conclude that FS is our preferred estimator. This is supported by our empirical application where FS gives better results than UH.

We now present the theoretical risk bounds derived in Theorem 2 of the main paper. These are shown for Scenarios 1-7 for the FM, FH and FS estimators. The results for Θ_0^2 are presented in Figure 2.7.11 and Θ_0^1 are presented in Figure 2.7.12. The bounds are derived assuming $\overline{W} = (D_\theta' D_\theta + \alpha I)$. This is not the case for UH or the simulated estimators with $\overline{W} = D_\theta' D_\theta$. Showing risk bounds in these case would require more theoretical derivations. For this reason, we exclude them. In addition, for the risk bounds presented we fix the value f_t in the function $\mu(\theta)$. These values are fixed at $f_{1t}^s = 2.5$ and $f_{2t}^s = 4$. This is in contrast to the simulations where f_t were random across realizations used to compute the estimators. Various f_t were tried and these values are representative of the results.

The contrast between the theoretical risk bounds and the simulated risk is striking. For the Θ_0^2 case, FH has a lower bound than FS for almost all values. For the Θ_0^1 case, the bounds

for FH and FS are almost indistinguishable. The upper risk bounds for both FH and FS are substantially below the FM risk. Considering the very poor performance of FH compared with FM and FS in simulations, the bounds are not very representative of actual risk. These results call into question the usefulness of these risk bounds in determining the risk properties of estimators. Further theoretical study is needed to better understand this large difference. We leave this to future research.

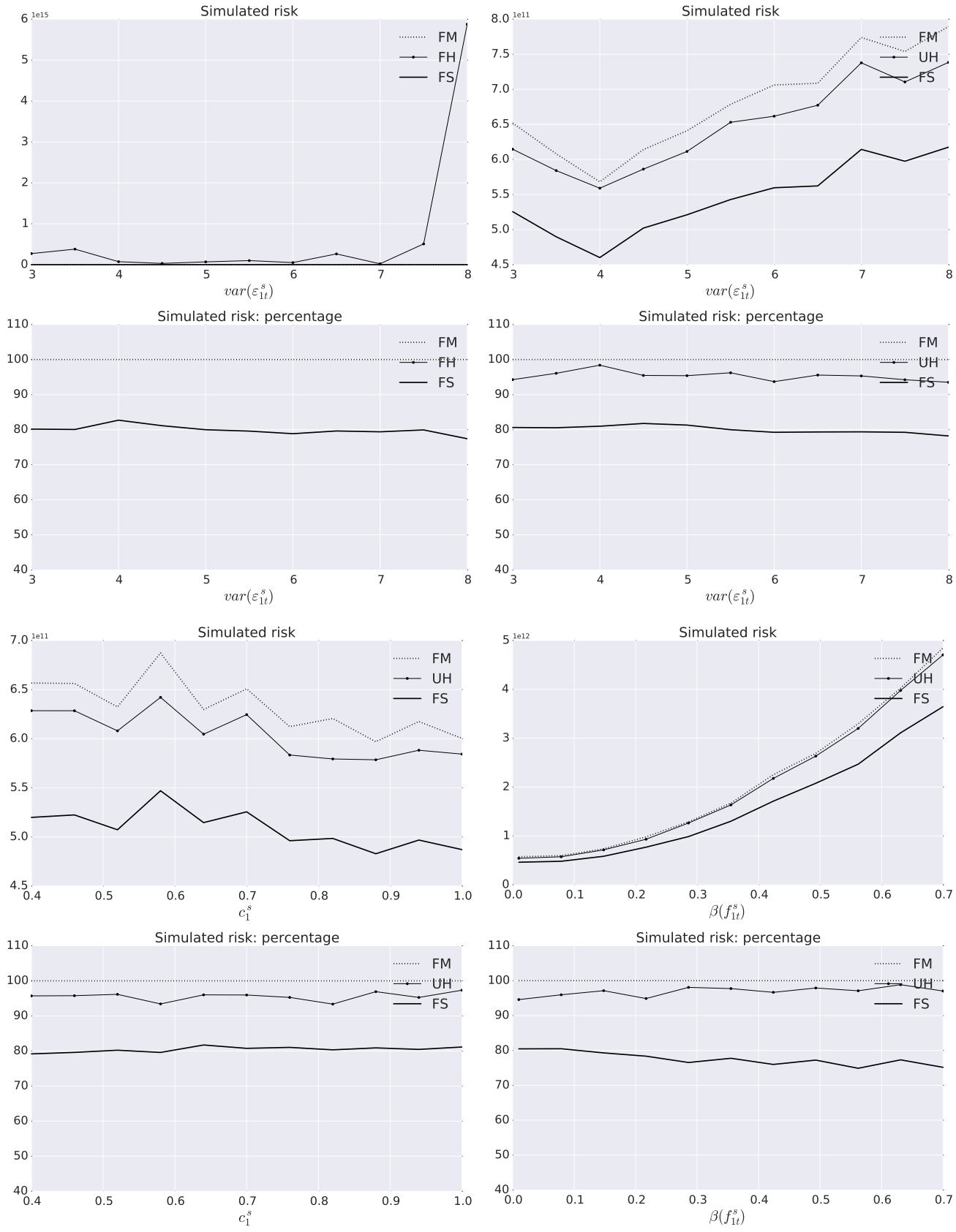

Figure 2.7.1 – Simulated risk, Θ_0^1 and $\overline{W} = D_\theta^{'} D_\theta$.

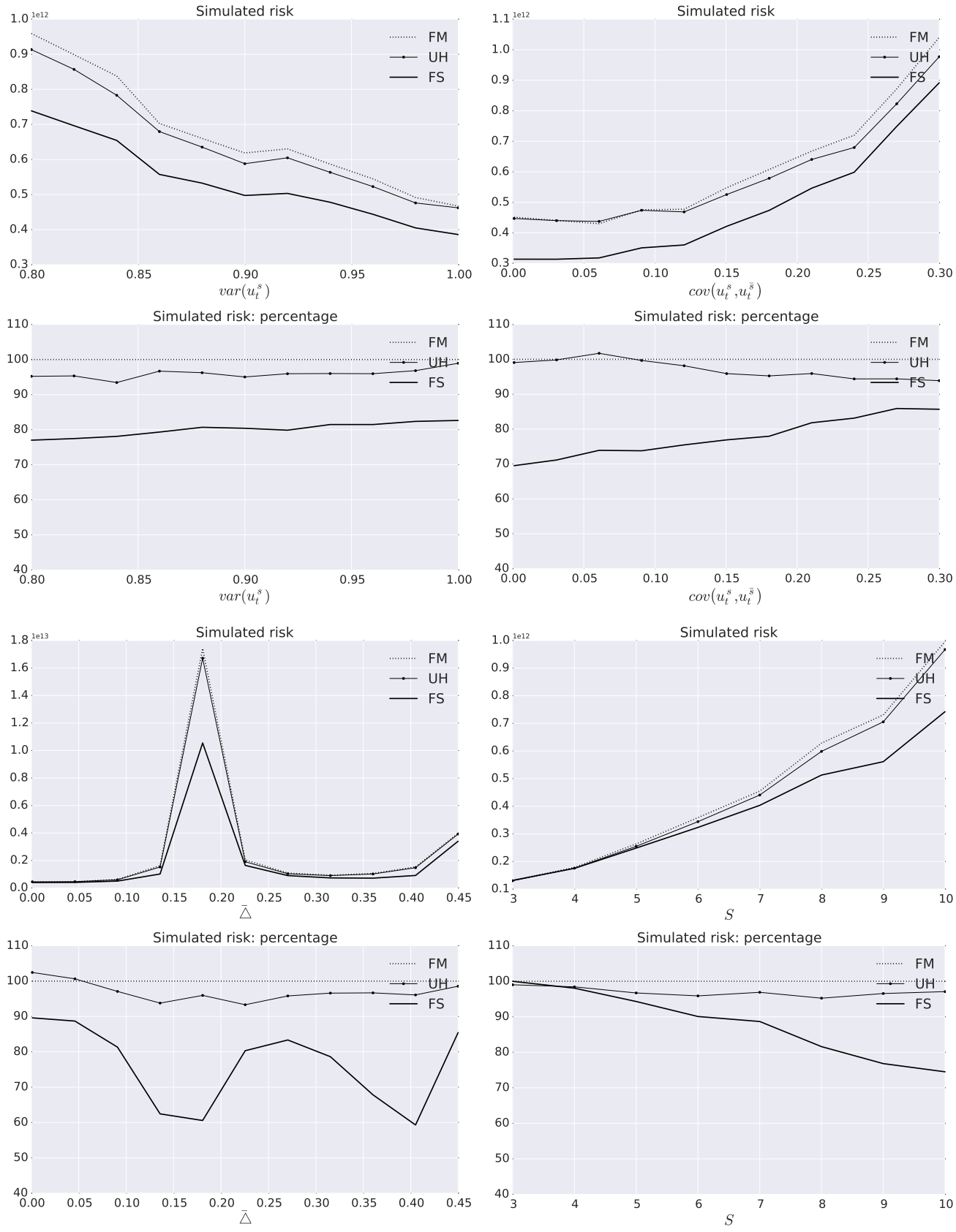

Figure 2.7.2 – Simulated risk, Θ_0^1 and $\overline{W} = D_\theta^{'} D_\theta$.

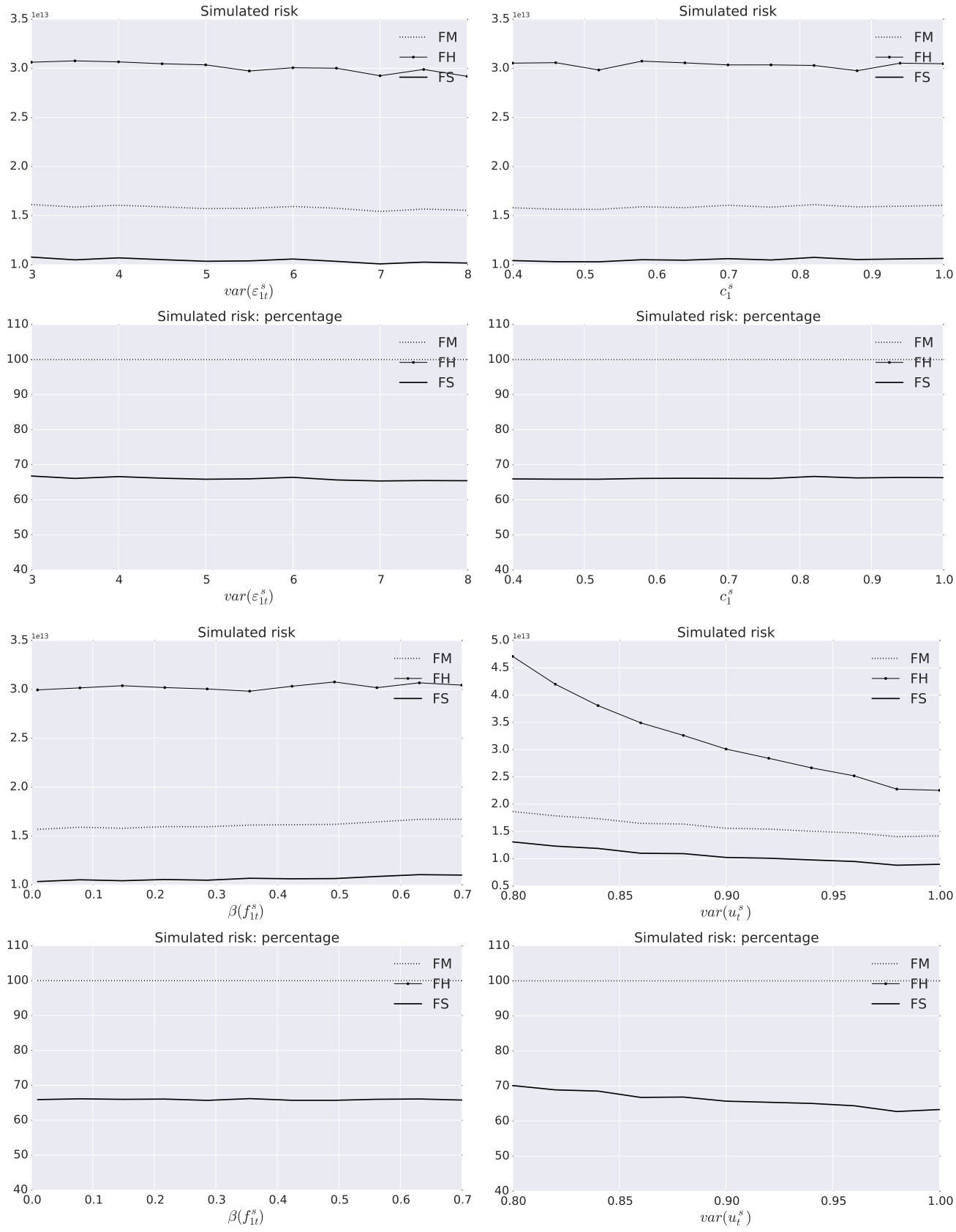

Figure 2.7.3 – Simulated risk, Θ_0^2 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

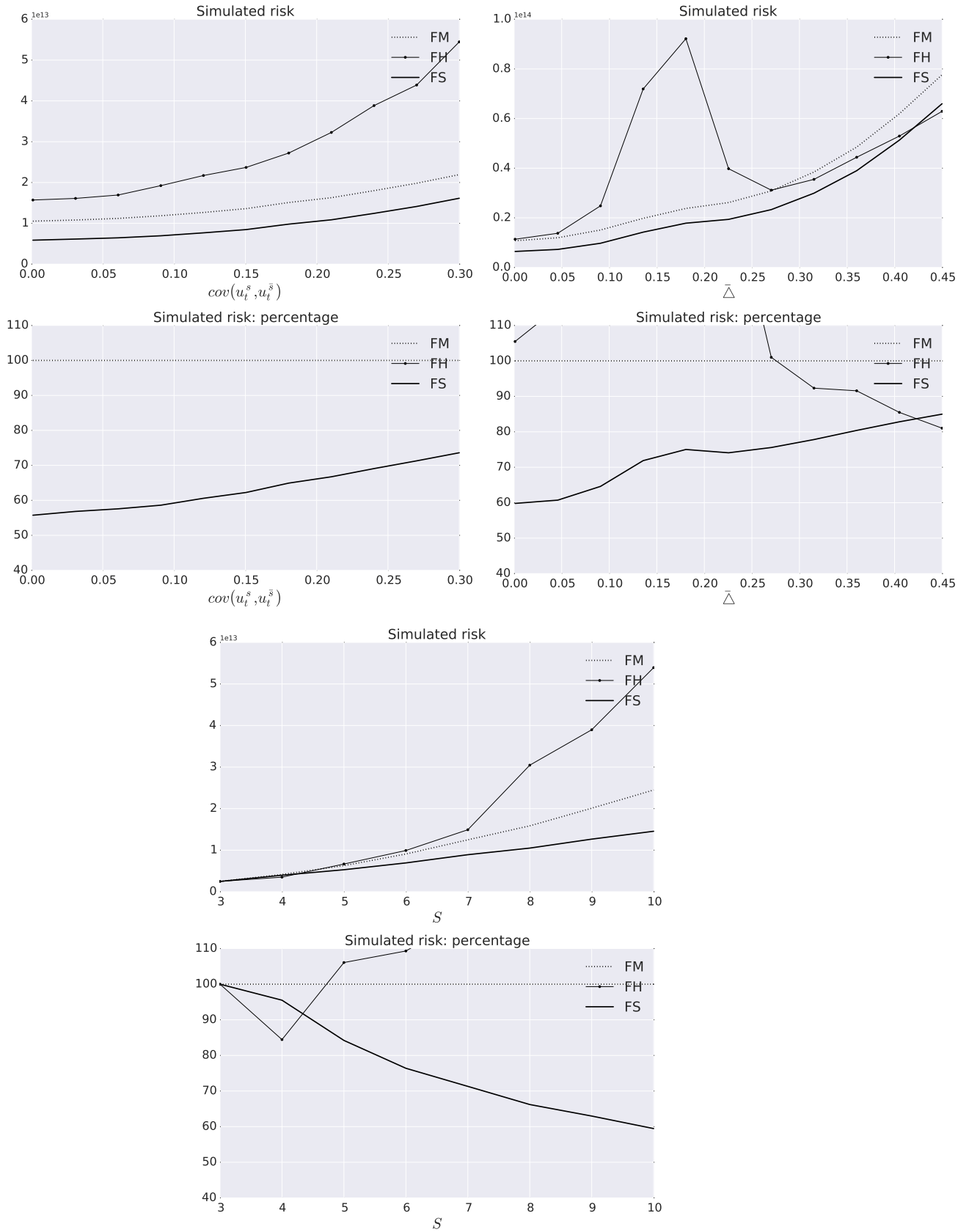

Figure 2.7.4 – Simulated risk, Θ_0^2 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

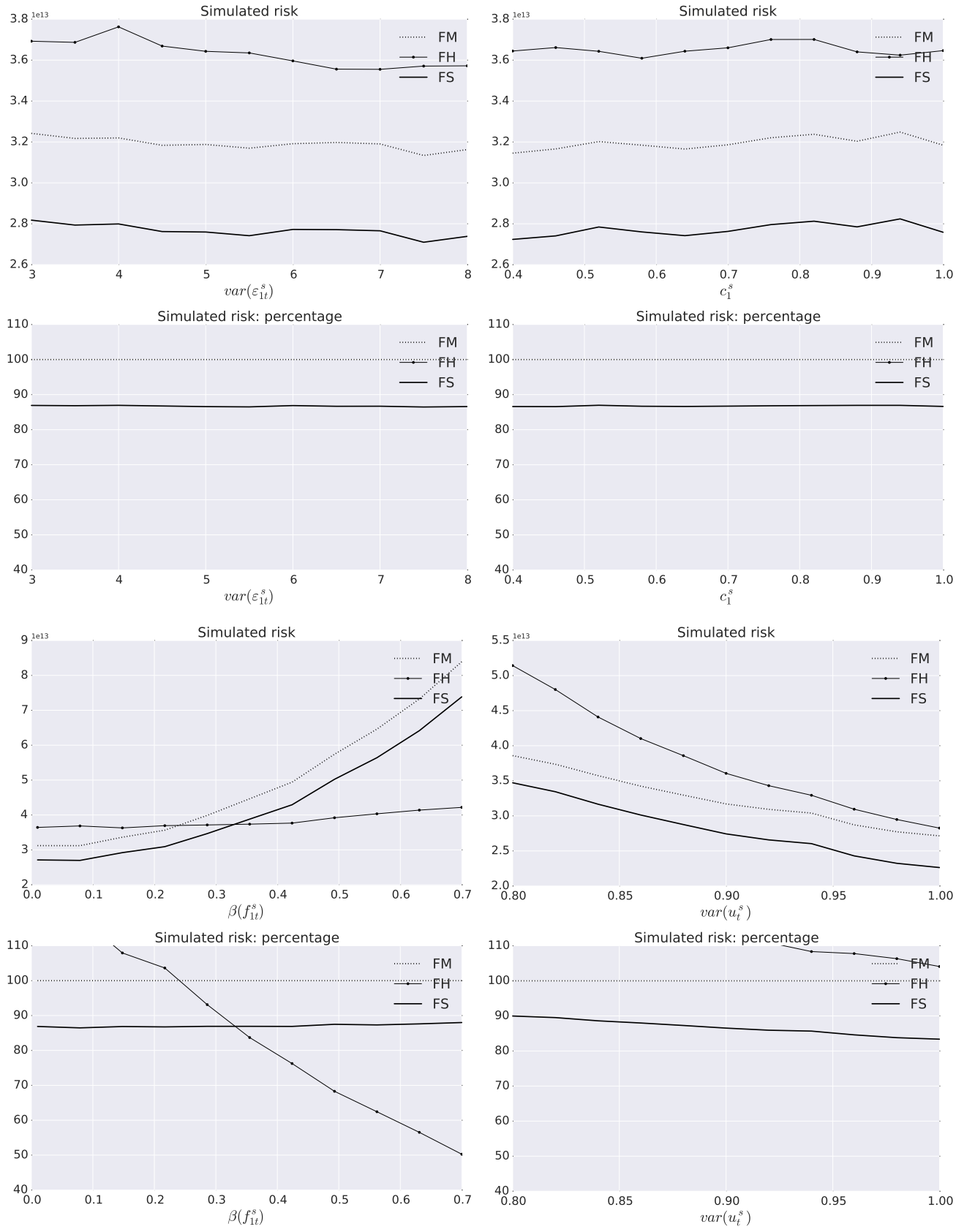

Figure 2.7.5 – Simulated risk, Θ_0^1 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

