# Statistical Methods for the Analysis of Financial Risk

Inauguraldissertation zur Erlangung des Doktorgrades der

Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln

2023

vorgelegt von

Philipp Christian Hansen

aus

Köln

| | |
|---|---|
| Referent: | Prof. Dr. Jörg Breitung |
| Korreferent: | Prof. Dr. Dominik Wied |
| Tag der Promotion: | 21. Dezember 2023 |

# Danksagung

An dieser Stelle möchte ich mich bei den folgenden Personen bedanken, deren Unterstützung wesentlich zur Fertigstellung dieser Dissertationsschrift beigetragen hat.

Mein besonderer Dank gilt meinem Betreuer Professor Dr. Jörg Breitung. Seine Vorlesung hat während meines Masterstudiums maßgeblich mein Interesse an ökonometrischen Themen geweckt. Ich bedanke mich für die kontinuierliche Unterstützung, die gute Zusammenarbeit sowie die unschätzbaren Anmerkungen und Ratschläge während meines Promotionsstudiums.

Ich danke Professor Dr. Dominik Wied für seine Unterstützung und Anmerkungen sowie für die Übernahme des Zweitgutachtens. Er stand jederzeit für meine Fragen zur Verfügung.

Ich bedanke mich bei allen (ehemaligen) Kollegen und Kolleginnen am Institut für Ökonometrie und Statistik für die äußerst angenehme Arbeitsatmosphäre und das hohe Maß an Hilfsbereitschaft. Es war eine wundervolle Zeit, die mir immer in besonderer Erinnerung bleiben wird.

Ganz besonders möchte ich mich bei meinen Eltern, Großeltern, meiner Schwester und meinen Schwiegereltern für jegliche Art der Unterstützung bedanken. Besonders hervorheben möchte ich die zahlreichen Diskussionen mit meinem Vater, die maßgeblich zu meiner Motivation beigetragen haben. Zudem bedanke ich mich bei meinen Freunden aus Köln, die mir oft den notwendigen Ausgleich verschafft haben.

Nicht zuletzt gebührt der größte Dank meiner Frau Sarah. Sie verdient weit mehr als nur Dankbarkeit für ihre bedingungslose Unterstützung, ihren unerschütterlichen Glauben an mich sowie ihr Verständnis für lange Arbeitszeiten.

# Contents

# List of Tables

# List of Figures

# Chapter 1.

# Introduction

This thesis consists of three self-contained essays on statistical methods for modeling financial risk. Chapter 2 corresponds to the paper "Alternative estimation methods for the factor augmented panel data model with small $T$" by Breitung and Hansen (2021), published in *Empirical Economics*. Chapter 3 corresponds to the working paper "Empirical Challenges for Optimal Portfolio Selection", which is also a joint work with Jörg Breitung. Chapter 4 contains my single-authored working paper "Quantifying Downside Risk: A comparative Study of Value at Risk and Expected Shortfall".

The topics in Chapters 3 and 4 are directly related to the modeling of financial risks. In Chapter 3, we analyze portfolio models, which are closely linked to the minimization of portfolio risk. Chapter 4 examines the estimation of risk measures to quantify the downside risk of financial assets. Chapter 2 does not contain an empirical application, since the primary focus of this paper lies in the theoretical differences of alternative estimation methods and their comparative performance in simulation studies. However, the considered estimation methods are also relevant in the field of financial risk. Application examples can be found, for instance, in the context of exchange rate risks in Breitung and Mann (2017) or in the case of dynamic panel data models for leverage in Westerlund et al. (2022). In the following, the main findings of the three papers will be summarized and my contribution to Chapters 2 and 3 will be outlined.

In Chapter 2, alternative estimation methods for the factor augmented panel data model are compared. In contrast to traditional panel data models, factor augmented panel data models offer a highly flexible approach to account for cross-sectional dependence and time-varying heterogeneity in the error term. For instance, panel data often exhibit cross-sectional dependence, even after conditioning on relevant independent variables (Karabiyik et al., 2019a). Ignoring the cross-sectional dependence in the errors can lead to serious consequences such as misleading inference or even inconsistent estimates.

In the factor augmented panel data model, unobservable time-varying individual effects are modeled though a factor structure in the error term, where the factors affect

all cross-section units with different intensities. In Chapter 2, the focus lies on panel datasets where the number of cross-sections ($N$) is large relative to the number of time periods ($T$). In the comparison of different estimation methods, we include the principal component (PC) estimator of Bai (2009) and the common correlated effects (CCE) estimator proposed by Pesaran (2006). These estimators were originally developed for panel data with large $N$ and $T$. Additionally, we consider the GMM approaches introduced by Ahn et al. (2013) and Robertson and Sarafidis (2015), which assume that $T$ is small (that is $T$ is fixed in the asymptotic analysis).

Our comparison of these existing methods addresses three different issues. First, we analyze the possibility of an inappropriate normalization of the factor space (referred to as the normalization failure). The results indicate that the normalization conditions for the CCE estimator of Pesaran (2006) and the original ALS estimator of Ahn et al. (2013) can be problematic when the factors and loadings are close to a normalization failure. However, it is possible to adapt the estimation methods to improve the performance in such cases. In particular, we propose a variant of the CCE estimator that avoids the normalization failure by adapting a weighting scheme inspired by the analysis of Mundlak (1978). Secondly, we examine the impact of estimating versus fixing the number of factors in advance. We find that for small $T$, the selection criteria proposed by Bai and Ng (2002) and Ahn and Horenstein (2013) can yield inconsistent results, whereas the BIC criteria of Ahn et al. (2013) and Robertson and Sarafidis (2015) demonstrate robust performance. Thirdly, we demonstrate that the relative performance of these alternative estimation methods is highly influenced by the specific design of the Monte Carlo experiment, which helps to explain the conflicting findings from previous Monte Carlo studies.

My contribution to Chapter 2 is as follows: I developed the MATLAB codes and performed all Monte Carlo simulations. Regarding the writing process, I authored Section 2.7 on the Monte Carlo simulations and created the appendix. Prior to and during the publication process, I revisited the draft of the paper multiple times.

In Chapter 3, various challenges encountered in practical portfolio selection are examined. The maximum Sharpe ratio (MSR) portfolio, as proposed by Markowitz (1952), requires reliable estimates of expected returns and the covariance matrix of returns. Estimating these moments via their sample counterparts (the so-called plug-in method) yields extreme portfolio weights that fluctuate excessively over time and typically perform poorly out-of-sample (see, e.g., Michaud, 1989; Best and Grauer, 1991; Chopra and Ziemba, 1993; DeMiguel et al., 2009b). Obtaining reliable estimates is particularly problematic in case the number of investable assets ($N$) is of a similar magnitude as the

available amount of time series data ($T$). A common procedure is to ignore the information on the first moment, resulting in the estimation of the global minimum variance (GMV) portfolio, which generally improves the out-of-sample performance.

We analyze the impact of estimation uncertainties in both moments on portfolio performance. To mitigate the effects of errors in estimating the covariance matrix, several regularization methods have been proposed, such as shrinkage estimators or dimensionality reduction techniques like factor models. These methods can significantly improve the out-of-sample performance of the GMV portfolio, particularly in high-dimensional applications (e.g., Ledoit and Wolf, 2003, 2004a,b). Another approach to enhance the out-of-sample performance of both portfolios, MSR and GMV, is by regularizing the weights. Both portfolio models can be represented as regression models (Britten-Jones, 1999; Kempf and Memmel, 2006). This opens up the possibility, for example, to constrain the L1-norm of the weights through the application of LASSO techniques in the regression estimation. Another issue arises from the large number of negative weights that result from the estimation of the MSR and GMV portfolios. Traditionally, negative weights imply a short-selling strategy, which, however, leads to high risk exposure and turnover rates for the portfolio. As an alternative, we propose a put option strategy that reduces portfolio risk, turnover, and extreme weights.

In an empirical application, we analyze the out-of-sample performance of 15 portfolio models using two different datasets characterized by different concentration ratios ($N/T$). In cases where $T$ is substantially larger than $N$, the performance of the plug-in GMV estimator is competitive in comparison to other models. However, when the number of assets is considerable relative to the number of time periods, we demonstrate how alternative regularization approaches, such as LASSO or shrinkage methods, help to improve portfolio allocation in practice.

My contributions to Chapter 3 are as follows: Firstly, I performed all Monte Carlo simulations of Sections 3.4.1 and 3.4.3. Secondly, I collected and processed all the data for the empirical application. Thirdly, I am fully responsible for the empirical application in Section 3.5, where I implemented and performed all empirical exercises using the financial data set. Fourthly, I contributed to the theoretical development of our LASSO versions and introduced the concept of blocking strategies. In terms of the writing process, I authored Section 3.5 on the empirical application, the introduction, conclusion, and several paragraphs in Sections 3.2 and 3.4. Additionally, I created the appendix.