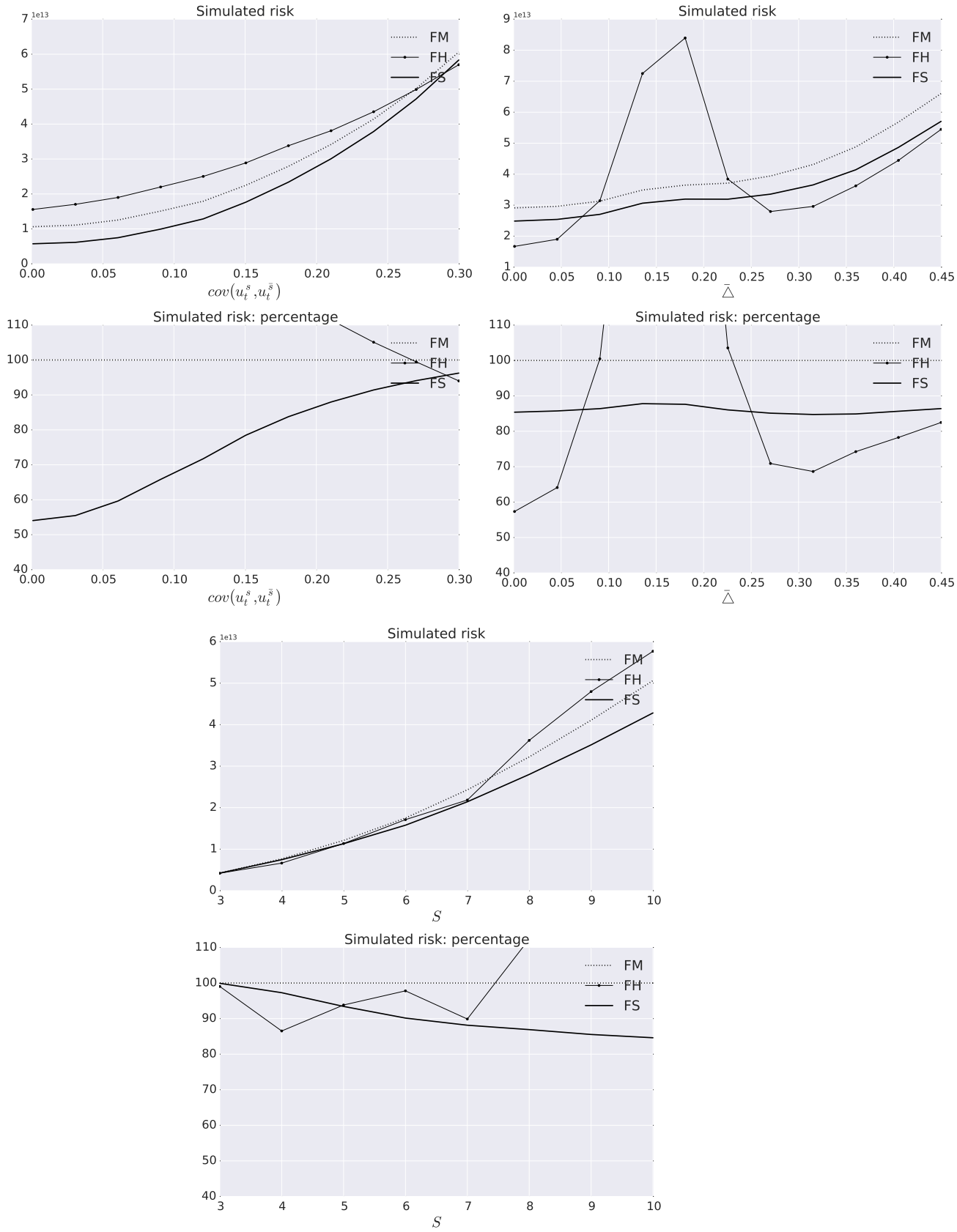

Figure 2.7.6 – Simulated risk, Θ_0^1 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

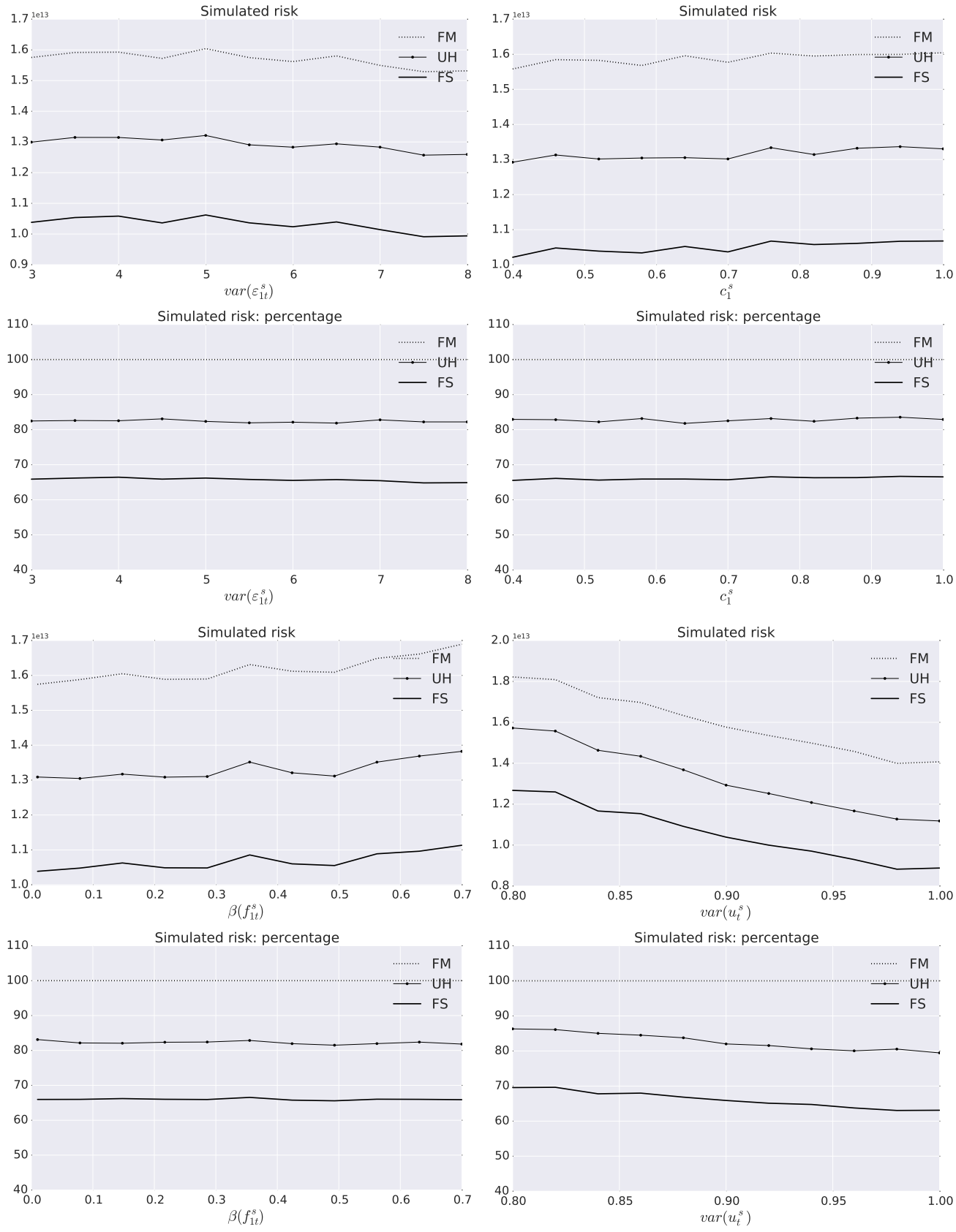

Figure 2.7.7 – Simulated risk, Θ_0^2 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

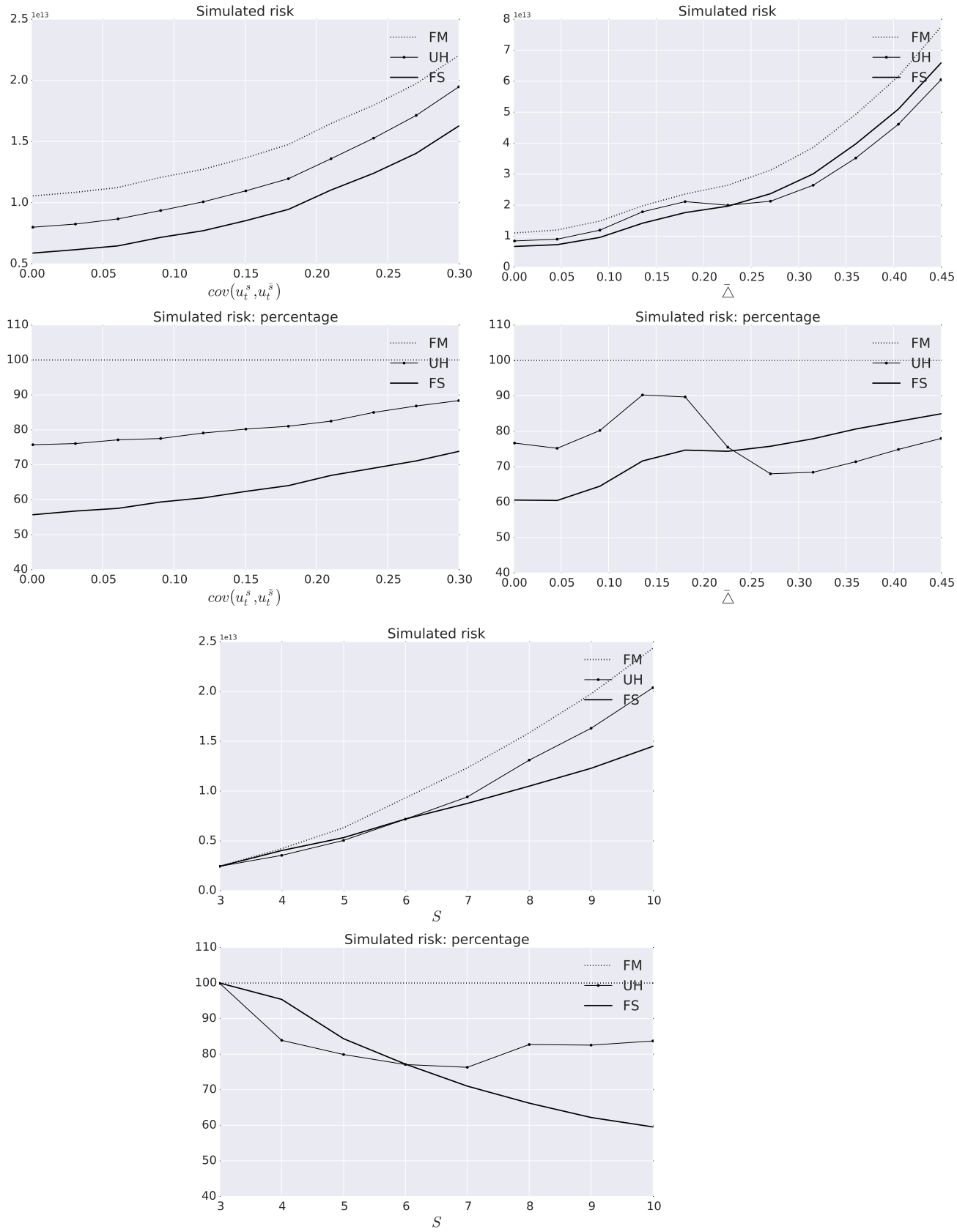

Figure 2.7.8 – Simulated risk, Θ_0^2 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

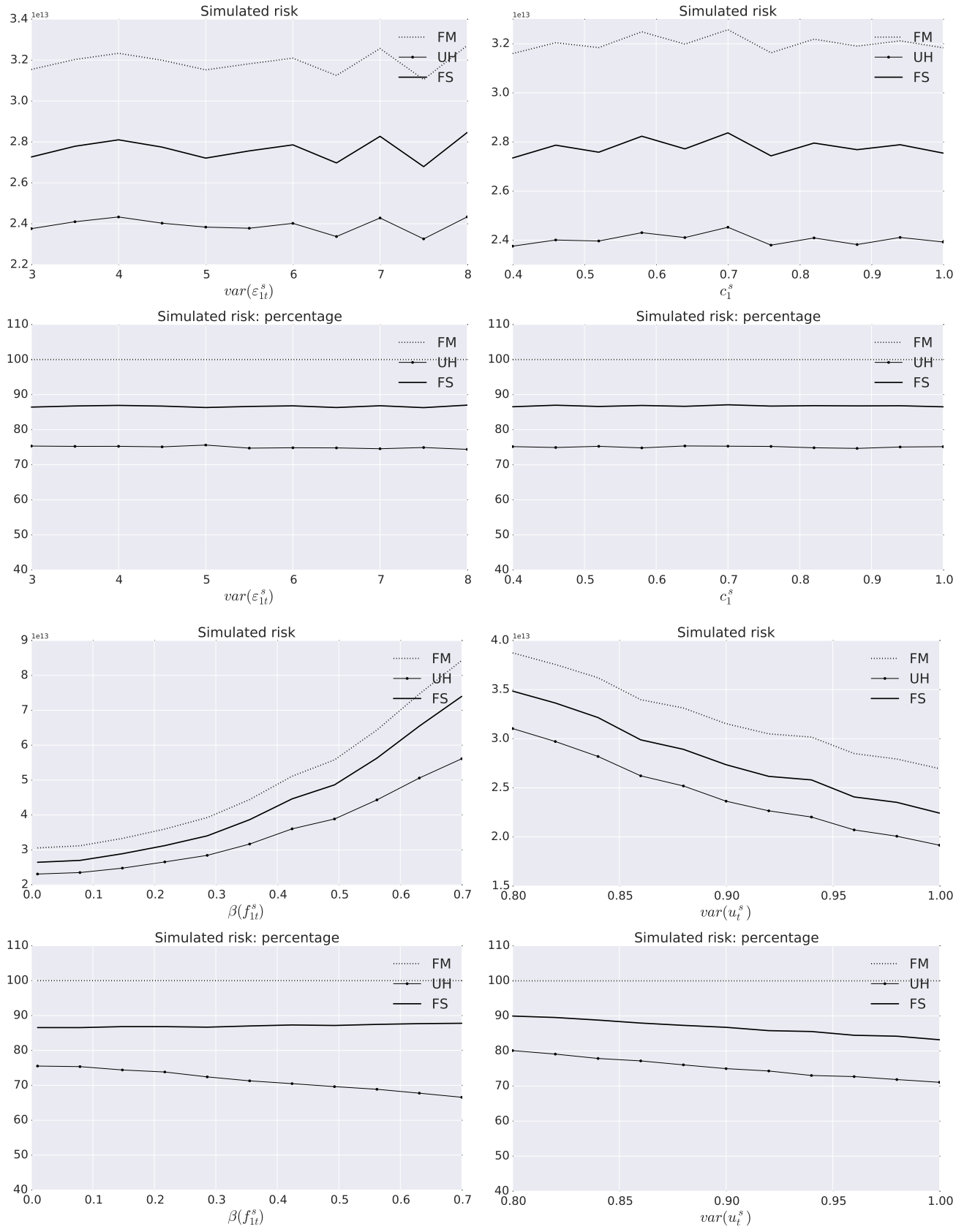

Figure 2.7.9 – Simulated risk, Θ_0^1 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

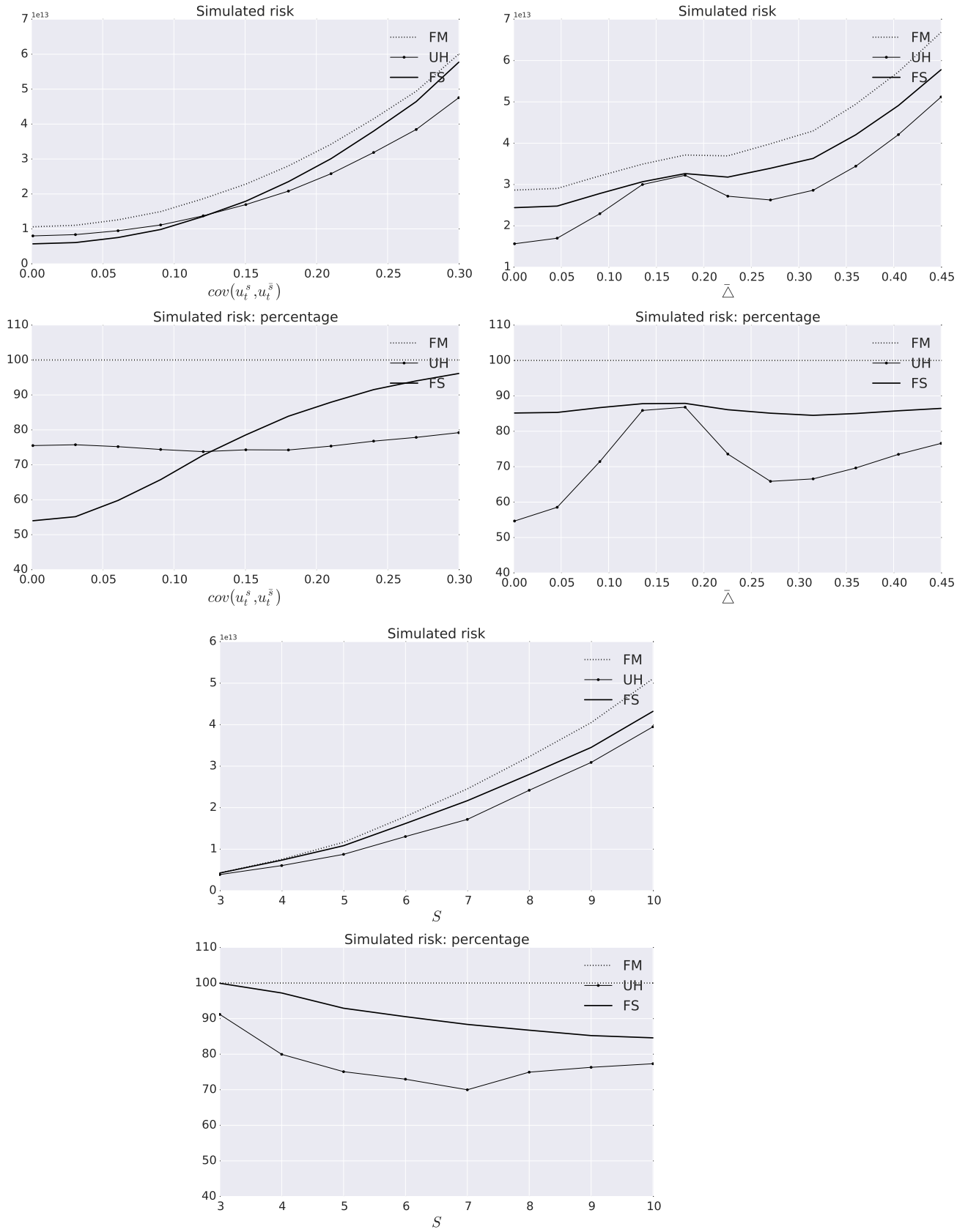

Figure 2.7.10 – Simulated risk, Θ_0^1 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

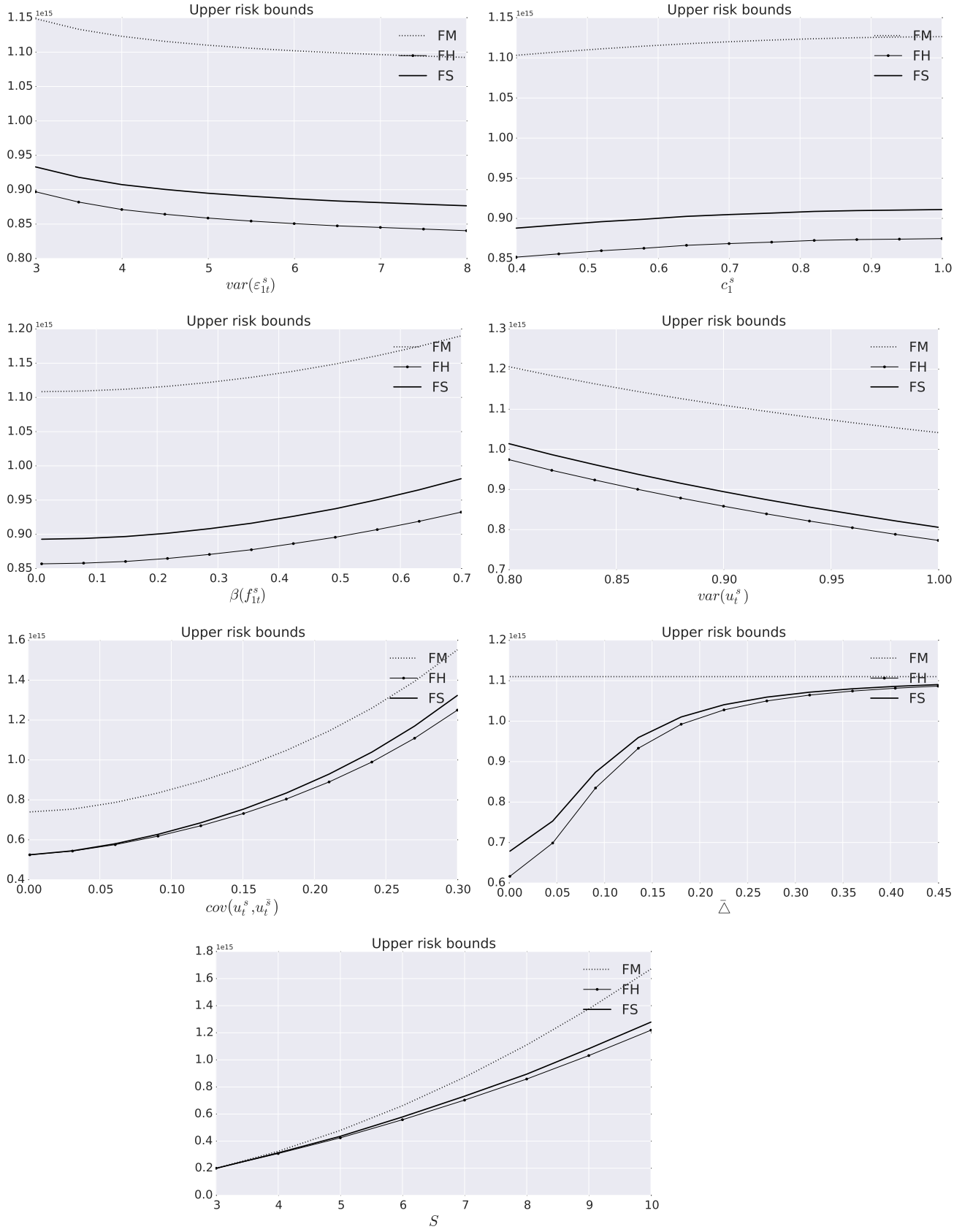

Figure 2.7.11 – Upper risk bounds, Θ_0^2 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

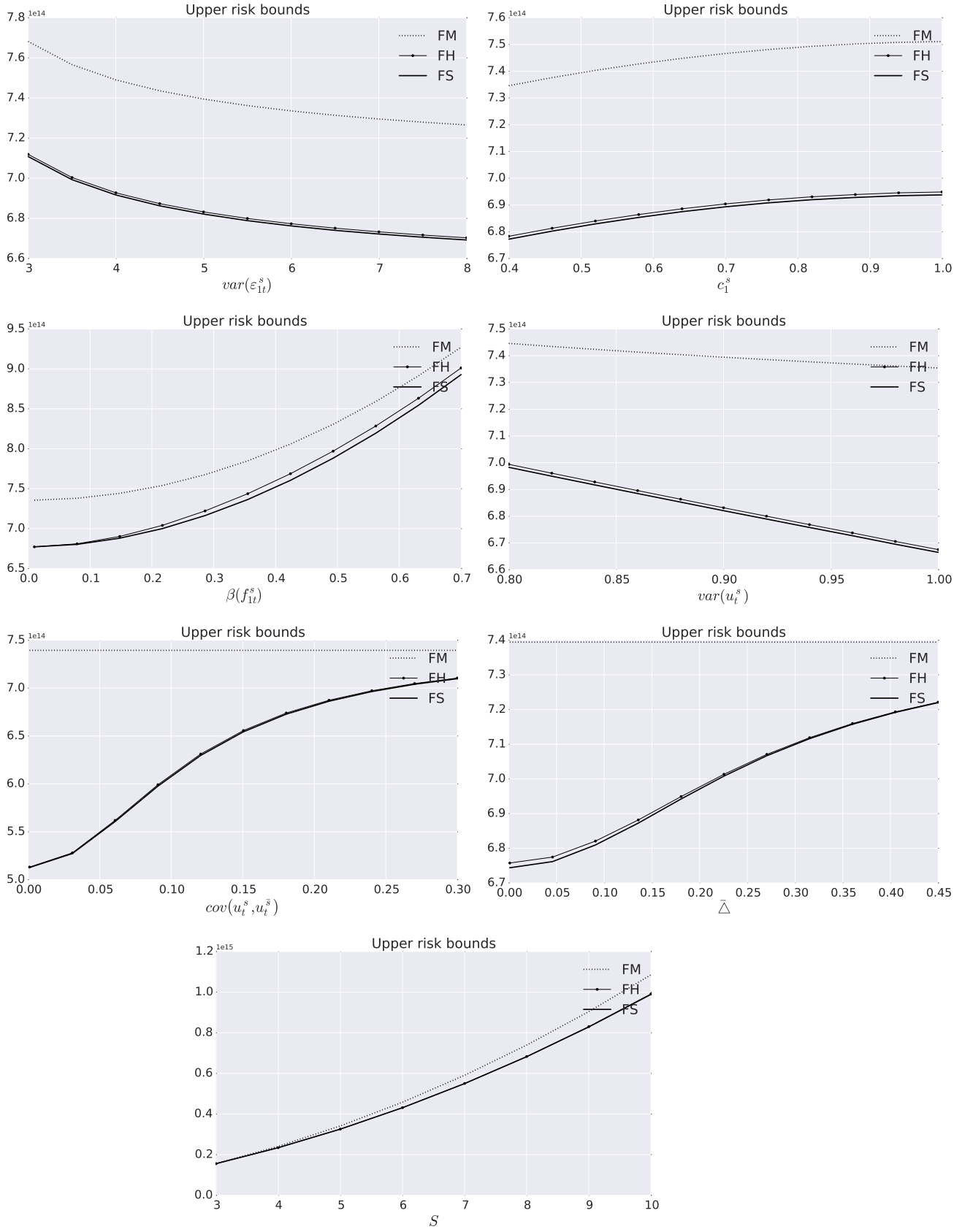

Figure 2.7.12 – Upper risk bounds, Θ_0^1 and $\overline{W} = (D_\theta^{'} D_\theta + \alpha I)$.

2.7.2 Proof of Theorem 1

First, consider $var(W, \tilde{\theta}^V)$. Define

$$\begin{aligned} D &= \Omega R(R'\Omega R)^{-1}R' - VR(R'VR)^{-1}R', \\ M &= I - \Omega R(R'\Omega R)^{-1}R'. \end{aligned}$$

.

Consider:

$$\begin{aligned} var(\tilde{\theta}^V) &= (I - P_V)\Omega(I - P_V)', \\ &= (M + D)\Omega(M + D)', \\ &= M\Omega M + M\Omega D' + D\Omega M' + D\Omega D', \\ &= var(\tilde{\theta}^\Omega) + M\Omega D' + D\Omega M' + D\Omega D'. \end{aligned}$$

The second term satisfies

$$\begin{aligned} M\Omega D' &= \left(I - \Omega R\left(R'\Omega R\right)^{-1}R'\right)\Omega\left(\Omega R\left(R'\Omega R\right)^{-1}R' - VR\left(R'VR\right)^{-1}R'\right)' \\ &= \left(\Omega - \Omega R\left(R'\Omega R\right)^{-1}R'\Omega\right)\left(\Omega R\left(R'\Omega R\right)^{-1}R' - VR\left(R'VR\right)^{-1}R'\right)' \\ &= \left(\Omega - \Omega R\left(R'\Omega R\right)^{-1}R'\Omega\right)\left(R\left(R'\Omega R\right)^{-1}R'\Omega - R\left(R'VR\right)^{-1}R'V\right) \\ &= \Omega R\left(R'\Omega R\right)^{-1}R'\Omega - \Omega R\left(R'VR\right)^{-1}R'V \\ &- \Omega R\left(R'\Omega R\right)^{-1}R'\Omega R\left(R'\Omega R\right)^{-1}R'\Omega + \Omega R\left(R'\Omega R\right)^{-1}R'\Omega R\left(R'VR\right)^{-1}R'V \\ &= \Omega R\left(R'\Omega R\right)^{-1}R'\Omega - \Omega R\left(R'VR\right)^{-1}R'V - \Omega R\left(R'\Omega R\right)^{-1}R'\Omega + \Omega R\left(R'\Omega R\right)^{-1}R'V \\ &= 0 \end{aligned}$$

As the third term is just the second transposed, we have

$$var(\tilde{\theta}^V) = var(\tilde{\theta}^\Omega) + D\Omega D',$$

which is

$$(I - P_V)\Omega(I - P_V)' = (I - P_\Omega)\Omega(I - P_\Omega)' + D\Omega D'.$$

Using the above equality we have:

$$var(W, \theta^V) - var(W, \theta^\Omega) \;=\; tr(W var(\theta^V)) - tr(W var(\theta^\Omega)) \;=\; tr(WD\Omega D^{'})$$

We note that W and $D\Omega D^{'}$ are PSD matrices. Using the inequality (1) from Fang et al. (1994) we have

$$tr(WD\Omega D^{'}) \;\geq\; \lambda_{min}(W) tr(D\Omega D^{'}) \;\geq\; 0.$$

The last line follows because the eigenvalues of a PSD matrix are greater than or equal to zero. In addition, the trace of a matrix is the sum of its eigenvalues. When $V = \Omega$, $D = 0$. Therefore, $var(W, \theta^V)$ is minimized if $V = \Omega$.

Now we consider $bias^2(W, \theta^V)$.

$$bias^2(W, \theta^V) \;=\; \Delta^{'} R(R^{'}VR)^{-1} R^{'} VWVR(R^{'}VR)^{-1} R^{'}\Delta \;=\; \Delta^{'} B_V \Delta$$

where B_V is the B matrix with a particular V. Define the matrices

$$\begin{aligned} J \;&=\; R(R^{'}VR)^{-1} R^{'} VW^{1/2}, \\ L \;&=\; W^{-1/2} R(R^{'}W^{-1}R)^{-1} R^{'} W^{-1/2}, \end{aligned}$$

and consider

$$\begin{aligned} B_V - B_{W^{-1}} \;&=\; R(R^{'}VR)^{-1} \left\{ R^{'}VWVR - [(R^{'}VR)(R^{'}W^{-1}R)^{-1}(R^{'}VR)] \right\} (R^{'}VR)^{-1} R^{'}, \\ &=\; R(R^{'}VR)^{-1} R^{'} VW^{1/2} \left\{ I - W^{-1/2} R(R^{'}W^{-1}R)^{-1} R^{'} W^{-1/2} \right\} W^{1/2} VR(R^{'}VR)^{-1} R^{'}, \\ &=\; J \left\{ I - L \right\} J^{'}. \end{aligned}$$

L is symmetric and idempotent. Therefore, $(I - L)$ is symmetric and idempotent and $(I - L) = (I - L)(I - L)^{'}$. Therefore,

$$\begin{aligned} B_V - B_{W^{-1}} \;&=\; J[I - L]J^{'} \\ &=\; J(I - L)(I - L)^{'} J^{'} \\ &=\; (J(I - L))(J(I - L))^{'}. \end{aligned}$$

As a result, $B_V - B_{W^{-1}}$ is PSD. Therefore,

$$bias^2(W, \theta^V) - bias^2(W, \theta^{W^{-1}}) \;=\; \Delta^{'}(B_V - B_{W^{-1}})\Delta \;\geq\; 0$$

Thus, $V = W^{-1}$ minimizes $bias^2(W, \theta^V)$.

2.7.3 Proof of Theorem 2

The risk of the full model estimator θ is easily seen to be $tr(W\Omega)$. Arguing as in Hansen's proof of his Theorem 2, we have

$$\begin{aligned}\rho(W, \theta^*, \tau) \;\leq\; & tr(W\Omega) + \tau^2 E\left\{[(1/((Z+\Delta)^{'}B(Z+\Delta)))]\right\}, \\ & -2\tau E\left\{(((Z+\Delta)^{'}R(R^{'}VR)^{-1}R^{'}VWZ)/((Z+\Delta)^{'}B(Z+\Delta)))\right\}.\end{aligned} \tag{2.22}$$

Define the function $\eta : \mathbb{R} \to \mathbb{R}$

$$\eta(x) = \left(\frac{1}{x^{'}Bx}\right)x.$$

Applying Stein's Lemma as in Hansen (2016), but accounting for the fact that $V \neq \Omega$, leads to the two terms

$$\begin{aligned} & E\left\{(((Z+\Delta)^{'}R(R^{'}VR)^{-1}R^{'}VWZ)/((Z+\Delta)^{'}B(Z+\Delta)))\right\} \\ =\; & E\left\{tr[(\partial/(\partial x))\eta(Z+\Delta)^{'}R(R^{'}VR)^{-1}R^{'}VW\Delta]\right\} \\ =\; & E\left\{tr[((R(R^{'}VR)^{-1}R^{'}VW\Omega)/((Z+\Delta)^{'}B(Z+\Delta)))]\right\} \\ & -2E\left\{tr[((B(Z+\Delta)(Z+\Delta)^{'}R(R^{'}VR)^{-1}R^{'}VW\Omega)/(((Z+\Delta)^{'}B(Z+\Delta))^2))]\right\}\end{aligned} \tag{2.23}$$

We use the cyclical property of the trace for:

$$\begin{aligned} & tr[B(Z+\Delta)(Z+\Delta)^{'}R(R^{'}VR)^{-1}R^{'}VW\Omega] \\ =\; & tr[(Z+\Delta)^{'}R(R^{'}VR)^{-1}R^{'}VW\Omega B(Z+\Delta)]\end{aligned}$$

The interior of this is:

$$M \;=\; R(R^{'}VR)^{-1}R^{'}VW\Omega B \;=\; R(R^{'}VR)^{-1}R^{'}VW\Omega R(R^{'}VR)^{-1}R^{'}VWVR(R^{'}VR)^{-1}R^{'}$$

so we need an upper bound on:

$$E\left[\frac{tr[(Z+\Delta)^{'}M(Z+\Delta)}{((Z+\Delta)^{'}B(Z+\Delta))^2}\right]$$

We can use a Cauch-Schwartz type inequality to bound $|(Z+\Delta)^{'}M(Z+\Delta)|$. Define the matrix

$$F \;=\; R(R^{'}VR)^{-1}R^{'}$$

which is symmetric and PSD. Define also:

$$\begin{aligned} x^{'} \;&=\; R(R^{'}VR)^{-1}R^{'}VW\Omega, \\ y \;&=\; VWVR(R^{'}VR)^{-1}R^{'}. \end{aligned}$$

Using a Cauch-Schwartz type inequality (Abadir and Magnus (2005) Exercise 12.2) we have:

$$\begin{aligned} |(Z+\Delta)^{'}M(Z+\Delta)| \;&=\; |(Z+\Delta)^{'}x^{'}Fy(Z+\Delta)| \\ &\leq\; \sqrt{([(Z+\Delta)^{'}x^{'}Fx(Z+\Delta)]\,[(Z+\Delta)^{'}y^{'}Fy(Z+\Delta)])} \end{aligned} \tag{2.24}$$

Define

$$\begin{aligned} B_1^{'} \;&=\; R(R^{'}VR)^{-1}R^{'}VW^{1/2} \\ K \;&=\; W^{'1/2}\Omega R(R^{'}VR)^{-1}R^{'}\Omega W^{1/2} \end{aligned}$$

and note that

$$B_1^{'}B_1 \;=\; B.$$

We write:

$$((Z+\Delta)^{'}x^{'}Fx(Z+\Delta)) \;=\; (Z+\Delta)^{'}B_1^{'}KB_1(Z+\Delta)$$

As the center term is symmetric, we can apply the a standard inequality to get:

$$\begin{aligned} (Z+\Delta)^{'}B_1^{'}KB_1(Z+\Delta) \;&\leq\; (Z+\Delta)^{'}B_1^{'}B_1(Z+\Delta)\lambda_{max}(K) \\ &=\; (Z+\Delta)^{'}B(Z+\Delta)\lambda_{max}(K) \end{aligned}$$

A similar argument gives

$$((Z+\Delta)^{'}y^{'}Fy(Z+\Delta)) \;\leq\; (Z+\Delta)^{'}B(Z+\Delta)\lambda_{max}(A^{*})$$

These results give a bound on (2.24):

$$(Z+\Delta)^{'}M(Z+\Delta) \;\leq\; (Z+\Delta)^{'}B(Z+\Delta)\lambda_{max}^{1/2}(A^{*})\lambda_{max}^{1/2}(K)$$

This gives the final bound of:

$$\begin{aligned} & E\left\{tr[((B(Z+\Delta)(Z+\Delta)^{'}R(R^{'}VR)^{-1}R^{'}VW\Omega)/(((Z+\Delta)^{'}B(Z+\Delta))^{2}))]\right\} \\ \leq\; & E\left\{((\lambda_{max}^{1/2}(A^{*})\lambda_{max}^{1/2}(K))/((Z+\Delta)^{'}B(Z+\Delta)))\right\}. \end{aligned}$$

Therefore, the term (2.23) is bounded below by:

$$(2.23) \geq E\left\{\frac{tr(A)-2\lambda_{max}^{1/2}(A^{*})\lambda_{max}^{1/2}(K)}{((Z+\Delta)^{'}B(Z+\Delta)))}\right\}.$$

This means that (2.22) is bounded above by:

$$tr(W\Omega)-\tau E\left\{\frac{2[tr(A)-2\lambda_{max}^{1/2}(A^{*})\lambda_{max}^{1/2}(K)]-\tau}{(Z+\Delta)^{'}B(Z+\Delta)}\right\}$$

All of the statements of the proposition follow from this.

2.7.4 Proof of Theorem 3

We use the following results in our proof of Theorem 3.

RES1. (Problem 18, page 423, Horn and Johnson, Matrix Analysis). Let A and B be Hermitian positive definite we have

$$\lambda_{max}(AB) \;\leq\; \lambda_{max}(A)\,\lambda_{max}(B)$$

and

$$\lambda_{min}(AB) \;\geq\; \lambda_{min}(A)\,\lambda_{min}(B)$$

RES2. (Inequality for the Trace of Matrix Product, Fang et al. (1994)). For any positive semidefinite matrices A and B

$$\lambda_{min}(A)\, tr(B) \;\leq\; tr(AB) \;\leq\; \lambda_{max}(A)\, tr(B)$$

We bound each of the components of the bound in Theorem 2. This gives us an upper bound on the previous risk bound.

Bounding $\lambda_{\max}(\mathbf{A}^*)$ and $\lambda_{\max}(\mathbf{K})$

Using RES1, we have the following inequalities:

$$\begin{aligned}
\lambda_{max}(A^*) &= \lambda_{max}\left(W^{'\frac{1}{2}} VR\left(R'VR\right)^{-1} R'VW^{\frac{1}{2}}\right) \\
&= \lambda_{max}\left(VR\left(R'VR\right)^{-1} R'VW\right) && \text{(using PSDness)} \\
&\leq \lambda_{max}\left(VR\left(R'VR\right)^{-1} R'V\right)\lambda_{max}(W) && \text{(using RES1)} \\
&\leq \lambda_{max}\left(R\left(R'VR\right)^{-1} R'\right)\lambda_{max}(VV)\,\lambda_{max}(W) \\
&\leq \lambda_{max}\left(R\left(R'VR\right)^{-1} R'\right)\lambda^2_{max}(V)\,\lambda_{max}(W)
\end{aligned}$$

Now inspecting the sandwich-type matrix $R\left(R'VR\right)^{-1}R'$, we have the following:

$$\begin{aligned}
\lambda(V) &= \{\lambda_{min}(V) \leq \ldots \leq \lambda_{max}(V)\} \\
\lambda\left(R'VR\right) &= \left\{\lambda_{min}\left(R'VR\right) \leq \ldots \leq \lambda_{max}\left(R'VR\right)\right\} \\
&\quad \text{(using Poincare separation theorem (Horn and Johnson (2012))} \\
&\quad \text{with } \left\{\lambda_{min}(V) \leq \lambda_{min}\left(R'VR\right) \leq \ldots \leq \lambda_{max}\left(R'VR\right) \leq \lambda_{max}(V)\right\} \\
&\quad \text{(positive } \lambda_i\text{)} \\
\lambda\left(\left(R'VR\right)^{-1}\right) &= \left\{\frac{1}{\lambda_{max}(R'VR)} \leq \ldots \leq \frac{1}{\lambda_{min}(R'VR)}\right\} \\
&\quad \text{with } \left\{\frac{1}{\lambda_{max}(V)} \leq \frac{1}{\lambda_{max}(R'VR)} \leq \ldots \leq \frac{1}{\lambda_{min}(R'VR)} \leq \frac{1}{\lambda_{min}(V)}\right\} \\
\lambda\left(R\left(R'VR\right)^{-1}R'\right) &= \left\{0 \leq \ldots \leq \frac{1}{\lambda_{min}(R'VR)}\right\} \\
&\quad \text{with } \left\{0 \leq \frac{1}{\lambda_{max}(V)} \leq \frac{1}{\lambda_{max}(R'VR)} \leq \ldots \leq \frac{1}{\lambda_{min}(R'VR)} \leq \frac{1}{\lambda_{min}(V)}\right\}
\end{aligned} \tag{2.25}$$

The last equality follows since matrix $R\left(R^{'}VR\right)^{-1}R^{'}$ has dimension of matrix V and contains the elements of $\left(R^{'}VR\right)^{-1}$ and zeros on the rows and columns that are excluded by R. Consequently, $R\left(R^{'}VR\right)^{-1}R^{'}$ can be transformed into the block-diagonal matrix using permutation matrix P

$$P^{'}\left[R\left(R^{'}VR\right)^{-1}R^{'}\right]P \;=\; \begin{bmatrix} \left(R^{'}VR\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and since this partition in case of the exclusion restriction is such that $P^{'}P = I$, this resulting matrix has the eigenvalues of $\left(R^{'}VR\right)^{-1}$ and $\mathbf{0}$ implying the final bound. This in turn implies that

$$\begin{aligned} \lambda_{max}\left(A^{*}\right) \;&\leq\; \lambda_{max}\left(R\left(R^{'}VR\right)^{-1}R^{'}\right)\lambda_{max}^{2}\left(V\right)\lambda_{max}\left(W\right) \\ &\leq\; \frac{1}{\lambda_{min}\left(V\right)}\lambda_{max}^{2}\left(V\right)\lambda_{max}\left(W\right) \end{aligned}$$