Chapter 4 focuses on the analysis of the two risk measures Value at Risk (VaR) and Expected Shortfall (ES) with regard to their theoretical differences and practical estimation. This comparison is particularly relevant, since ES at the 97.5% confidence level

replaces VaR at the 99% confidence level as the regulatory risk measure for calculating capital requirements according to the Basel III Accords. This transition is motivated by the fact that ES captures tail risks in contrast to VaR. Furthermore, ES is a subadditive risk measure unlike VaR (Artzner et al., 1999), a theoretical property related to risk diversification. On the other hand, backtesting ES appears to be more challenging due to its lack of elicitability (Gneiting, 2011). ES is also less robust with regard to model misspecifications and noise in the data in comparison to VaR (see, e.g., Cont et al., 2010; Kou et al., 2013; Kellner and Rösch, 2016).

Despite these theoretical differences, both risk measures have to be estimated in practice. This leaves the question whether the more complex estimation of ES provides any additional insights with regard to the quantification of risk. The results of the simulation and application study in Chapter 4 reveal that ES at the 97.5% confidence level is estimated with considerable more uncertainty compared to VaR, even when considering VaR at the higher 99% confidence level.

Under certain distributional assumptions, there exists a relationship between the two risk measures. For instance, under the assumption of a normal distribution, ES does not provide additional insights compared to VaR, as the corresponding VaR can be simply multiplied by a constant to obtain ES. In more realistic scenarios, the relationship between ES and VaR depends on the tail thickness of the distribution. In an empirical application, the performance of certain ratio models is examined. These models involve multiplying VaR by a factor to obtain ES estimates. Such ratio models either exhibit improved performance compared to ES benchmark models or offer comparable quality.

# Chapter 2.

# Alternative estimation approaches for the factor augmented panel data model with small T

## 2.1. Introduction

The seminal work of Holtz-Eakin et al. (1988) has provided two important contributions to the statistical analysis of panel data. First, it proposes a GMM framework for estimating dynamic panel data models that were further developed and popularized by Arellano and Bond (1991). This approach has become standard in the dynamic analysis of panel data. The second contribution, the introduction of time varying individual effects, was less influential and went largely unnoticed for many years. For example, the excellent monograph of Baltagi (2005) – as all other textbooks on panel data analysis of the early 2000s – does not consider time varying individual effects or any other factor structure. Bai (2009) pointed out that time varying individual effects are just a special case of a factor structure and provided a general framework for estimating a panel data model with "interactive fixed effects", which is also referred to as the *factor-augmented panel data model*.

With the work of Ahn et al. (2001, 2013), Pesaran (2006), and Bai (2009) the interest in models that account for time varying heterogeneity and cross-section correlation surged considerably and the 25th. International Conference on Panel Data in Vilnius 2019 included a large number of papers dealing with factor-augmented panel data models. In empirical practice, the Common Correlated Effects (CCE) approach proposed by Pesaran (2006) has recently become very popular among empirical researchers. This is due to the fact that this estimator is easy to understand and implement, a STATA routine (xtmg) and a Gretl add-on (xtcsd) is available and it performs well in Monte Carlo studies. It is however not clear, whether the CCE approach is similarly attractive

in empirical applications where the number of time periods $T$ is small (say 5 – 15). Ahn et al. (2013) and Robertson and Sarafidis (2015) proposed a GMM approach that is shown to be consistent for finite $T$, whereas the CCE and the Principal Component (PC) estimator were developed for samples with large $T$ and $N$. Su and Jin (2012) and Westerlund et al. (2019) showed that the CCE approach is consistent and asymptotically (mixed) normal if $T$ is fixed and $N \to \infty$, whereas the consistency of the PC estimator requires quite restrictive assumptions (such as i.i.d. errors across time) in this case. It is, however, not clear how large $T$ should be in order to ensure reliable estimation and inference.