Next, analyzing $\lambda_{max}\left(K\right)$:

$$\begin{aligned} \lambda_{max}\left(K\right) \;&=\; \lambda_{max}\left(W^{'\frac{1}{2}}\Omega R\left(R^{'}VR\right)^{-1}R^{'}\Omega W^{\frac{1}{2}}\right) \\ &=\; \lambda_{max}\left(\Omega R\left(R^{'}VR\right)^{-1}R^{'}\Omega W\right) && \text{(using PSDness)} \\ &\leq\; \lambda_{max}\left(\Omega R\left(R^{'}VR\right)^{-1}R^{'}\Omega\right)\lambda_{max}\left(W\right) && \text{(using RES1)} \\ &\leq\; \lambda_{max}\left(R\left(R^{'}VR\right)^{-1}R^{'}\right)\lambda_{max}\left(\Omega\Omega\right)\lambda_{max}\left(W\right) \\ &\leq\; \lambda_{max}\left(R\left(R^{'}VR\right)^{-1}R^{'}\right)\lambda_{max}^{2}\left(\Omega\right)\lambda_{max}\left(W\right) \\ &\leq\; \frac{1}{\lambda_{min}\left(V\right)}\lambda_{max}^{2}\left(\Omega\right)\lambda_{max}\left(W\right) && \text{(using (2.25) above)} \end{aligned}$$

Finally, combining these two bounds we have:

$$\begin{aligned} \lambda^{\frac{1}{2}}\left(A^{*}\right)\lambda^{\frac{1}{2}}\left(K\right) \;&\leq\; \left[\frac{1}{\lambda_{min}\left(V\right)}\lambda_{max}^{2}\left(V\right)\lambda_{max}\left(W\right)\right]^{\frac{1}{2}}\left[\frac{1}{\lambda_{min}\left(V\right)}\lambda_{max}^{2}\left(\Omega\right)\lambda_{max}\left(W\right)\right]^{\frac{1}{2}} \\ &=\; \frac{\lambda_{max}\left(V\right)}{\lambda_{min}\left(V\right)}\lambda_{max}\left(\Omega\right)\lambda_{max}\left(W\right) \end{aligned}$$

Bounding $(\mathbf{Z}+\mathbf{\Delta})^{'}\mathbf{B}(\mathbf{Z}+\mathbf{\Delta})$

We can also further loosen the bound for the denominator part using the max eigenvalue property (Raleigh-Ritz, p. 176):

$$\begin{aligned}
(Z+\Delta)^{'}B(Z+\Delta) &= (Z+\Delta)^{'}R\left(R^{'}VR\right)^{-1}R^{'}VWVR\left(R^{'}VR\right)^{-1}R^{'}(Z+\Delta) \\
&= (Z+\Delta)^{'}\left\{R\left(R^{'}VR\right)^{-1}R^{'}V^{\frac{1}{2}}\right\}\left\{V^{\frac{1}{2}}WV^{\frac{1}{2}}\right\}\left\{V^{\frac{1}{2}}R\left(R^{'}VR\right)^{-1}R^{'}\right\}(Z+\Delta) \\
&\leq (Z+\Delta)^{'}R\left(R^{'}VR\right)^{-1}R^{'}V^{\frac{1}{2}}V^{\frac{1}{2}}R\left(R^{'}VR\right)^{-1}R^{'}(Z+\Delta)\lambda_{max}\left(V^{\frac{1}{2}}WV^{\frac{1}{2}}\right) \\
&= (Z+\Delta)^{'}R\left(R^{'}VR\right)^{-1}R^{'}(Z+\Delta)\lambda_{max}\left(V^{\frac{1}{2}}WV^{\frac{1}{2}}\right) \\
&\leq (Z+\Delta)^{'}R\left(R^{'}VR\right)^{-1}R^{'}(Z+\Delta)\lambda_{max}(V)\lambda_{max}(W) \\
&\leq (Z+\Delta)^{'}(Z+\Delta)\lambda_{max}\left(R\left(R^{'}VR\right)^{-1}R^{'}\right)\lambda_{max}(V)\lambda_{max}(W) \\
&\leq (Z+\Delta)^{'}(Z+\Delta)\frac{1}{\lambda_{min}(V)}\lambda_{max}(V)\lambda_{max}(W)
\end{aligned}$$

The value of tr $\left(\bar{\mathbf{A}}\right)$ for different choices of V

I. If $V=W^{-1}$, we have

$$\begin{aligned}
tr\left(\bar{A}\right) &= tr\left(\left(R^{'}W^{-1}R\right)^{-1}R^{'}W^{-1}W\Omega R\right) \\
&= tr\left(\left(R^{'}W^{-1}R\right)^{-1}R^{'}\Omega R\right) \\
&\geq \lambda_{min}\left(\left(R^{'}W^{-1}R\right)^{-1}\right)tr\left(R^{'}\Omega R\right) \quad \text{(using RES2.)} \\
&\geq \frac{1}{\lambda_{max}(W^{-1})}tr\left(R^{'}\Omega R\right) \qquad \text{(from (2.25))} \\
&= \frac{1}{\lambda_{max}(W^{-1})}tr\left(\Omega RR^{'}\right) \qquad \text{(cyclical property)} \\
&\geq \frac{1}{\lambda_{max}(W^{-1})}\lambda_{min}(\Omega)\,tr\left(RR^{'}\right) \\
&= \lambda_{min}(W)\lambda_{min}(\Omega)\,Rank\left(RR^{'}\right) \\
&= \lambda_{min}(\Omega)\lambda_{min}(W)\,Rank\left(RR^{'}\right)
\end{aligned}$$

II. If $V = \Omega$, we have

$$\begin{aligned}
tr\left(\bar{A}\right) &= tr\left(\left(R'\Omega R\right)^{-1} R'\Omega W\Omega R\right) \\
&= tr\left(\left(R'\Omega R\right)^{-1} R'\Omega W\Omega R\right) \\
&\geq \lambda_{min}\left(\left(R'\Omega R\right)^{-1}\right) tr\left(R'\Omega W\Omega R\right) && \text{(using RES2)} \\
&\geq \frac{1}{\lambda_{max}(\Omega)} tr\left(R'\Omega W\Omega R\right) && \text{(from (2.25))} \\
&= \frac{1}{\lambda_{max}(\Omega)} tr\left(\Omega W\Omega RR'\right) && \text{(cyclical property)} \\
&\geq \frac{1}{\lambda_{max}(\Omega)} \lambda_{min}(\Omega W\Omega)\, tr\left(RR'\right) && \text{(using RES2)} \\
&= \frac{1}{\lambda_{max}(\Omega)} \lambda_{min}(W\Omega\Omega)\, Rank\left(RR'\right) && \text{(using PSDness)} \\
&\geq \frac{1}{\lambda_{max}(\Omega)} \lambda_{min}(W)\, \lambda_{min}(\Omega\Omega)\, Rank\left(RR'\right) && \text{(using RES1)} \\
&\geq \frac{1}{\lambda_{max}(\Omega)} \lambda_{min}(W)\, \lambda^2_{min}(\Omega)\, Rank\left(RR'\right) \\
&= \frac{\lambda^2_{min}(\Omega)}{\lambda_{max}(\Omega)} \lambda_{min}(W)\, Rank\left(RR'\right) \\
&= \frac{\lambda_{min}(\Omega)}{\lambda_{max}(\Omega)} \lambda_{min}(\Omega)\, \lambda_{min}(W)\, Rank\left(RR'\right)
\end{aligned}$$

Final risk bounds

Numerator

I. If $V = W^{-1}$

$$\begin{aligned}
tr\left(\bar{A}\right) - 2\lambda^{\frac{1}{2}}(A^*)\,\lambda^{\frac{1}{2}}(K) &\geq \lambda_{min}(\Omega)\,\lambda_{min}(W)\,Rank\left(RR'\right) - 2\frac{\lambda_{max}(W^{-1})}{\lambda_{min}(W^{-1})}\lambda_{max}(\Omega)\,\lambda_{max}(W) \\
&= \lambda_{min}(\Omega)\,\lambda_{min}(W)\,Rank\left(RR'\right) - 2\frac{\lambda_{max}(W)}{\lambda_{min}(W)}\lambda_{max}(\Omega)\,\lambda_{max}(W) \\
&= \lambda_{min}(\Omega)\,\lambda_{min}(W)\,Rank\left(RR'\right) - 2\frac{\lambda_{max}(W)}{\lambda_{min}(W)}\lambda_{max}(\Omega)\,\lambda_{max}(W) \\
&= \lambda_{min}(\Omega)\,\lambda_{min}(W)\left\{Rank\left(RR'\right) - 2\kappa(\Omega)\,\kappa^2(W)\right\}
\end{aligned}$$

II. If $V = \Omega$

$$\begin{aligned}
tr\left(\bar{A}\right) - 2\lambda^{\frac{1}{2}}(A^*)\,\lambda^{\frac{1}{2}}(K) &\geq \frac{\lambda_{min}(\Omega)}{\lambda_{max}(\Omega)}\lambda_{min}(\Omega)\,\lambda_{min}(W)\,Rank\left(RR'\right) \\
&- 2\frac{\lambda_{max}(\Omega)}{\lambda_{min}(\Omega)}\lambda_{max}(\Omega)\,\lambda_{max}(W) \\
&= \lambda_{min}(\Omega)\,\lambda_{min}(W)\left\{\frac{1}{\kappa(\Omega)}Rank\left(RR'\right) - 2\kappa^2(\Omega)\,\kappa(W)\right\}
\end{aligned}$$

Denominator

I. If $V = W^{-1}$

$$(Z+\Delta)^{'}(Z+\Delta)\frac{\lambda_{max}\left(W^{-1}\right)}{\lambda_{min}\left(W^{-1}\right)}\lambda_{max}(W) \;=\; (Z+\Delta)^{'}(Z+\Delta)\frac{\lambda_{max}(W)}{\lambda_{min}(W)}\lambda_{max}(W)$$

II. If $V = \Omega$

$$(Z+\Delta)^{'}(Z+\Delta)\frac{\lambda_{max}(\Omega)}{\lambda_{min}(\Omega)}\lambda_{max}(W)$$

Numerator and denominator combined

I. If $V = W^{-1}$

$$\begin{aligned}
&\frac{2\left[tr\left(\bar{A}\right)-2\frac{\lambda_{max}(V)}{\lambda_{min}(V)}\lambda_{max}(\Omega)\lambda_{max}(W)\right]-\tau}{(Z+\Delta)^{'}(Z+\Delta)\frac{\lambda_{max}(V)}{\lambda_{min}(V)}\lambda_{max}(W)}\\
=\;&\frac{2\left[\lambda_{min}(\Omega)\lambda_{min}(W)\left\{Rank\left(RR^{'}\right)-2\kappa(\Omega)\kappa^{2}(W)\right\}\right]-\tau}{(Z+\Delta)^{'}(Z+\Delta)\frac{\lambda_{max}(W)}{\lambda_{min}(W)}\lambda_{max}(W)}\\
=\;&\frac{\left(\frac{1}{\frac{\lambda_{max}(W)}{\lambda_{min}(W)}\lambda_{max}(W)}\right)\left(2\left[\lambda_{min}(\Omega)\lambda_{min}(W)\left\{Rank\left(RR^{'}\right)-2\kappa(\Omega)\kappa^{2}(W)\right\}\right]-\tau\right)}{(Z+\Delta)^{'}(Z+\Delta)}\\
=\;&\frac{\left(\frac{\lambda_{min}(W)}{\lambda_{max}(W)\lambda_{max}(W)}\right)\left(2\left[\lambda_{min}(\Omega)\lambda_{min}(W)\left\{Rank\left(RR^{'}\right)-2\kappa(\Omega)\kappa^{2}(W)\right\}\right]-\tau\right)}{(Z+\Delta)^{'}(Z+\Delta)}\\
=\;&\frac{2\left[\frac{\lambda_{min}(\Omega)\lambda_{min}(W)\lambda_{min}(W)}{\lambda_{max}(W)\lambda_{max}(W)}\left\{Rank\left(RR^{'}\right)-2\kappa(\Omega)\kappa^{2}(W)\right\}\right]-\left(\frac{\lambda_{min}(W)}{\lambda_{max}(W)\lambda_{max}(W)}\right)\tau}{(Z+\Delta)^{'}(Z+\Delta)}\\
=\;&\frac{2\left[\lambda_{min}(\Omega)\frac{1}{\kappa^{2}(W)}\left\{Rank\left(RR^{'}\right)-2\kappa(\Omega)\kappa^{2}(W)\right\}\right]-\left(\frac{\lambda_{min}(W)}{\lambda_{max}(W)\lambda_{max}(W)}\right)\tau}{(Z+\Delta)^{'}(Z+\Delta)}\\
=\;&\frac{2\left[\frac{\lambda_{min}(\Omega)}{\kappa^{2}(W)}\left\{Rank\left(RR^{'}\right)-2\kappa(\Omega)\kappa^{2}(W)\right\}\right]-\left(\frac{1}{\kappa(W)\lambda_{max}(W)}\right)\tau}{(Z+\Delta)^{'}(Z+\Delta)}
\end{aligned}$$

II. If $V = \Omega$

$$
\begin{aligned}
& \frac{2\left[tr\left(\bar{A}\right) - 2\frac{\lambda_{max}(V)}{\lambda_{min}(V)}\lambda_{max}(\Omega)\lambda_{max}(W)\right] - \tau}{(Z+\Delta)'(Z+\Delta)\frac{\lambda_{max}(V)}{\lambda_{min}(V)}\lambda_{max}(W)} \\
= & \frac{2\left[\lambda_{min}(\Omega)\lambda_{min}(W)\left\{\frac{1}{\kappa(\Omega)}Rank\left(RR'\right) - 2\kappa^2(\Omega)\kappa(W)\right\}\right] - \tau}{(Z+\Delta)'(Z+\Delta)\frac{\lambda_{max}(\Omega)}{\lambda_{min}(\Omega)}\lambda_{max}(W)} \\
= & \frac{\left(\frac{1}{\frac{\lambda_{max}(\Omega)}{\lambda_{min}(\Omega)}\lambda_{max}(W)}\right)\left(2\left[\lambda_{min}(\Omega)\lambda_{min}(W)\left\{\frac{1}{\kappa(\Omega)}Rank\left(RR'\right) - 2\kappa^2(\Omega)\kappa(W)\right\}\right] - \tau\right)}{(Z+\Delta)'(Z+\Delta)} \\
= & \frac{\left(\frac{\lambda_{min}(\Omega)}{\lambda_{max}(\Omega)\lambda_{max}(W)}\right)\left(2\left[\lambda_{min}(\Omega)\lambda_{min}(W)\left\{\frac{1}{\kappa(\Omega)}Rank\left(RR'\right) - 2\kappa^2(\Omega)\kappa(W)\right\}\right] - \tau\right)}{(Z+\Delta)'(Z+\Delta)} \\
= & \frac{2\left[\frac{\lambda_{min}(\Omega)}{\lambda_{max}(\Omega)\lambda_{max}(W)}\lambda_{min}(\Omega)\lambda_{min}(W)\left\{\frac{1}{\kappa(\Omega)}Rank\left(RR'\right) - 2\kappa^2(\Omega)\kappa(W)\right\}\right] - \left(\frac{\lambda_{min}(\Omega)\tau}{\lambda_{max}(\Omega)\lambda_{max}(W)}\right)}{(Z+\Delta)'(Z+\Delta)} \\
= & \frac{2\left[\frac{\lambda_{min}(\Omega)\lambda_{min}(W)}{\lambda_{max}(\Omega)\lambda_{max}(W)}\lambda_{min}(\Omega)\left\{\frac{1}{\kappa(\Omega)}Rank\left(RR'\right) - 2\kappa^2(\Omega)\kappa(W)\right\}\right] - \left(\frac{\lambda_{min}(\Omega)}{\lambda_{max}(\Omega)\lambda_{max}(W)}\right)\tau}{(Z+\Delta)'(Z+\Delta)} \\
= & \frac{2\left[\lambda_{min}(\Omega)\frac{1}{\kappa(\Omega)\kappa(W)}\left\{\frac{1}{\kappa(\Omega)}Rank\left(RR'\right) - 2\kappa^2(\Omega)\kappa(W)\right\}\right] - \left(\frac{\lambda_{min}(\Omega)}{\lambda_{max}(\Omega)\lambda_{max}(W)}\right)\tau}{(Z+\Delta)'(Z+\Delta)} \\
= & \frac{2\left[\frac{\lambda_{min}(\Omega)}{\kappa^2(\Omega)\kappa(W)}\left\{Rank\left(RR'\right) - 2\kappa^3(\Omega)\kappa(W)\right\}\right] - \left(\frac{1}{\kappa(\Omega)\lambda_{max}(W)}\right)\tau}{(Z+\Delta)'(Z+\Delta)}
\end{aligned}
$$

These results give the final risk bounds in the Theorem 3.

Chapter 3

Focused shrinkage estimators for the global minimum variance portfolio

Abstract. We propose a shrinkage estimator for the covariance matrix designed to minimize the mean squared error (MSE) of the Global Minimum Variance portfolio (GMV). Our proposed shrinkage estimator is a weighted average between an unrestricted estimator of the covariance matrix and a shrinkage target. While selecting a shrinkage target for the full covariance matrix is difficult, the residual covariance matrix from a factor model is expected to have a clear structure that we can exploit. This motivates our choice of the factor covariance model so that our shrinkage weights between the residual covariance matrix and its shrinkage target. Existing studies derive the optimal shrinkage weight between these to minimize the MSE of the covariance matrix itself without taking into account its subsequent use to estimate the GMV portfolio. Our estimator exploits the form of the solution to the GMV problem for determining the optimal weight. We conduct extensive simulations to compare the performance of our estimator with the existing competitors. Our estimator shows the most robust and superior performance in portfolios of different sizes. Similar improvements are found in empirical applications.

JEL: C31, C52, C58.

3.1 Introduction

Markowitz (1952, 1959) portfolio theory describes optimal decisions of an investor who cares about the mean and variances of static portfolio returns. While theoretically sound, the actual implementation of the mean-variance approach proved to be a very challenging task in practice. This is because the estimation errors of means and covariances outweigh the value

added by including them in a portfolio choice model (see Frost and Savarino (1986, 1988), Michaud (1989), Best and Grauer (1991), Chopra and Ziemba (1993), Broadie (1993), and Litterman et al. (2004) among others). The estimation error of the mean is particularly large and Jagannathan and Ma (2003) argued that nothing much is lost by excluding it from the optimization problem. This led to the so-called global minimum variance (GMV) portfolios. Haugen (1990), Haugen and Baker (1991) and Winston (1993) and more recently Clarke et al. (2006) and Clarke et al. (2011) illustrated the benefits of using GMV portfolios compared to stock-index or market capitalization weighted portfolios.

Implementing the GMV portfolio involves estimating the covariance matrix and then using it to obtain the variance-minimizing weights subject to the constraint that weights add up to one. A natural choice is to use the sample covariance matrix to obtain GMV weights. The estimation error of this 'plug-in' estimator can, however, be large in smaller samples (i.e. 60-120 of monthly observations used for GMV portfolios) when the number of assets is large. Using larger samples is also problematic because the distribution of returns changes over time that in turn translates into changing covariances. Time-varying properties of covariances are well-documented (Engle, 2002).

Subsequently, the financial and statistical literature has offered a number of approaches to improve the plug-in estimator. One line of research is focused on reducing estimation error in covariance matrices via shrinkage methods. These involve taking a weighted average between a sample estimator (of a covariance matrix) and some low-variance target estimator. The key idea is that this shrinking induces bias but reduces the variance of the resulting estimator. The appropriate choice of the weight (shrinkage intensity) provides the optimal trade-off between these effects such that the resulting estimator of the covariance matrix has minimum risk (expected error). Ledoit and Wolf (2003, 2004) propose shrinking the covariance matrix to a low-variance target estimator such as the identity matrix or 1-factor model covariance. Earlier studies on shrinkage in the context of portfolio choice include Jorion (1986) who proposed Bayes-Stein estimators of the mean of asset returns.

Another strand of literature considers reducing estimation error in the covariance matrix using factor models. Sharpe (1963) first suggested using a covariance matrix from a single factor market model in the mean-variance problem. Subsequently, factor models have been widely used in finance, see Ross (1976, 1977), Engle and Watson (1981), Chamberlain and Rothschild (1982), Chamberlain (1983), Fama and French (1992, 1993), Aguilar and West

(2000), Bai (2003), Ledoit and Wolf (2003), Stock and Watson (2005) among many others. Recent studies include Fan et al. (2008) that developed high-dimensional covariance matrix estimators using factor models. They further provided the bounds on the difference between the portfolio variance computed using their estimator and the true global minimum variance portfolio.

The methods discussed so far primarily target estimation of the moments of asset returns (i.e. covariance) rather then the optimal GMV portfolio weights. In contrast, a more recent line of research offers several estimators where the shrinkage or constraints are applied directly to portfolio weights. Jagannathan and Ma (2003) suggested imposing short sale constraints on portfolio weights in the GMV problem to mitigate the sampling error of the plug-in estimators. DeMiguel et al. (2009) extended this approach to a wider class of norm-constrained portfolios. In particular they consider adding 1-Norm and A-Norm constraints to the standard GMV framework. For high-dimensional portfolios Fan et al. (2012) analyzed 1-Norm constraints on weights to reduce the estimation error of weights inherited from estimating high dimensional covariance matrices; they then show that the resulting loss in portfolio variance is bounded. Finally, Frahm and Memmel (2010) propose dominating estimators of GMV portfolios which minimize the out-of-sample variance of portfolio return. Their estimator of GMV weights is the weighted average between the plug-in estimator and a chosen target portfolio.

While these studies target portfolio weights by 'disciplining' them using different constraints, they are in fact equivalent to solving the standard unconstrained GMV problem with the covariance matrix replaced by the appropriate shrinkage version. They are therefore similar to the methods discussed before with the difference that the shrinkage on the covariance matrix is done in an implicit way. For example, Jagannathan and Ma (2003) showed that the solution to the short sale-constrained problem coincides with the standard GMV problem in which the sample covariance matrix is replaced by $\hat{\Sigma} - \lambda \mathbf{1}^{'} - \mathbf{1}\lambda^{'}$ where λ is the vector of Lagrange multipliers for the short sale constraint. Similar results are shown by DeMiguel et al. (2009) and Fan et al. (2012) for general 1-Norm constrained GMV portfolios. For A-Norm constrained portfolios[1], DeMiguel et al. (2009) show that when matrix A equals identity then there is one-to-one correspondence between A-Norm-Constrained portfolios and the shrinkage portfolio proposed in Ledoit and Wolf (2004). Further, if A is chosen to be 1-factor covariance matrix Σ_F, then the solution to A-Norm-Constrained GMV portfolio coincides with

[1] A-Norm constraint is defined as $w^{'} A w \leq \delta$

the shrinkage portfolio in Ledoit and Wolf (2003). Next, for 2-Norm-Constrained portfolios, if the threshold parameter is set to $1/p$, where p is the number of assets, then the solution to this 2-Norm-Constrained GMV problem coincides with setting (i.e. fully shrinking) covariance matrix to a covariance matrix with identical variances and common covariances in the standard GMV problem. Finally, Frahm and Memmel (2010) show that their estimator of portfolio weights is equivalent to the solution to the standard GMV problem with the appropriate shrinkage of the sample covariance matrix.

To sum up, the approaches developed in the existing literature can be generally cast in terms of different shrinkage estimators of covariance matrices. Some of these shrinkage schemes are shown to be optimal in a sense of minimizing the risk of the shrinkage version of covariance matrix, some are shown to have bound on the resulting risk, whereas others modify the covariance matrix to reduce the extreme positions in certain assets. These approaches, however, ignore the final object of interest: GMV portfolio weights and its associated risk.

In this paper we propose a *focused* shrinkage estimator of the covariance matrix (Focused GMV). In contrast to the previous approaches which aim to improve upon estimation error in the covariance matrix or implicitly shrink it to reduce extreme weights, we tailor our shrinkage procedure to minimize the expected error on GMV portfolio weights - our final object of interest. More generally, in Chapter 2 we develop focused shrinkage estimators that minimize mean squared error of the chosen focus function $x(\theta)$ where θ is estimated from data. This approach is related to the Focused Information Criteria (see Claeskens and Hjort (2003)) because how we shrink depends on the functions of interest. In the current case our focus function $x(\theta)$ is the solution to the GMV portfolio and θ contains the elements of the covariance matrix.

It is known that shrinkage estimators show the best improvements when a chosen shrinkage target (or a restricted model) conforms to the data. Finding a shrinkage target for the raw covariance matrix is generally difficult. Panel A of figure 3.1.1 shows the rolling window (120 months) estimates of the off-diagonal elements of the sample covariance matrix for 25 Fama and French portfolios. As one can see, these covariances do not conform to some simplified structure or restriction like a diagonal or equicorrelation matrix. However, using a factor model to eliminate the effects of the common factors from the asset returns leaves us with the residual covariance matrix that conforms to more homogeneous structure that we can shrink towards. Residual covariances have been used for modeling the raw return covariance

in ARCH models Engle (1982) and GARCH models Bollerslev (1986). Panels B and C figure 3.1.1 display residual covariances from 1- and 3-factor models. As one can see, these look more similar than those in the raw covariances (Panel A) especially in the 3-factor case. This motivates our choice of the factor model to model covariance[2]. It is common in the literature on the factor models to makes residual covariance diagonal. Returning to figure 3.1.1 (Panels B and C) we see that although this assumption seems plausible, it may not always hold.

Figure 3.1.1 – Residual covariance estimates from the sample covariance and 1- and 3-factor models. Note. Panel A of figure 3.1.1 shows the rolling window (120 months) estimates of the off-diagonal sample covariance estimates for 25 Fama and French stocks. These covariances do not conform to some simplified structure or restriction like diagonal or equicorrelation matrix. Panels B and C figure 3.1.1 display residual covariances from 1- and 3-factor models. These are more homogeneous than that of the raw covariances (Panel A) especially in case of 3-factor model. This motivates our choice of the factor model so that our shrinkage weights between the residual covariance matrix estimator and its shrinkage target (see (3.11)).

[2]It is important to note that our use of the factor models is to reduce estimation error rather than to deal with dimensionality. When $p > n$, we cannot compute our unrestricted sample covariance matrix by the method of moments since it will not be positive-semidefinite.

Instead of setting the residual covariance matrix to diagonal, our focused shrinkage estimator weights between an unrestricted (method of moments) and a restricted estimators of the residual covariance. The weight between these two models (or shrinkage intensity) depends on a test of how far are restricted and unrestricted estimators from each other. In the focused approach, our test uses the right measure of 'closeness' between unrestricted and restricted models which depends on our focus i.e. GMV portfolio weights. The key idea here is that if the sensitivity of the focus function $x(\theta)$ to some parameter θ_i is small, then, then we may accept more estimation error in this parameter in order to improve upon the estimation error of a parameter θ_j that has a large impact on our focus function. The shrinkage intensity in our focused approach takes this trade-off into account providing a formal way to control for the estimation error in our focus function - GMV portfolio weights. This contrasts with the existing approaches that can be viewed as focusing solely on parameter θ (i.e. covariance matrix elements) regardless on how different parameters in θ affect the object of interest $x(\theta)$.

Apart from controlling for the estimation error in the chosen function, our proposed method has an advantage that we can consider a wide range of restricted models. These include shrinking to the equicorrelation model (after normalizing it with standard deviations), diagonal covariance model or possibly restricting the covariances to be common between some stock i and other stocks. This flexibility in terms of customization addresses a known weakness of shrinkage estimator whose optimality properties are strongly dependent on the shrinkage target.

Many proposed shrinkage estimators assume normal distribution of the asset prices. This is the case with Frahm and Memmel (2010) who derive shrinkage estimators of GMV portfolio weights and are most closely related to our approach in a sense of targeting some object of interest i.e. in their case - out-of-sample variance of the portfolio return. The normality assumption is, however, often very restrictive in economic and financial applications. Our focused approach does not impose any distributional assumptions that makes it applicable to a wide range of cases. Its asymptotic properties are developed in the localization framework (see Chapter 2) for more details).

It is important to emphasize that different focus functions can lead to different estimators of the covariance matrices. For example, the covariance matrix estimator that we propose for the GMV portfolio may be different from the covariance matrix that is optimal for the mean-variance portfolio (for the latter see Chapter 2). This is because the effect of the error

in means is considerably larger than that in variances and our focused approach will take that into account. Focused shrinkage is shown to work especially well in smaller samples where controlling for the estimation errors is important. These cases are typically encountered in practice when using monthly data.

The rest of the paper is organized as follows. We first describe GMV portfolio problem set up, its solution and the existing estimation approaches (Section 3.2). We then introduce Focused-GMV estimators in Section 3.3. We conduct extensive simulations designed to realistically represent the U.S. stock price dynamics in Section 3.4. Finally, we apply our method to different data sets in Section 3.5 and conclude in Section 3.6.

3.2 Global minimum variance problem and existing estimation approaches

Consider the global minimum variance (GMV) portfolio which is the the solution to the following problem:

$$\begin{aligned} &\min_{x} x^{'}\Sigma x \\ &\text{s.t. } x^{'}\mathbf{1} = 1 \end{aligned} \tag{3.1}$$

with x is $p \times 1$ the vector of portfolio weights and $\mathbf{1}$ is $p \times 1$ vector of ones. The solution is given by

$$x\left(\Sigma\right) \;=\; \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \tag{3.2}$$

Denote x^* to be the true weights, i.e. the solution to this problem based on the true covariance matrix $x^* = x\left(\Sigma\right)$. Let $R_t = \left(R_{1t}, \ldots, R_{pt}\right)^{'}$ for $t = 1, \ldots, n$ be $p \times 1$ vectors of returns for t periods. Denote the associated $p \times p$ sample covariance matrix by $\hat{\Sigma}$. The plug-in estimator of GMV $\hat{w}$ is calculated by using the sample covariance matrix $\hat{\Sigma}$ instead of the true one i.e. $\hat{x} = x\left(\hat{\Sigma}\right)$.

Ledoit and Wolf (2003, 2004) estimate the GMV portfolio weights by replacing the sample covariance matrix with the shrinkage estimator of the covariance matrix $\hat{\Sigma}_{LW}$ defined as

$$\hat{\Sigma}_{LW} \;=\; \frac{k}{n}\,\hat{\Sigma}_{\text{target}} + \left(1 - \frac{k}{n}\right)\hat{\Sigma} \tag{3.3}$$

which is the weighted average of the sample covariance and the chosen low-variance target estimator $\hat{\Sigma}_{\text{target}}$. They consider several candidates for the target matrix such as the identity matrix or the 1-factor covariance matrix with market as a factor. Their proposed shrinkage estimators are suitable for large dimensional portfolios i.e. when $p > n$. The authors provide an optimal way of estimating k by minimizing the expected Frobenius norm between $\hat{\Sigma}_{LW}$ and the true covariance matrix. This loss function, however, does not take into account the fact that $\hat{\Sigma}_{LW}$ will be subsequently used to estimate GMV portfolio weights. The optimal value of k for 1-factor covariance as a target matrix is outlined in Appendix 3.7.1.

Frahm and Memmel (2010) develop the estimators for GMV portfolios that minimize out-of-sample variance of the portfolio return. Their estimator has the following form:

$$\hat{x}_{FM} \;=\; k_S x_R + (1 - k_S)\,\hat{x} \tag{3.4}$$

where x_R is the reference portfolio chosen to be the equally-weighted portfolio and k_S is the shrinkage intensity estimated from data and specified in Appendix 3.7.2. Although their estimator is a shrinkage estimator of portfolio weights, it is equivalent to the solution to the standard GMV problem where the sample covariance matrix is replaced by the appropriate shrinkage version of the covariance matrix (see Theorem 8, Frahm and Memmel (2010)). Further, the properties of this estimator rely on the assumption of normally and serially independent returns.

Next we outline the Norm-Constrained portfolios. DeMiguel et al. (2009) consider adding an extra constraint to the GMV problem in (3.1). They consider two types of constraints: Norm-1 and Norm-A defined as follows:

$$\begin{aligned} \text{Norm-1:}\;\; \|x\|_1 &= \textstyle\sum_{i=1}^{N} |x_i| \;\leq\; \delta \\ \text{Norm-A:}\;\; \|x\|_A &= \left(x' A x\right)^{1/2} \;\leq\; \delta \end{aligned} \tag{3.5}$$

Let $\tilde{x}_{NC1}$ and $\tilde{x}_{NCA}$ be the solutions to 1-Norm and A-Norm GMV portfolios correspondingly based on the sample covariance matrix $\hat{\Sigma}$ and the threshold parameter δ. DeMiguel et al. (2009) suggest determining the optimal δ by cross-validation methods either by minimizing the portfolio variance or maximizing the last period portfolio return (see details in Appendix 3.7.3). They also show that the short sale-constrained GMV portfolio analyzed by Jagannathan and Ma (2003) is the special case of 1-Norm constrained portfolio when $\delta = 1$.

Although 1-Norm and A-Norm constrained GMV portfolios use the sample covariance matrix as the input they in fact coincide with solving the standard GMV problem (3.1) with the sample covariance matrix replaced by the appropriate shrinkage version of the covariance matrix as shown in Proposition 6 of DeMiguel et al. (2009).

To sum up, the existing estimators of the GMV portfolio can be cast in terms of different shrinkage estimators of the covariance matrix. These estimators are shown to be optimal either in terms of minimizing the risk (expected error) between the proposed estimator and the true covariance matrix or having a bounded risk. Some of them also rely on the assumptions of normal distributions and serial independence. We now introduce the focused shrinkage estimator of the covariance matrix that is designed to minimize the risk of the estimated GMV weights.

3.3 Focused GMV estimator

Consider the problem of minimizing the risk of an estimator of GMV portfolio weights $w\left(\tilde{\Sigma}\right)$ based on some estimator of the covariance matrix $\tilde{\Sigma}$. We choose this risk to be the asymptotic mean square error (MSE):

$$\lim_{n\to\infty} E\left[n\left(x\left(\tilde{\Sigma}\right)-x\left(\Sigma\right)\right)^{'}\left(x\left(\tilde{\Sigma}\right)-x\left(\Sigma\right)\right)\right] \tag{3.6}$$

The estimation setting is augmented by the belief that the true covariance matrix Σ may be close to a restricted (or highly-structured) matrix Σ_0. A formal way to analyze this situation is the localization framework (see Van der Vaart (2000)). Let θ denote the $m=\frac{p(p+1)}{2}$ stacked vector of the covariance matrix Σ elements, with variances coming first, and covariances second:

$$\theta = \begin{bmatrix} \sigma_{11} \\ \vdots \\ \sigma_{pp} \\ \sigma_{12} \\ \vdots \\ \sigma_{(p-1)p} \end{bmatrix} \tag{3.7}$$

When n observations are available, our data is assumed to be distributed according to

$$\theta_n \;=\; \theta_0 + n^{-1/2}\Delta$$

where θ_0 lies in a restricted subspace Θ_0. Thus with n observations, the true covariance is θ_n, the restricted covariance (under the null hypothesis) is θ_0 and Δ is a localizing parameter. This captures our hypothesis that the true covariance θ_n is close to its restricted version θ_0: as $n \to \infty$, we expect them to coincide. In this situation a natural estimator for the covariance matrix is to weight between an unrestricted estimator of the covariance matrix $\hat{\theta}$ and its restricted version $\hat{\theta}_R$. Our shrinkage estimator of the covariance matrix thus has the following form:

$$\tilde{\theta} \;=\; w\hat{\theta} + (1-w)\,\hat{\theta}^R$$

where w defines the weight on the unrestricted model. It is easy to see that our shrinkage estimator should put more weight on the unrestricted model when Δ is large i.e. when the true parameter θ_n is far away from its restricted version θ_0. In our focused shrinkage approach, the weight between θ_n and θ_0 is based on the distance between the implied focus functions i.e. $x(\theta_n)$ and $x(\theta_0)$. This translates to controlling for the error of of $x\left(\tilde{\theta}\right)$, rather than error of $\tilde{\theta}$ that is commonly considered in the literature. We now describe the unrestricted and restricted estimators of the covariance matrix as well as the optimal shrinkage intensity w for the GMV portfolio problem.

3.3.1 Unrestricted estimator

Our unrestricted estimator of the covariance matrix is based on a factor model. As discussed before, factor models offer us a sensible shrinkage target - the residual covariance matrix. By construction, effects of the common factors are removed and therefore it may conform to some homogeneous structure that we can shrink towards. We start by assuming that asset returns R_t are driven by a set of factors:

$$R_t \;=\; Bf_t + u_t \tag{3.8}$$

where

$$f_t \quad (k \times 1)$$

$$B \quad (p \times k)$$

$$u_t \quad (p \times 1)$$

that correspond to k factors at period t, a matrix of factor loadings, and residuals. The matrix of factor can also be represented as follows:

$$B \;=\; \begin{bmatrix} b_1 \;\ldots\; b_p \end{bmatrix}', \quad b_i \;=\; \begin{bmatrix} b_{i1} \\ \vdots \\ b_{ik} \end{bmatrix}$$

We assume that $E\left(u_t | \, f_t\right) = 0$ and that u_t are uncorrelated across t, that is a common assumption in the literature (see Fan et al. (2008)). Now let

$$cov\left(f_t\right) \;=\; \Sigma_f$$

$$cov\left(u_t\right) \;=\; \Sigma_u$$

Under model (3.8), covariance is expressed as

$$\Sigma \;=\; B\Sigma_f B' + \Sigma_u \tag{3.9}$$

This is estimated using plug-in least-squares estimators of B, Σ_f, and Σ_u:

$$\begin{aligned} \hat{B}' \;&=\; \left(\textstyle\sum_{t=1}^{n} R_t f_t'\right)\left(\textstyle\sum_{t=1}^{n} f_t f_t'\right)^{-1} \\ \hat{\Sigma}_f \;&=\; \frac{1}{n}\textstyle\sum_{t=1}^{n}\left(f_t - \bar{f}\right)\left(f_t - \bar{f}\right)' \quad , \quad \bar{f} = \begin{bmatrix} \bar{f_1} \\ \vdots \\ \bar{f_k} \end{bmatrix} = \frac{1}{n}\textstyle\sum_{t=1}^{n} f_t \\ \hat{\Sigma}_u \;&=\; \frac{1}{n}\textstyle\sum_{t=1}^{n} \hat{u}_t \hat{u}_t' \end{aligned} \tag{3.10}$$

It is important to note here that the expression for $\hat{\Sigma}_u$ is different from the one used in the existing literature. The latter makes $\hat{\Sigma}_u$ diagonal in (3.10) which ensures it positive

semi-definiteness even when $p > n$. As discussed before in the context of figure 3.1.1, this restriction may not always conform to the data. In contrast, our approach uses the method of moments estimator of the residual covariance (3.10) as an unrestricted model and thus applicable for $p < n$. We consider 3 types of factor models:

1. 1-factor model with the excess return on the market as a factor.
2. 3-factor model by Fama and French (1992, 1993).
3. 4-factor model including 3 Fama and French factors as well as the momentum factor introduced by Carhart (1997).

Note that the solution to the GMV problem (3.2), now depends on B, Σ_f, and Σ_u i.e.

$$x\left(\Sigma\right) \;=\; x\left(B, \Sigma_f, \Sigma_u\right).$$

Our shrinkage estimator is a weighted average between the unrestricted estimator of Σ_u specified in (3.10) and a restricted estimator Σ_u^R. The resulting estimator of the return covariance matrix has the following form:

$$\tilde{\Sigma} \;=\; \hat{B}\hat{\Sigma}_f\hat{B}' + w\hat{\Sigma}_u + \left(1-w\right)\hat{\Sigma}_u^R \tag{3.11}$$

The unrestricted estimator of the residual covariance matrix $\hat{\Sigma}_u$ is consistent and asymptotically normal under standard assumptions. Stacking the elements of $\hat{\Sigma}_u$ in a vector as in (3.7), we have:

$$\sqrt{n}\left(\hat{\theta}_u - \theta_{u,n}\right) \;\rightarrow\; Z \;\sim\; N\left(0, \Omega\right)$$

with

$$\hat{\theta}_u \;=\; \begin{bmatrix} \hat{\sigma}_{u,11} \\ \vdots \\ \hat{\sigma}_{u,pp} \\ \hat{\sigma}_{u,12} \\ \vdots \\ \hat{\sigma}_{u,(p-1)p} \end{bmatrix} \tag{3.12}$$

where $\theta_{u,n}$ is the true sample-size dependent residual covariances. Note here that few distributional assumptions on the data are required for this result to hold. Asymptotic normality holds in both i.i.d. and non-i.i.d. settings (for latter see Bosq (2012)).