An important assumption for the CCE estimator is that the (weighted) mean of the factor loadings is different from zero. This assumption is difficult to verify as the factor loadings are typically unknown. Furthermore, we show that the CCE estimator is already biased if the mean of the factor loadings is $O(N^{-1/2})$. To escape such a "normalization failure", we suggest a data dependent weighting scheme that is inspired by the Mundlak (1978) approach. In our Monte Carlo simulations we show that this simple weighting scheme performs well, whenever the original CCE estimator suffers from a normalization failure.

The rest of the paper is organized as follows. Section 2.2 compares the existing estimation methods and Section 2.3 reviews and complements the asymptotic results for fixed $T$ and $N \to \infty$. Possible problems with the normalization of the estimators are analyzed in Section 2.4. An extension to multiple factors is considered in Section 2.5 and empirical approaches for selecting the number of common factors are examined in Section 2.6. We argue that popular selection rules for the number of factors are generally inconsistent if $T$ is fixed. The small sample properties of alternative estimation procedures are investigated in Section 2.7. Specifically, we illustrate the detrimental effect of a normalization failure and demonstrate the robustness of the Mundlak type CCE estimator. Furthermore, we investigate the effects of estimating the number of factors on the performance of the estimation procedures. Finally, we employ three general model setups from the literature in order to compare the competing methods in more challenging and realistic scenarios. Section 2.8 concludes.

## 2.2. Existing estimation approaches

Consider the factor augmented panel data model:[1]

$$y_{it} = \boldsymbol{\beta}' \boldsymbol{x}_{it} + e_{it} \tag{2.1}$$

$$\text{with} \quad e_{it} = \lambda_i f_t + u_{it} \ , \tag{2.2}$$

where $\boldsymbol{x}_{it}$ and $\boldsymbol{\beta}$ are $k \times 1$ vectors. For the ease of exposition, we first consider a single factor with $r = 1$, that is, $f_t$ and $\lambda_i$ are scalars. The extension to multiple factors is considered in Section 2.5.

We adopt a "classical" panel data framework where the coefficient vector $\boldsymbol{\beta}$ is the same for all cross-section units (homogeneous panel). Furthermore, we assume that $T$ may be small relative to $N$, which is typical for many panel data applications. It should be noted that the asymptotic framework of Pesaran (2006) and Bai (2009) assumes that $N$ and $T$ tend to infinity, whereas Ahn et al. (2013) and Robertson and Sarafidis (2015) suppose that $T$ is small and fixed. Furthermore, the latter approach treats $f_t$ as parameters and thereby avoids making any assumptions on these parameters, whereas Pesaran (2006) and Bai (2009) assume that the factors are weakly correlated random variables and the loadings are treated as parameters (or also as random variables). We make the assumption that $u_{it}$ is independent (strictly exogenous) of $\boldsymbol{x}_{it}$, $f_t$ and $\lambda_i$. This rules out dynamic specifications.[2]

It is well known that in the two way panel data model the individual and time specific effects (which result as special cases of the factor model with constant factor and loading, respectively) can be removed by a simple data transformation, where the variables are adjusted by the individual and time specific averages. It is not difficult to see that a similar transformation exists for the model with interactive fixed effects, which is given by

$$y_{it} - \lambda_i \overline{y}_t(\boldsymbol{\lambda}) = \boldsymbol{\beta}' \left[ \boldsymbol{x}_{it} - \lambda_i \overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}) \right] + u_{it} - \lambda_i \overline{u}_t(\boldsymbol{\lambda}), \tag{2.3}$$

[1] The model may include further terms such as $\boldsymbol{\gamma}_i' \boldsymbol{d}_t$, where $\boldsymbol{d}_t$ is some observed strictly exogenous regressor, cf. Pesaran (2006). As such additional terms are easily accounted for without affecting the main results, these extensions are ignored.

[2] In panels with individual specific parameters and fixed $T$, including weakly dependent regressors (such as lagged dependent variables) results in a bias of order $1/T$ (the incidental parameter problem). The GMM based estimators of Section 2.2.3 are able to cope with this bias by introducing time dependent vectors of instruments. In this paper we abstract from such complications. The reader interested in dynamic models is referred to Juodis and Sarafidis (2018).