3.3.2 Restricted estimator

We now consider the restricted estimator of the residual covariance $\hat{\theta}_u^R$. Given that we choose residual covariance as our shrinkage target, we will now rewrite our GMV solution as a function of θ_u only. This means that we exclusively focus on reducing the estimation error of GMV weights via controlling the estimation error of Σ_u. A more complete approach to shrinkage should also account for the uncertainty in estimating B and Σ_f. This, however, requires accounting for additional $pk + k^2$ parameters that is computationally costly when dimension p or number of factors k is large. We leave this to future research. Thus:

$$x \;=\; x\left(\Sigma_u\right) \;\equiv\; x\left(\theta_u\right)$$

Note further, that because of the delta method, the risk of GMV portfolio weights (3.6) is asymptotically equivalent to

$$\lim_{n\to\infty} E\left[n\left(\tilde{\theta}_u - \theta_{u,n}\right)' W \left(\tilde{\theta}_u - \theta_{u,n}\right)\right] \tag{3.13}$$

where $D_{\theta_u} = \frac{\partial}{\partial \theta_u} x\left(\theta_{u,0}\right)$ is the matrix of partial derivatives of the solution to GMV problem (3.2) with respect to the residual covariances and

$$W \;=\; D_{\theta_u}' D_{\theta_u}$$

Thus, the risk we want to minimize is equivalent to the MSE with a weight matrix W. This matrix is important in mapping the distance between $\theta_{u,n}$ and $\theta_{u,0}$ into that between $x\left(\theta_{u,n}\right)$ and $x\left(\theta_{u,n}\right)$. Our restricted estimator of the residual covariance is the following minimum distance estimator:

$$\begin{aligned} \hat{\theta}_u^R \;&=\; \arg\min_{\theta} \left(\hat{\theta}_u - \theta_u\right)' V^{-1} \left(\hat{\theta}_u - \theta_u\right) \\ &\text{s.t. } \theta_u \in \Theta_0 \end{aligned} \tag{3.14}$$

This maps into the covariance matrix $\hat{\Sigma}_u^R$ and in combination with the unrestricted

estimator $\hat{\Sigma}_u$ and the factor model estimates (3.10) gives the final shrinkage estimator of the covariance matrix (3.11). For the case of linear restrictions we have

$$\Theta_0 = \left\{ \theta | \, R^{'} \theta_u - a = r \right\}$$

where R is a restriction matrix and a is a vector. Our restricted estimator $\hat{\theta}_u^R$ has the following form[3]:

$$\hat{\theta}_u^R \; = \; \hat{\theta}_u - VR \left(R^{'} V R \right)^{-1} \left(R^{'} \hat{\theta}_u - a \right) \tag{3.15}$$

where V is $m \times m$ positive semi-definite matrix. The choice of V is important in determining the properties of the resulting shrinkage estimator. Note that for $V = I$, this maps the distance between unrestricted estimator $\hat{\theta}_u$ and the restricted estimator $\hat{\theta}_u^R$ into the Euclidean space. Hansen (2016) suggests setting $V = \Omega$. In Theorem 1 of Chapter 2 we show that if $\Delta \neq 0$, then the variance of $\hat{\theta}_u^R$ is minimized by setting $V = \Omega$ whereas the bias part is minimized by setting $V = W^{-1}$. In smaller samples, it is possible that the bias part will dominate, and Chapter 2 illustrates that the latter choice works well in simulations and applications to the mean-variance framework with transaction costs. Importantly, in the context of the minimum distance estimator outlined in (3.14), setting $V = W^{-1}$ leads to projecting the distance between our full and restricted estimators ($\hat{\theta}_u$ and $\hat{\theta}_u^R$) into the *focused* space i.e. the space defined by our chosen focus function i.e. GMV portfolio weights. Our restricted estimator $\hat{\theta}_u^R$ will thus be defined relative to our object of interest, and incorporate the sensitivity of the focus function to different parameters.

The invertibility of matrix W is important for these results. Since the dimension of the focus $x(\theta_u)$ is smaller than the dimension of θ_u, the resulting matrix $W = D_{\theta_u}^{'} D_{\theta_u}$ will be rank deficient. For that reason, following Chapter 2, we consider regularizing matrix W towards the identity i.e. $W = D_{\theta_u}^{'} D_{\theta_u} + \lambda I$ that was found to work well in practice. We will discuss the choice of the regularization parameter λ below in the context of GMV portfolio weights.

[3]Note that in case when the shrinkage is toward the full space of the unrestricted model i.e. $R = I$, the restricted model is fixed $\hat{\theta}_R = \theta_0$ (no estimation needed). When the shrinkage is toward the subspace of the full model, the restricted model needs to be estimated as described in the main text.

3.3.3 Optimal focused shrinkage intensity w

We now describe how to obtain the shrinkage weight w for GMV portfolio weights. Our shrinkage estimator of the residual covariance is defined as

$$\tilde{\theta}_u \;=\; w_n \hat{\theta}_u + (1 - w_n)\, \hat{\theta}_u^R \tag{3.16}$$

with the resulting residual covariance matrix $\tilde{\Sigma}_u$ obtained by mapping the vector $\tilde{\theta}_u$ into matrix $\tilde{\Sigma}_u$. This translates into the following shrinkage estimator of the full covariance matrix:

$$\tilde{\Sigma} \;=\; \hat{B}\hat{\Sigma}_f\hat{B}' + \tilde{\Sigma}_u.$$

The shrinkage intensity w_n in (3.16) has a standard positive-rule form[4]:

$$w_n \;=\; \left(1 - \frac{\tau_n}{D_n}\right)_+ \tag{3.17}$$

where D_n is a test that the restriction $R'\theta_u = r$ holds[5]:

$$D_n \;=\; n\left(\hat{\theta}_u - \hat{\theta}_u^R\right)' W \left(\hat{\theta}_u - \hat{\theta}_u^R\right) \tag{3.18}$$

Note that this test statistic is a delta approximation of the difference between our focus functions i.e.

$$n\left(x\left(\hat{\theta}_u\right) - x\left(\hat{\theta}_u^R\right)\right)' \left(x\left(\hat{\theta}_u\right) - x\left(\hat{\theta}_u^R\right)\right) \to n\left(\hat{\theta}_u - \hat{\theta}_u^R\right)' W \left(\hat{\theta}_u - \hat{\theta}_u^R\right)$$

Thus, our measure of closeness between the restricted and unrestricted models is defined by the distance between the implied values of our focus functions - GMV portfolio weights. A larger difference between the restricted and unrestricted models translates into the bigger value of D_n and hence lower weight on the restricted model. Finally, $\tau \geq 0$ controls the degree of shrinkage and is chosen by the researcher. The optimal value of τ (under the condition

[4] $(x)_+ = \begin{cases} x \text{ for } x \geq 0 \\ 0 \text{ for } x < 0 \end{cases}$

[5] we consider linear restrictions in our applications

that $\tau > 0$) is derived in Theorem 2 in Chapter 2 and has the following form:

$$\tau^* = \left[tr\left(\bar{A}\right) - 2\lambda_{max}^{1/2}\left(A^*\right)\lambda_{max}^{1/2}\left(K\right)\right], \tag{3.19}$$

with λ_{max} denoting the maximum eigenvalue and

$$\begin{aligned} \bar{A} &= \left(R^{'}VR\right)^{-1}R^{'}VW\Omega R \\ A^* &= W^{1/2}VR\left(R^{'}VR\right)^{-1}R^{'}VW^{1/2} \\ K &= W^{1/2}\Omega R\left(R^{'}VR\right)^{-1}R^{'}\Omega W^{1/2} \end{aligned} \tag{3.20}$$

In practice, we use the sample analogues of of these.

3.3.4 Computation of the Focused GMV estimator of the covariance matrix

Our focused estimator of the covariance is constructed in three steps.

1. Compute the unrestricted model estimator and a consistent estimator of its asymptotic covariance. Our unrestricted model estimator is the residual covariance Σ_u computed from (3.10) and presented in the vector form θ_u as in (3.12). A consistent estimate of its asymptotic covariance Ω has different expressions depending on how returns are distributed. For i.i.d. returns we have

$$\hat{\Omega} = \frac{1}{n}\sum_{t=1}^{n}\hat{\lambda}_t\hat{\lambda}_t^{'}, \quad \hat{\lambda}_t = \begin{bmatrix} \hat{u}_{1t}^2 - \hat{\sigma}_{u,11} \\ \vdots \\ \hat{u}_{(p-1)t}\hat{u}_{pt} - \hat{\sigma}_{u,(p-1)p} \end{bmatrix}$$

For non-i.i.d. data, estimators of $\hat{\Omega}$ are also available and should be computed using limit theorems for strongly mixing processes (see Bosq (1998)).

2. Compute the restricted model using (i) the chosen restriction R and (ii) the estimator of the projection matrix $V = W^{-1}$. We start with the restriction matrix R. For the case of linear restrictions we have $R^{'}\theta_u - a = r$. There are several reasonable restrictions we can consider for our application. First, we can shrink the residual covariance to a diagonal matrix. Another option is to pool covariances to the same value. This is usually done together with

normalizing the data by standard deviations in order to put all the variables on the same scale. We illustrate these cases below for $p = 3$ assets such that we have $m = \frac{p\,(p+1)}{2} = 6$ residual covariance elements:

$$R^{'}\theta_u - a \;=\; \overset{\text{diagonal}}{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{'} \begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{23} \end{bmatrix}} - \mathbf{0}_3, \qquad R^{'}\theta_u - a \;=\; \overset{\text{pooling}}{\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}^{'} \begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{13} \\ \sigma_{23} \end{bmatrix}} - \mathbf{0}_2$$

Next, our projection matrix is $V = W^{-1}$. Since W is not invertible we regularize it i.e. $V = W^{-1} = \left(D_{\theta_u}^{'} D_{\theta_u} + \lambda I\right)^{-1}$. To construct V we need the derivative of our focus function - solution to the GMV portfolio problem $x\,(\theta_u)$. This has the following form:

$$D_{\theta_u} \;=\; \frac{\partial x}{\partial \sigma_{u,ij}} \;=\; \frac{-\Sigma^{-1}\dfrac{\partial \Sigma_u}{\partial \sigma_{u,ij}}\Sigma^{-1}\mathbf{1}\mathbf{1}^{'}\Sigma^{-1}\mathbf{1} + \Sigma^{-1}\mathbf{1}\mathbf{1}^{'}\Sigma^{-1}\dfrac{\partial \Sigma_u}{\partial \sigma_{u,ij}}\Sigma^{-1}\mathbf{1}}{\left(\mathbf{1}^{'}\Sigma^{-1}\mathbf{1}\right)^2}$$

We evaluate this derivative at the initial restricted model estimate: $\hat{D}_{\theta_u} = \hat{D}_{\theta_u}\left(\hat{\theta}_u^{R,0}\right)$ where $\hat{\theta}_u^{R,0}$ is computed with $V = I$:

$$\hat{\theta}_u^{R,0} \;=\; \hat{\theta}_u - R\left(R^{'}R\right)^{-1}\left(R^{'}\hat{\theta}_u - a\right)$$

We then construct projection $\hat{V} = \left(\hat{D}_{\theta_u}^{'}\hat{D}_{\theta_u} + \lambda I\right)^{-1}$. We choose the amount of regularization λ by considering a range of values. Typically, we find that $\lambda = 3 \cdot \max\left|\hat{D}_{\theta_u}^{'}\hat{D}_{\theta_u}\right|$ (where $|X|$ denotes component-wise absolute values of X) works well for GMV problem with other choices giving us stable results across different simulation scenarios. Finally, we compute restricted estimator of covariance $\hat{\theta}_u^R$ as specified in (3.15) using the unrestricted model estimator $\hat{\theta}_u$, restriction matrix R, vector a, and projection $\hat{V}$.

3. Compute the optimal shrinkage intensity. We weight between the full and restricted estimators of the residual covariance using the positive-rule as outlined in (3.17). For that we

compute the test D_n as specified in (3.18) and the sample analogue of the optimal value of τ from (3.19). For the latter we use previously estimated $\hat{\Omega}$ and $\hat{V}$ to compute primitive matrices specified in (3.20). We then compute our focused shrinkage estimator of the covariance as outlined in (3.16).

3.4 Simulations

3.4.1 Set up

We illustrate the performance of our focused covariance estimator in a series of simulations that are designed to capture the properties of the data used in our applications. We assume that the returns on the assets R_t are driven by a 3-factor model. A similar factor model design was used by Fan et al. (2008) for the asset return covariance estimation and Jagannathan and Ma (2003) for computing their estimator of GMV portfolio.

To inform our choice of simulation parameters B, Σ_f, and Σ_u, we fit a 3-factor model (3.8) using 10 years of monthly data (120 observations) of 25 Fama and French portfolios sorted by size and book-to-market from 2005/12 to 2015/12. We estimate B, Σ_f, and Σ_u as specified in (3.10). We then compute the following statistics:

$$\begin{aligned}
\mu_f &= \frac{1}{n}\sum_{t=1}^{n} f_t \\
cov_f &= \frac{1}{n}\sum_{t=1}^{n} (f_t - \mu_f)(f_t - \mu_f)' \\
\mu_B &= \frac{1}{p}\sum_{i=1}^{p} \hat{b}_i \\
cov_B &= \frac{1}{p}\sum_{i=1}^{p} \left(\hat{b}_i - \mu_B\right)\left(\hat{b}_i - \mu_B\right)' \\
\mu_{var(\Sigma_u)} &= \frac{1}{p}\sum_{\substack{i=1 \\ i=j}}^{p} \hat{\sigma}_{u,ij}, \\
\mu_{cov(\Sigma_u)} &= \frac{1}{d}\sum_{\substack{i=1 \\ i\neq j}}^{d} \hat{\sigma}_{u,ij}, \\
var_{cov(\Sigma_u)} &= \frac{1}{d}\sum_{\substack{i=1 \\ i\neq j}}^{d} \left(\hat{\sigma}_{u,ij} - \mu_{cov(\Sigma_u)}\right)^2 \\
d &= \tfrac{p(p-1)}{2}
\end{aligned}$$

We report the value of these for the 25FF data set described in Table 3.1.

Panel A: factors and factor loadings			
μ_f	cov_f		
0.6318	20.3548	3.8735	3.3170
0.0547	3.8735	5.1157	0.9915
−0.1315	3.3170	0.9915	5.7993
μ_B	cov_B		
1.0175	0.0050	−0.0013	−0.0101
0.5602	−0.0013	0.2084	−0.0032
0.1654	−0.0101	−0.0032	0.1515

Panel B: residual covariances		
$\mu_{var(\Sigma_u)}$	$\mu_{cov(\Sigma_u)}$	$var_{cov(\Sigma_u)}$
2.0829	0.0321	0.1678

Panel C: scaling constant						
p	5	10	20	50	70	100
c_p	5	5	5	10	15	17

Table 3.1 – Parameters used in simulations. This table summarizes parameter values used in simulations. These are computed from (3.1). Panel A presents the means and covariances of the factors f_t and factor loadings B that correspond to the empirical values of 25 Fama and French portfolios sorted by size and book-to-market in the period 2005/12 to 2015/12 (10 years of monthly data i.e. 120 observations). We use multivariate normal distributions to generate factor and factor loading with these means and covariances. Panel B presents the summary statistics on the residual covariance matrix Σ_u. The mean values of diagonal terms (variances) and off-diagonal terms (covariances) are reported as well as the variances of the off-diagonal terms. These are then used to generate the residual covariance matrix by setting all the diagonal terms to the empirical mean value of the variances and then simulating the off-diagonal terms from the normal with the corresponding means and variances. To ensure that the simulate residual covariance matrix is positive-semidefinite we scale the variance in the simulation by c_p that is reported in Panel C. This scaling c_p increases with dimension p.

For each simulation, we carry the following steps:

1. Generate a random sample of factors of size $n = 120$ following a 3-variate normal distribution $f_t \sim N(\mu_f, cov_f)$.

2. Then for dimensionality $p \in \{5, 10, 20, 50, 70, 100\}$ we do the following.

3. Generate a random sample of size p of market factor loadings following from the 3-variate normal distribution $B \sim N(\mu_B, cov_B)$.

4. Construct the residuals covariance matrix Σ_u by setting all its variances (diagonal elements) to $\mu_{var(\Sigma_u)}$ and generating its covariances (off-diagonal terms) from a normal distribution $\sigma_{u,ij} \sim N\left(\mu_{cov(\Sigma_u)}, var_{cov(\Sigma_u)}/c_p\right)$ for $p \in \{5, 10, 20, 50, 70, 100\}$. We use normal distribution to control the dispersion of the residual covariances. The scaling

of the variance by c_p is done to ensure that the residual covariance matrix is positive semi-definite. This scaling is increasing with the dimension (see table 3.1).

5. Obtain the true return covariance matrix Σ by plugging cov_f and generated B and Σ_u into (3.9). Compute the true GMV weights using Σ from (3.2).

6. Generate a random sample of residuals of size $n = 120$ from p-variate normal distribution $u_t \sim N(0, \Sigma_u)$.

7. Obtain a random sample of $R = (R_1, \ldots, R_n)'$ which is $n \times p$ matrix with $n = 120$ and $p \in \{5, 10, 20, 50, 100\}$ by plugging the previously generated f_t, B, and u_t into (3.8).

8. Compute our focused-shrinkage estimators based on 3-factor model by choosing different shrinkage targets such as diagonal and pooled. We also compute a number of competitors that we outline below.

We repeat the above simulations 500 times and report the mean squared error of the computed weights and the global minimum variance computed as follows:

$$\begin{aligned} MSEw &= \frac{1}{500} \sum_{i=1}^{500} (\hat{x}_i - x_i)' (\hat{x}_i - x_i) \\ MSEvar &= \frac{1}{500} \sum_{i=1}^{500} \left(\hat{x}_i' \Sigma \hat{x}_i - x_i' \Sigma x_i \right)^2 \end{aligned}$$

where $\hat{w}_i$ is an estimator of GMV portfolio weights computed from simulation i and w_i is the true GMV weights in simulation i.

3.4.2 Competitors

We consider a number of competitors to our proposed focused estimator. The first four are simple benchmarks such as equally-weighted portfolio defined as $x_{EW} = \frac{1}{p}\mathbf{1}$ as well as GMV portfolios based on the sample, 1-factor and 3-factor covariance matrices i.e. $\hat{\Sigma}$, $\hat{\Sigma}_{1F}$ and $\hat{\Sigma}_{3F}$. The latter two are obtained from (3.10) with the difference that all off-diagonal elements of the residual covariance matrix $\hat{\Sigma}_u$ are set to zero [6]. We also compute two competing shrinkage estimators: Ledoit and Wolf (2003) estimator of the covariance matrix (3.3) with 1-factor model covariance as a target matrix, and the Frahm and Memmel (2010) dominating estimator of GMV portfolio weights (3.4) with equally-weighted reference portfolio. Details are provided

[6] This is in keeping in line with the the existing literature (e.g. Fan et al. (2008))

in Appendices 3.7.1 and 3.7.2. Finally, we consider GMV portfolios with an additional constraint. We first compute Jagannathan and Ma (2003) short-sale constrained portfolio based on sample covariance matrix $\hat{\Sigma}$. Next, we consider the 1-Norm-Constrained portfolios (3.5) studied in DeMiguel et al. (2009). These were shown to perform well both in terms of minimizing the out-of-sample variance and maximizing the Sharpe Ratio. 1-Norm-Constrained portfolios are computed by solving the GMV problem (3.1) using the sample covariance matrix $\hat{\Sigma}$ with an extra 1-Norm constraint (3.5). Following DeMiguel et al. (2009), we choose the threshold parameter δ by either (i) using cross-validation to minimize the portfolio variance or (ii) maximize the last period return to exploit positive autocorrelation in portfolio returns. The details of cross-validation are provided in Appendix 3.7.3.

Table 3.2 summarizes portfolios to which we compare the performance of our proposed focused estimators.

No.	**Model**	**Abbreviation**
	Panel A: Portfolio strategies developed in this paper	
1	Focused shrinkage of covariance to diagonal covariance in a 1-factor model	fd1
2	Focused shrinkage of covariance to equicorrelation model in a 1-factor model	fp1
3	Focused shrinkage of covariance to diagonal covariance in a 3-factor model	fd3
4	Focused shrinkage of covariance to equicorrelation model in a 3-factor model	fp3
5	Focused shrinkage of covariance to diagonal covariance in a 4-factor model	fd4
6	Focused shrinkage of covariance to equicorrelation model in a 4-factor model	fp4
	Panel B: Portfolio strategies from the existing literature used for comparison	
Simple benchmarks		
1	Equally-weighted portfolio	EW
2	GMV portfolio with sample covariance matrix	S
3	GMV with 1-factor covariance matrix	1F
4	GMV with 3-factor covariance matrix	3F
GMV using shrinkage estimators		
1	Shrinking covariance to 1-factor covariance (Ledoit and Wolf, 2003)	LW
2	Shrinking weights to equally-weighted portfolio (Frahm and Memmel, 2010)	FM
GMV with additional constraints		
1	Short sale-constrained portfolio (Jagannathan and Ma, 2003)	JM
2	1-Norm constraint with δ calibrated over portfolio variance	NC1V
3	1-Norm constraint with δ calibrated over last period return	NC1R

Table 3.2 – List of Portfolios Considered. This table presents the methods computed in simulations and empirical applications. Panel A presents the estimators developed in this paper and Panel B reports commonly used competitors from the existing literature. In simulations we only use focused shrinkage estimators based on 3-factor models (fd3 and fp3) whereas in the empirical applications we compute all the listed focused shrinkage estimators.

3.4.3 Simulation results

Figure 3.4.1 reports the Mean Squared Error of the GMV portfolio weights estimated using different methods. Figure 3.4.2 reports four best competitors. Focused shrinkage estimators (fd3 and fp3) that lie on the top of each show the best performance across all the dimensions. This is in stark contrast with the competitors. Highest losses are shown by equally-weighted and Jagannathan and Ma (2003) for portfolios of up to 50 assets and by sample covariance and Frahm and Memmel (2010) for portfolios from 50 assets. This is to be expected, since equally-weighted typically portfolio does well when the estimation error is too large to be reduced by a statistical model; it appears that for portfolios of up to 50 assets existing statistical methods are able to capture the covariance structure reasonably well. As the dimension gets larger, equally-weighted estimators starts to perform better relative to its competitors. Jagannathan and Ma (2003) estimator's performance is also not surprising since this method forces weights to be sparse through the 1-Norm constraint which is expected to do better in portfolios of large sizes. The performance of the sample covariance matrix in a large portfolio is also anticipated, since estimation error of the sample covariance is large when p is close to n.

Performance of the 3-factor model and 1-Norm constrained portfolio (NC1V) is worse compared to our focused estimators. Factor model performance is in line with expectations since our data-generating process is based on this specification. It does worse than focused estimator which does not enforce residual covariance matrix to be diagonal. 1-Norm constrained portfolio (NC1V) also does well relative to our estimators and other competitors. It also does particularly well relative to 1-Norm-Constrained portfolio with return-maximizing cross-validation (NC1R) and Jagannathan and Ma (2003) estimators. This is because it involves calibrating the threshold parameter δ toward variance minimization.

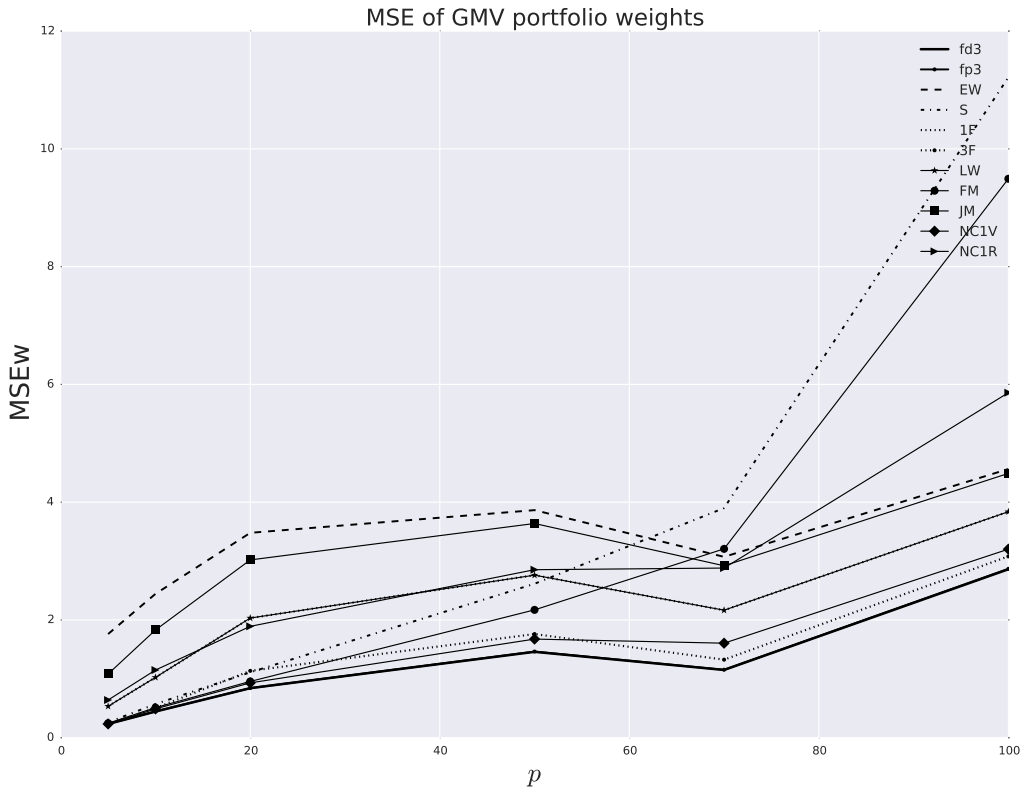

Figure 3.4.1 – Mean squared error (MSE) of the GMV portfolio weights of different estimators for portfolios of different sizes p (X-axis).

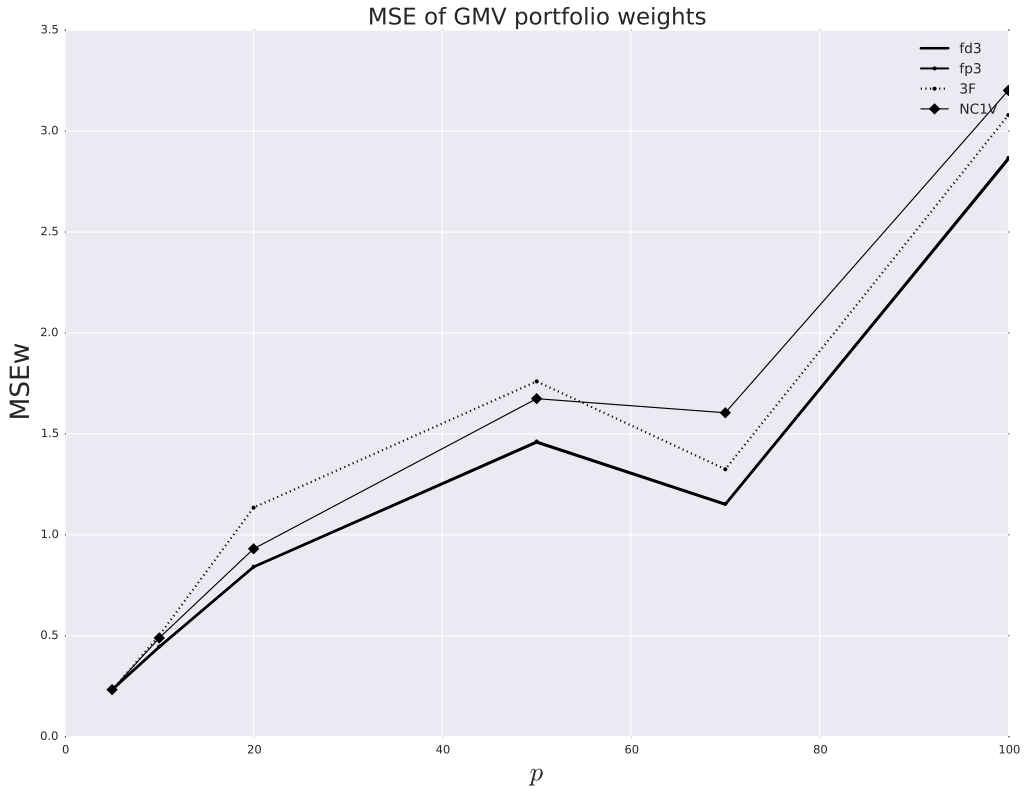

Figure 3.4.2 – Mean squared error (MSE) of the GMV portfolio weights of the four best estimators for portfolios of different sizes p (X-axis).

We now turn to the resulting losses in the GMV portfolio variances reported in figure 3.4.3. The four best competitors are reported in figure 3.4.4. The results confirm the superior performance of the focused shrinkage estimators (fd3 and fp3) which have uniformly lower loss compared to all the competitors in portfolios of various sizes. The highest losses are displayed by equally weighted and Jagannathan and Ma (2003) portfolios and Ledoit and Wolf (2003) for the portfolios of larger sizes. As before, the 3-factor model and 1-Norm-Constrained portfolio with variance-minimizing cross-validation are second best.

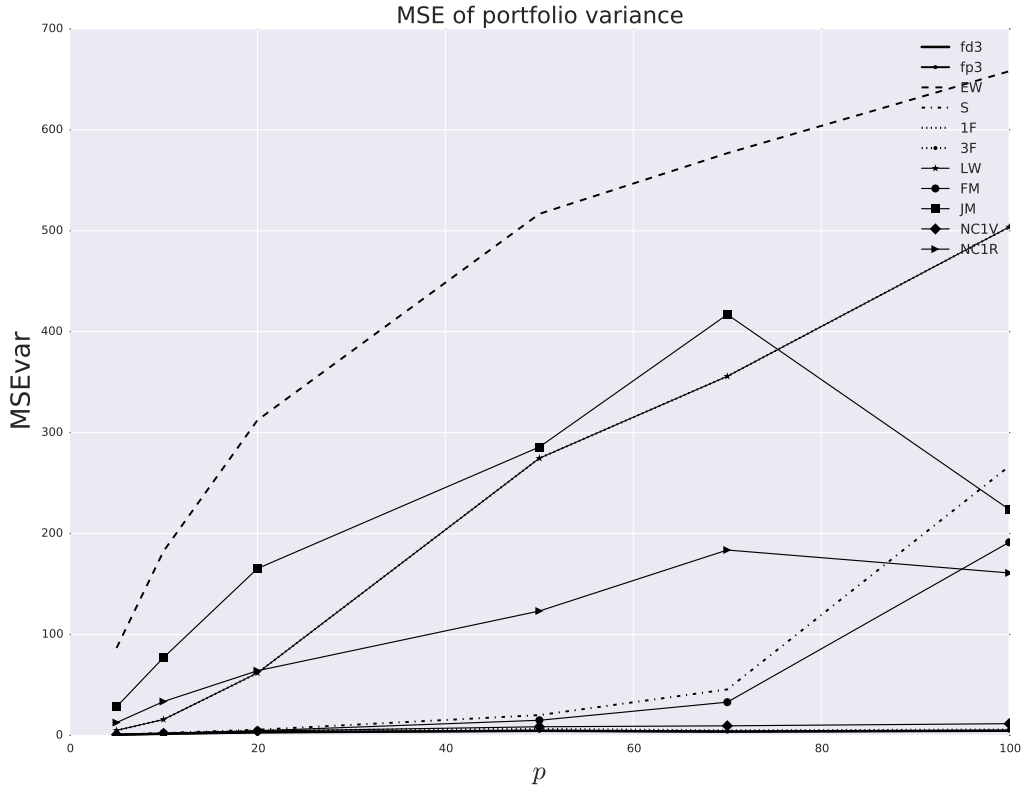

Figure 3.4.3 – Mean squared error (MSE) of the GMV portfolio variance of different estimators for portfolios of different sizes p (X-axis).

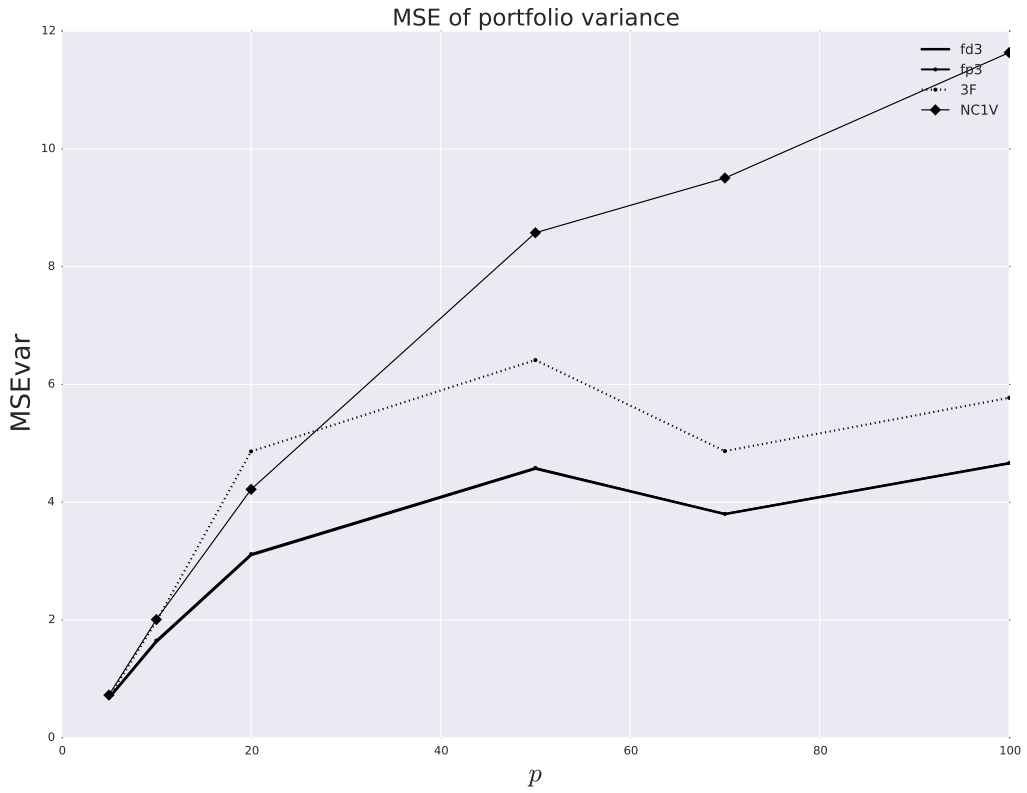

Figure 3.4.4 – Mean squared error (MSE) of the GMV portfolio variance of different estimators for portfolios of different sizes p (X-axis)

To sum up, focused shrinkage estimators show superior performance compared to competitors. They have the lowest mean squared error of the GMV portfolio weights and the resulting portfolio variance for the portfolios of all sizes that we consider.

3.5 Empirical application

3.5.1 Data

In this section we illustrate performance of our focused estimator and its competitors on 3 different data sets outlined in Table 3.3. For 10Ind and 25FF data sets the starting period is 1963/07. For the 48Ind data set the starting period is 1970-01 because some of the 48 industry portfolios have missing observations before that date[7].

[7]We also considered similar starting dates that had no effect on our results.

No.	Data set	Abbreviation	p	Time period
1	10 industry portfolios representing the U.S. stock market	10Ind	10	1963/07 - 2015/12
2	48 industry portfolios representing the U.S. stock market	48Ind	48	1970/01 - 2015/12
3	25 Fama and French (1992) portfolios of firms sorted by size and book-to-market	25FF	25	1963/07 - 2015/12

Table 3.3 – Datasets. This table presents the data sets used in the empirical application. All of them are taken from the Kenneth French data library: `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. For 10Ind, 25FF and 100FF data sets the starting period is 1963/07. For 48Ind data set the starting period is 1970-01 because some of the 48 industry portfolios have missing observations.

3.5.2 Methodology of performance evaluation

We use "the rolling-window" approach for our evaluation. First, we choose the window size $w < n$ over which we perform the estimation. In our case we use $w = 120$ of monthly observations which corresponds to 10 years of data. Next, using the return data corresponding to the chosen window we compute various estimators of the GMV portfolios outlined in Table 3.2. Finally, we repeat this "rolling-window" estimation by adding the data for the next month and dropping the data for the earliest month. We continue doing this till we use all the available data. As a result of this procedure, we obtain $n - w$ portfolio-weight vectors for each strategy i.e. x_t for $t = w, \ldots, n-1$. The out-of-sample return at time $t+1$ is given by $x_t^{'} R_{t+1}$. Using these, we compute the following performance metrics: out-of-sample portfolio variance, out-of-sample portfolio Sharpe ratio, and portfolio turnover

$$\begin{aligned} Var_P &= \frac{1}{n-w-1} \sum_{t=w}^{n-1} \left(x_t^{'} R_{t+1} - \mu \right)^2 \\ \mu &= \frac{1}{n-w} \sum_{t=w}^{n-1} x_t^{'} R_{t+1} \\ SR &= \frac{\mu}{Var_P} \\ Turnover &= \frac{1}{n-w-1} \sum_{t=w}^{n-1} \sum_{t=w}^{p} \left(\left| x_{j(t+1)} - x_{jt} \right| \right) \end{aligned} \tag{3.21}$$

where x_{jt} denotes portfolio weight for asset $j = 1, \ldots, p$ at time t.

3.5.3 Empirical results

Out-of-sample portfolio variances

Table 3.4 report out-of-sample variances (in 10^4) for the estimated portfolios. For all the data sets considered, focused shrinkage estimators based on 3-factor and 4-factor models are consistently top-ranked being first best in 10Ind and second best in 25FF and 48Ind. Focused shrinkage estimators based on the 1-factor model display worse performance. This is due to the fact that in 1-factor model, residual covariance contains the co-movement between assets related to common factors like size and value and therefore does not conform to both diagonal and pooling restricted models considered in our application. In contrast, the residual covariance matrix in 3-factor and 4-factor models has more 'white-noise' type of structure that makes it a reasonable shrinkage target for our focused shrinkage. This results in considerable improvements in performance. Further, using diagonal covariance model has a slightly better performance compared to the pooled covariance model for 25FF and 48Ind, but worse for 10Ind.

Interestingly, GMV estimators based on 1- and 3-factor model covariances display worse performance compared to the focused shrinkage estimators. The former ones set off-diagonal elements of the residual covariance in (3.10) to zero which contrasts with the focused shrinkage that weights between a restricted and an unrestricted models. This suggests of the misspecification of the factor models that is alleviated using our proposed approach.

Turning to other competitors, the 1-Norm-Constrained estimator with cross-validation for variance (NC1V) is another top-ranked competitor that emerges as first best in 25FF and 48Ind and fourth best in 10Ind. We expect Norm-constrained estimators to perform better in larger portfolios i.e. 25FF and 48Ind data sets. For 25FF data set there are 325 unique covariance matrix elements and with 120 observations the focused shrinkage estimator that relies on standard asymptotics ($n \to \infty$ with p fixed) may not provide good approximations. The same argument applies for 48Ind data set that implies 1176 unique covariance matrix terms. Beyond that, the second-best performance of the focused shrinkage in these data sets indicates that the chosen shrinkage target - the residual covariance matrix - does not always conform to the diagonal or equicorrelation matrices that we used as our restricted models. This might indicate that 3- and 4-factor models employed are misspecified to a larger degree in these data sets.

Frahm and Memmel (2010) estimator appears is the third best. This to be expected, since this estimator is designed to minimize out-of-sample portfolio variance and hence can be viewed as focused estimator. However, since its properties rely and normally distributed and i.i.d. returns which are violated in the data sets, their estimator performs worse than our focused shrinkage estimator which does not impose any assumptions on the data. The plug-in estimator using sample covariance matrix has reasonable performance followed by Ledoit and Wolf (2003) and the short sale-constrained portfolio by Jagannathan and Ma (2003). The equally-weighted portfolio also does not display good relative performance.

Portfolio	25FF	10Ind	48Ind
Panel A: Focused shrinkage estimators			
fd1	15.51	12.80	12.43
fp1	15.52	12.79	12.45
fd3	14.74	12.82	12.31
fp3	14.75	12.80	12.34
fd4	14.74	12.78	12.32
fp4	14.75	12.76	12.35
Panel B: Existing estimators			
EW	25.89	18.57	21.93
S	15.24	12.99	15.91
1F	26.81	13.92	16.11
3F	15.09	12.93	13.22
LW	19.53	13.46	14.35
FM	14.90	12.86	14.16
JM	19.06	13.96	18.31
NC1V	14.32	12.99	11.85
NC1R	18.48	13.94	16.85

Table 3.4 – Out-of-sample portfolio variances$\times 10^4$. Note. This table reports Out-of-sample portfolio variances (3.21) reported in 10^4 computed for different different competitors (table 3.2) across different data sets (table 3.3). Focused shrinkage estimators (Panel A) based on 3-factor and 4-factor models are consistently top-ranked being first best in 10Ind and second best in 25FF and 48Ind. 1-Norm constrained estimator with cross-validation for variance (NC1V) is another top-ranked competitor that emerges as first best in 25FF and 48Ind and fourth-best in 10Ind. This estimator, however, has considerably larger turnover that makes is costly to trade (table 3.6). Overall, focused estimators display most robust and stable performance.

Out-of-sample Sharpe Ratios

Out-of-sample Sharpe Ratios are reported in table 3.5. Focused estimators show reasonable performance on this metric across all the data sets being in the middle of the ranking. This is not surprising since our focused estimator is designed to for variance minimization, not Sharpe

Ratio maximization. In Chapter 2 we show that applying focused shrinkage method to the portfolio optimizer that targets Sharpe Ratio after transaction costs results in our estimator being consistently top-ranked on Sharpe Ratios and turnover levels. Focused shrinkage outperforms the market which has a Sharpe Ratio of 0.1248 over the corresponding period. In terms of other competitors, the sample covariance matrix and Frahm and Memmel (2010) estimators show the best performance in 25FF, 3-factor model and Jagannathan and Ma (2003) display the best performance in 10Ind and 1- and 3-factors are top-performing in 48Ind.

Portfolio	25FF	10Ind	48Ind
Panel A: Focused shrinkage estimators			
fd1	0.2191	0.1803	0.1505
fp1	0.2190	0.1806	0.1488
fd3	0.2226	0.1796	0.1482
fp3	0.2229	0.1796	0.1472
fd4	0.2232	0.1797	0.1484
fp4	0.2234	0.1798	0.1473
Panel B: Existing estimators			
EW	0.1885	0.1759	0.1869
S	0.2707	0.1812	0.1316
1F	0.1239	0.1847	0.2079
3F	0.2271	0.1970	0.2062
LW	0.1578	0.1792	0.1797
FM	0.2679	0.1839	0.1578
JM	0.1708	0.1911	0.2052
NC1V	0.2568	0.1827	0.1755
NC1R	0.1816	0.1575	0.1379

Table 3.5 – Out-of-sample Sharpe Ratios. This table reports out-of-sample Sharpe Ratios (3.21) computed for different competitors (table 3.2) across different data sets (table 3.3). Focused estimators of covariance matrix show reasonable performance on this metric across all the data sets being in the middle of the ranking. This is not surprising since our focused estimator is designed to for variance minimization, not Sharpe Ratio maximization. In Chapter 2 we show that applying focused shrinkage method to the portfolio optimizer that targets Sharpe Ratio after transaction costs results in our estimator being consistently top-ranked on Sharpe Ratios and turnover levels. Focused shrinkage outperforms market that has Sharpe Ratio of 0.1248 over the same period.

Turnovers

Table 3.6 displays the turnover for different estimators. For the focused shrinkage estimators based on 3- and 4-factor models, turnover levels are at the middle of the ranking. In particular, they are higher compared to the equally-weighted, 1- and 3-factor models, and Ledoit and Wolf (2003) estimator. Focused shrinkage estimators outperform sample covariance and Frahm

and Memmel (2010) estimators by a small margin and 1-Norm constraint portfolio with variance minimizing cross-validation (NC1V) by a large margin in 25FF and 10Ind portfolios. Jagannathan and Ma (2003) and 1-Norm-Constrained portfolio with return maximizing cross-validation (NC1R) show the highest turnovers in 25FF and 10Ind; the former improves considerably in 48Ind.

To sum up, empirical results are generally consistent with simulations. Our focused estimator displays the lowest and second-lowest out-of-sample portfolio variance and reasonable out-of-sample Sharpe Ratios and turnovers. It is only outperformed in 25FF and 48Ind by 1-Norm constraint portfolio with variance minimizing cross-validation (NC1V) that is, however, 4th best in 10Ind and has considerably larger turnover levels which makes it costly to trade with implications for its Sharpe Ratios.

Portfolio	25FF	10Ind	48Ind
Panel A: Focused shrinkage estimators			
fd1	0.7249	0.1284	0.4061
fp1	0.7260	0.1279	0.4104
fd3	0.5584	0.1273	0.3940
fp3	0.5602	0.1260	0.3989
fd4	0.5643	0.1254	0.3874
fp4	0.5652	0.1238	0.3921
Panel B: Existing estimators			
EW	0.0016	0.0016	0.0016
S	0.7120	0.1368	0.7175
1F	0.1290	0.0649	0.1165
3F	0.3034	0.0863	0.1744
LW	0.1794	0.0728	0.1486
FM	0.6408	0.1277	0.6033
JM	0.6687	0.1579	0.0441
NC1V	0.8926	0.2643	0.3941
NC1R	3.5054	0.8013	2.5119

Table 3.6 – Portfolio Turnovers. This table reports turnovers (3.21) computed for different competitors (table 3.2) across different data sets (table 3.3). For the focused shrinkage estimators based on 3- and 4-factor models, turnover levels are at medium of the ranking In particular, they are higher compared to the equally-weighted, 1- and 3-factor models, and Ledoit and Wolf (2003) estimator. Focused shrinkage estimators outperform sample covariance and Frahm and Memmel's (2010) estimators by a small margin and 1-Norm constraint with cross-validation for variance (NC1V) by a large margin in 25FF and 10Ind portfolios.

3.6 Conclusion

In this paper we proposed a shrinkage estimator for the covariance matrix based on a factor model. Our estimator is designed to minimize the mean squared error of the Global Minimum Variance portfolio weights. Focused shrinkage estimator weights between an unrestricted and restricted estimators of the residual covariance - our chosen shrinkage target. We provide the optimal shrinkage intensity for the specific case of the Global Minimum Variance portfolio. We then illustrate the performance of our proposed method by conducting extensive simulations designed to realistically represent U.S. stock market dynamics. Our focused shrinkage estimator shows the best performance in terms of out-of-sample portfolio variances compared to all the competitors considered. Empirical applications to the sorted Fama and French and industry portfolios of different sizes show similar improvements. Our focused estimator displays the lowest and second-lowest out-of-sample portfolio variance and reasonable out-of-sample Sharpe Ratios and turnovers.

3.7 Appendix

3.7.1 Ledoit and Wolf (2003) estimator

This is defined as:

$$\hat{S}_{LW} \;=\; \frac{k}{n}\,\hat{\Sigma}_{1F} + \left(1 - \frac{k}{n}\right)\hat{\Sigma}$$

with

$$k \;=\; \frac{p-r}{c}$$

and

$$
\begin{aligned}
p &= \textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij} \\
p_{ij} &= \frac{1}{T}\textstyle\sum_{t=1}^{T}\left\{\left(R_{it}-\bar{R}_i\right)\left(R_{jt}-\bar{R}_j\right)-\hat{\sigma}_{ij}\right\}^2
\end{aligned}
$$

$$
r = \textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} r_{ij}
$$

$$
r_{ij} = \begin{cases}
\frac{1}{T}\sum_{t=1}^{T} r_{ijt} & \\
r_{ijt} = \dfrac{\hat{\sigma}_{jm}\sqrt{\hat{\sigma}_{mm}}\left(R_{it}-\bar{R}_i\right)+\hat{\sigma}_{im}\sqrt{\hat{\sigma}_{mm}}\left(R_{jt}-\bar{R}_j\right)-\hat{\sigma}_{im}\hat{\sigma}_{jm}\left(r_{mt}-\bar{r}_m\right)}{\hat{\sigma}_{mm}} & \text{for } i \neq j \\
\qquad \times\left(r_{mt}-\bar{r}_m\right)\left(R_{it}-\bar{R}_i\right)\left(R_{jt}-\bar{R}_j\right)-\hat{f}_{ij}\hat{\sigma}_{ij} & \\
p_{ii} & \text{for } i = j
\end{cases}
$$

$$
\begin{aligned}
c &= \textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij} \\
c_{ij} &= \left(\hat{f}_{ij}-\hat{\sigma}_{ij}\right)^2
\end{aligned}
$$

3.7.2 Frahm and Memmel (2010) estimator

The estimator has the following form:

$$
\hat{x}_{FM} = k_S x_{EW} + (1-k_S)\,\hat{x}
$$

with

$$
\begin{aligned}
k_S &= \frac{p-3}{n-p+2}\cdot\frac{1}{\hat{\tau}_R} \\
\hat{\tau}_R &= \frac{GMV\left(x_{EW},\hat{\Sigma}\right)-GMV\left(\hat{x},\hat{\Sigma}\right)}{GMV\left(\hat{x},\hat{\Sigma}\right)} \\
GMV\left(\hat{x},\hat{\Sigma}\right) &= \frac{1}{\mathbf{1}'\hat{\Sigma}^{-1}\mathbf{1}} \\
GMV\left(x_{EW},\hat{\Sigma}\right) &= \frac{1}{p^2}\mathbf{1}'\hat{\Sigma}\mathbf{1}
\end{aligned}
$$

3.7.3 De Miguel et al. (2009) cross-validation strategies

DeMiguel et al. (2009) consider adding extra constraint to the GMV problem in (3.1). They consider two types of constraints: Norm-1 and Norm-A defined as follows:

$$\begin{aligned} \text{Norm-1: } \|x\|_1 &= \textstyle\sum_{i=1}^{N} |x_i| \leq \delta \\ \text{Norm-A: } \|x\|_A &= \left(x' A x\right)^{1/2} \leq \delta \end{aligned}$$

The cross-validation procedure for determining optimal δ works as follows:

1. For each δ on an equally-spaced grid of 10 points in range $\delta \in \left[1, \bar{\delta}\right]$ with $\bar{\delta} = \|\hat{x}\|_1$[8] do the following.

2. Given $n \times p$ matrix of returns R, for each $t = 1, \ldots, n$:

   (a) Delete the tth sample return from R. Denote the resulting return matrix without row t by $R_{(t)}$.

   (b) Use $R_{(t)}$ to compute the corresponding sample covariance matrix; denote it by $\hat{\Sigma}_{(t)}$.

   (c) Use $\hat{\Sigma}_{(t)}$ and δ to compute the corresponding 1-Norm-Constrained GMV portfolio (3.5); denote it by $\hat{x}_{NC1(t)}$.

   (d) Compute out-of-sample return on that portfolio: $R_{\delta,(t)} = \hat{x}'_{NC1(t)} R_t$.

3. Compute the variance of the out-of-sample portfolio returns: $Var_{P,\delta} = \frac{1}{n} \sum_{t=1}^{n} \left(R_{\delta,(t)} - \bar{R}_\delta\right)^2$ where $\bar{R}_\delta = \frac{1}{n} \sum_{t=1}^{n} R_{\delta,(t)}$.

4. Find δ^* that minimizes the out-of-sample variance: $\delta^* = \arg\min_{\delta} Var_{P,\delta}$.

If the objective is to maximize the last portfolio return, then for each δ on an equally-spaced grid of 10 points in range $\delta \in \left[1, \bar{\delta}\right]$ we find δ^*such that

$$\delta^* = \arg\max_{\delta} \hat{x}_{NC1,\delta} r_n$$

where $\hat{w}_{NC1,\delta}$ is the 1-Norm-Constrained estimated using sample covariance matrix $\hat{\Sigma}$ and threshold δ, and R_n is the last period return for a given sample.

[8] we bound range of delta by $\bar{\delta} = \|\hat{x}\|_1$ since for value $\delta > \bar{\delta}$ 1-Norm-Constrained optimization produces solution equivalent to unconstrained GMV optimization (3.1) with the sample covariance matrix.

Bibliography

[1] Abadir, K.M., Magnus, J.R. (2005), *Matrix algebra*, Vol. 1, New York: Cambridge University Press.

[2] Aguilar, O. and West, M. (2000) 'Bayesian dynamic factor models and portfolio allocation', *Journal of Business and Economic Statistics*, 18, 338–357.

[3] Bai, J. (2003) 'Inferential theory for factor models of large dimensions', *Econometrica*, 71(1), 135–171.

[4] Berger, J. O. (1976a) 'Admissible Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss', *Annals of Statistics*, 4, 223-226.

[5] Berger, J. O. (1976b) 'Minimax Estimation of a Multivariate Normal Mean Under Arbitrary Quadratic Loss', *Journal of Multivariate Analysis*, 6, 256-264.

[6] Berger, J. O. (1982) 'Selecting a Minimax Estimator of a Multivariate Normal Mean', *Annals of Statistics*, 10, 81-92.

[7] Berger, J. O., Bock, M. E., Brown, L. D., Casella, G., and Gleser, L. (1977) 'Minimax Estimation of a Normal Mean Vector for Arbitrary Quadratic Loss and Unknown Covariance Matrix', *Annals of Statistics*, 5, 763-771.

[8] Best, M. J. and Grauer, R. R. (1991) 'On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results', *Review of Financial Studies* 4, 315–342.

[9] Bhattacharya, P. K. (1966) 'Estimating the Mean of a Multivariate Population with General Quadratic Loss Function', *Annals of Mathematical Statistics*, 37, 1819-1824.

[10] Bock, M. E. (1975) 'Minimax Estimators of the Mean of a Multivariate Normal Distribution', *Annals of Statistics*, 3, 209-218.

[11] Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics*, 31, 307–327.

[12] Bosq, D. (2012) *Nonparametric statistics for stochastic processes: estimation and prediction*, Vol. 110, Springer Science & Business Media.

[13] Broadie, M. (1993) 'Computing efficient frontiers using estimated parameters', *Annals of Operations Research*, 45, 21–58.

[14] Burnham, K., Anderson, D. R. (2006) *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*, New York: Springer-Verlag.

[15] Carhart, M. M. (1997) 'On persistence in mutual fund performance', *Journal of Finance*, 52, 57–82.

[16] Chamberlain, G. and Rothschild, M. (1982) 'Arbitrage, factor structure, and mean-variance analysis on large asset markets', *Econometrica*, 51, 1281-1304.

[17] Chamberlain, G. (1983) 'Funds, factors, and diversification in arbitrage pricing models', *Econometrica*, 51, 1305–1323.

[18] Chopra, V. K. and Ziemba, W. T. (2011) 'The effect of errors in means, variances, and covariances on optimal portfolio choice', *Journal of Portfolio Management*, 19, 6-11.

[19] Claeskens, G., and Hjort, N. L. (2003) 'The Focused Information Criterion', *Journal of the American Statistical Association*, 98, 900-916.

[20] Claeskens, G., Magnus, J. R., Vasnev, A. L., Wang, W. (2016) 'The forecast combination puzzle: A simple theoretical explanation', *International Journal of Forecasting*, 32, 754–762.

[21] Clark, T.E., McCracken, M.W. (2009) 'Combining Forecasts from Nested Models', *Oxford Bulletin of Economics and Statistics*, 73, 303–329.

[22] Clarke, R., De Silva, H. and Thorley, S. (2006) 'Minimum-variance portfolios in the us equity market', *Journal of Portfolio Management*, 33, 10-24.

[23] Clarke, R., De Silva, H. and Thorley, S. (2011) 'Minimum-variance portfolio composition', *Journal of Portfolio Management*, 37, 31-45.

[24] Collin-Dufresne, P., Daniel, K., Moallemi, C., Saglam, M. (2015) 'Dynamic Asset Allocation with Predictable Returns and Transaction Costs', available at [SSRN: http://ssrn.com/abstract=2618910 or http://dx.doi.org/10.2139/ssrn.2618910].

[25] DeMiguel, V., Garlappi, L., Nogales, F. J. and Uppal, R. (2009) 'A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms', *Management Science*, 55, 798–812.

[26] DeMiguel, V., Garlappi, L., Uppal, R. (2009) 'Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?', *The Review of Financial Studies*, 22, 1915-1953.

[27] DeMiguel, V., Mei, X., Nogales, F. (2016) 'Multiperiod Portfolio Optimization with Many Risky Assets and General Transaction Costs', *Journal of Banking and Finance*, 69, 108-120.

[28] DeMiguel, V., Nogales, F., Uppal, F (2014) 'Stock Return Serial Dependence and Out-of-Sample Portfolio Performance', *Review of Financial Studies*, 27, 1031-1073.

[29] Elliot, G., Gargano, A., Timmermann, A. (2013) 'Complete Subset Regressions', *Journal of Econometrics*, 177, 357-373.

[30] Elliot, G., Gargano, A., Timmermann, A. (2015) 'Complete Subset Regressions with Large - Dimensional Sets of Predictors', *Journal of Economic Dynamics and Control*, 54, 86-110.

[31] Engle, R. and Watson, M. (1981) 'A one-factor multivariate time series model of metropolitan wage rates', *Journal of the American Statistical Association*, 76, 774–781.

[32] Engle, R. (2002) 'Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models', *Journal of Business and Economic Statistics* 20, 339–350.

[33] Engle, R. F. (1982) 'Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation', *Econometrica*, 987–1007.

[34] Fama, E. F. & French, K. R. (1992) 'The cross-section of expected stock returns', *Journal of Finance*, 47, 427–465.

[35] Fama, E. F. & French, K. R. (1993) 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics*, 33, 3–56.

[36] Fan, J., Fan, Y. and Lv, J. (2008) 'High dimensional covariance matrix estimation using a factor model', *Journal of Econometrics* , 147, 186–197.

[37] Fan, J., Liao, Y. and Liu, H. (2016) 'An Overview of the Estimation of Large Covariance and Precision Matrices', *Econometrics Journal*, 19, C1-C32.

[38] Fan, J., Zhang, J. and Yu, K. (2012) 'Vast portfolio selection with gross-exposure constraints', *Journal of the American Statistical Association*, 107, 592–606.

[39] Fang, Y., Loparo, K.A. and Feng, X. (1994) 'Inequalities for the trace of matrix product', *IEEE Transactions on Automatic Control*, 39, 2489-2490.

[40] Fomby, T. B., and Hill, R. C. (1979) 'Multicollinearity and the Minimax Conditions of the Bock Stein-Like Estimate', *Econometrica*, 47, 211-212.

[41] Frahm, G. and Memmel, C. (2010) 'Dominating estimators for minimum-variance portfolios', *Journal of Econometrics*, 159, 289–302.

[42] Frost, P. A. and Savarino, J. E. (1986) 'An empirical bayes approach to efficient portfolio selection', *Journal of Financial and Quantitative Analysis*, 21, 293–305.

[43] Frost, P. A. and Savarino, J. E. (1988) 'For better performance: Constrain portfolio weights', *The Journal of Portfolio Management*, 15, 29–34.

[44] Garleanu, N., and Pedersen, L. H. (2013) 'Dynamic Trading with Predictable Returns and Transaction Costs', *Journal of Finance*, 68, 2309-2340.

[45] Geweke, J., Amisano, G. (2011) 'Optimal prediction pools', *Journal of Econometrics*, 164, 130-141.

[46] Hall, P., Heyde, C.C. (1980) *Martingale Limit Theory and its Applications*, New York: Academic Press. Hamilton, J. D. (1994) Time Series Analysis, Princeton: Princeton University Press.

[47] Hansen, B (2007) 'Least Squares Model Averaging', *Econometrica*, 75, 1175-1189.

[48] Hansen, B. E. (2016) 'Efficient Shrinkage in Parametric Models', *Journal of Econometrics*, 190, 115-132.

[49] T. Hastie, R. Tibshirani, and J. Friedman (2001) *The Elements of Statistical Learning*, New York: Springer Series in Statistics.

[50] Haugen, R. A. and Baker, N. L. (1991) 'The efficient market inefficiency of capitalization weighted stock portfolios', *Journal of Portfolio Management*, 17, 35–40.

[51] Haugen, R. A. (1990) 'Building a better index: Cap-weighted benchmarks are inefficient vehicles', *Pensions & Investments*, October 1.

[52] Hjort, N.L., Claeskens, G. (2003a) 'Frequentist model average estimators', *Journal of American Statistical Association*, 98, 879–899.

[53] Hjort, N.L., Claeskens, G. (2003b) 'Rejoinder to the focused information criterion and frequentist model average estimators', *Journal of American Statistical Association*, 98, 938–945

[54] Horn, R.A, Johnson, C.R. (2012) *Matrix analysis*, New York: Cambridge university press.

[55] Jagannathan, R. and Ma, T. (2003) 'Risk reduction in large portfolios: Why imposing the wrong constraints helps', *Journal of Finance*, 58, 1651–1684.

[56] James, W., and Stein, C. (1961) 'Estimation with Quadratic Loss', *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1, 361-379.

[57] Jorion, P. (1986) 'Bayes-stein estimation for portfolio analysis', *Journal of Financial and Quantitative Analysis*, 21, 279–292.

[58] Judge, G. G., and Bock, M. E. (1976) 'A Comparison of Traditional and Stein-Rule Estimators Under Weighted Squared Error Loss', *International Economic Review*, 17, 234-240.

[59] Kim, J., Pollard, D. (1990) 'Cube root asymptotics', *Annals of Statistics*, 18, 191–219.

[60] Ledoit, O. and Wolf, M. (2003) 'Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection', *Journal of Empirical Finance*, 10, 603–621.

[61] Ledoit, O. and Wolf, M. (2004) 'A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices', *Journal of Multivariate Analysis*, 88, 365–411.

[62] Liang, H., Zou, G., Wan, A., Zhang, X. (2011) 'Optimal weight choice for frequentist model average estimators', *Journal of American Statistical Association*, 106, 1053–1066.

[63] Litterman, B. (2004), Modern investment management: an equilibrium approach, John Wiley & Sons. Liu, C. (2014) 'Distribution theory of the least squares averaging estimator', *Journal of Econometrics*, 186, 142-159.

[64] Markowitz, H. (1952) 'Portfolio Selection', *Journal of Finance*, 7, 77-91.

[65] Michaud, R. O. (1989) 'The Markowitz optimization enigma: is' optimized'optimal?', *Financial Analysts Journal*, 45, 31–42.

[66] Moskowitz, T. J., Ooi, Y. H., and Pedersen, L. H. (2012) 'Time series momentum', *Journal of Financial Economics*, 104, 228-250.

[67] Newey, W. K., and McFadden, D. (1994) 'Large Sample Estimation and Hypothesis Testing', *Handbook of Econometrics*, Vol. 4, 2111-2245.

[68] Oman, S. D. (1982a) 'Contracting Towards Subspaces when Estimating the Mean of a Multivariate Normal Distribution', *Journal of Multivariate Analysis*, 12, 270-290.

[69] Oman, S. D. (1982b) 'Shrinking Towards Subspaces in Multiple Linear Regression', *Technometrics*, 24, 307-311.

[70] Ross, S. A. (1976) 'The arbitrage theory of capital asset pricing', *Journal of Economic Theory*, 13, 341–360.

[71] Ross, S. A. (1977) 'The capital asset pricing model (CAMP), short-sale restrictions and related issues', *Journal of Finance*, 32, 177–183.

[72] Saleh, A. K. Md. E. (2006) *Theory of Preliminary Test and Stein-Type Estimation with Applications*, Hoboken: Wiley.

[73] Sclove, S. L. (1968) 'Improved Estimators for Coefficients in Linear Regression', *Journal of the American Statistical Association*, 63, 596-606.

[74] Sharpe, W. F. (1963) 'A simplified model for portfolio analysis', *Management Science*, 9, 277–293.

[75] Stein, C. (1966) 'An Approach to the Recovery of Inter-Block Information in Balanced Incomplete Block Designs', *Research Papers in Statistics: Festschrift for J. Neyman*, 351-366.

[76] Stock, J. H. and Watson, M. W. (2005) 'Implications of dynamic factor models for var analysis', Technical report, National Bureau of Economic Research.

[77] Timmerman, A. (2006) 'Forecast Combinations', *Handbook of Economic Forecasting*, Amsterdam: Elsevier, Amsterdam, 135-195.

[78] van der Vaart, A. W. (1998) *Asymptotic Statistics*, New York: Cambridge University Press.

[79] Winston, K. J. (1993) 'The coefficient index and prediction of portfolio variance', *Journal of Portfolio Management*, 19, 27–34.