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ and

$$\overline{y}_t(\boldsymbol{\lambda}) = \frac{1}{N\overline{\lambda^2}} \sum_{i=1}^{N} \lambda_i y_{it}$$

with $\overline{\lambda^2} = N^{-1} \sum_{i=1}^{N} \lambda_i^2$. The weighted averages $\overline{\boldsymbol{x}}_t(\boldsymbol{\lambda})$ and $\overline{u}_t(\boldsymbol{\lambda})$ are constructed in an analogous manner. Note that $\overline{e}_t(\boldsymbol{\lambda}) = \overline{y}_t(\boldsymbol{\lambda}) - \boldsymbol{\beta}'\overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}) = f_t + \overline{u}_t(\boldsymbol{\lambda})$ serves as an estimate of $f_t$. Estimating the transformed regression (2.3) is equivalent to the least-squares estimator, treating $\boldsymbol{\beta}$ and $f_1, \ldots, f_T$ as parameters and $\boldsymbol{x}_{it}$ and $\lambda_i$ as regressors. Accordingly, the resulting estimator is efficient if $u_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

### 2.2.1. The PC estimator

For the PC approach suggested by Bai (2009), equation (2.3) is replaced by the feasible version

$$y_{it} - \widehat{\lambda}_i \overline{y}_t(\widehat{\boldsymbol{\lambda}}) = \boldsymbol{\beta}' \left[ \boldsymbol{x}_{it} - \widehat{\lambda}_i \overline{\boldsymbol{x}}_t(\widehat{\boldsymbol{\lambda}}) \right] + e_{it} - \widehat{\lambda}_i \overline{e}_t(\widehat{\boldsymbol{\lambda}}), \tag{2.4}$$

where $e_{it} = y_{it} - \boldsymbol{\beta}'\boldsymbol{x}_{it} = \lambda_i f_t + u_{it}$ and $\widehat{\lambda}_i$ denotes the PC estimator of the factor loading $\lambda_i$, which is equivalent to the eigenvector associated with the largest eigenvalue of the sample covariance matrix $\boldsymbol{\Omega}_{ee}(\boldsymbol{\beta}) = T^{-1} \sum_{t=1}^{T} \boldsymbol{e}_t(\boldsymbol{\beta}) \boldsymbol{e}_t(\boldsymbol{\beta})'$ with $\boldsymbol{e}_t(\boldsymbol{\beta}) = (y_{i1} - \boldsymbol{\beta}'\boldsymbol{x}_{i1}, \ldots, y_{iT} - \boldsymbol{\beta}'\boldsymbol{x}_{iT})'$. As shown by Moon and Weidner (2015) the sum of squared residuals can be obtained by minimizing the objective function

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin} \left\{ \mu_{\min} \left[ \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' \right] \right\} \tag{2.5}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$ and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$ and $\mu_{\min}\{\boldsymbol{A}\}$ denotes the smallest eigenvalue of the matrix $\boldsymbol{A}$. The minimum can be obtained by standard numerical methods, whereas Bai (2009) proposed to compute the (nonlinear) least-squares estimator of (2.4) sequentially by starting with the pooled OLS or within-group estimator of $\boldsymbol{\beta}$ (that is by ignoring the factor structure in the errors). The first principal component of the residual $e_{it}(\widehat{\boldsymbol{\beta}})$ yields a first estimator of the common factor and the associated loadings are used to obtain the estimated analog of the weighted averages in (2.4). The estimation procedure is iterated until the estimators converge to the least-squares estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$.

Moon and Weidner (2019) pointed out that the least-squares objective function may exhibit several local minima and therefore it is possible that the gradient based mini-

mization algorithm fails to find the global minimum. To cope with this problem, Moon and Weidner (2019) propose a nuclear norm penalty that results in a convex optimization problem. Another possibility is to initialize the minimization algorithm by a $\sqrt{NT}$-consistent initial estimator. In this case it is sufficient to assume convexity in the $1/\sqrt{NT}$ vicinity around the true value.

### 2.2.2. The CCE Estimator

In contrast to the PC estimator, the CCE approach proposed by Pesaran (2006) does not adopt an (asymptotically) efficient weighting scheme, but employs instead pre-specified weights $\boldsymbol{\lambda}_0$.[3] In practice $\boldsymbol{\lambda}_0 = (1, \dots, 1)'$ is the default option, but any other granular weighting scheme is possible. This gives rise to a modified transformation,

$$y_{it} - \lambda_i^* \, \overline{y}_t(\boldsymbol{\lambda}_0) = \boldsymbol{\beta}' \left[ \boldsymbol{x}_{it} - \lambda_i^* \, \overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}_0) \right] + u_{it} - \lambda_i^* \, \overline{u}_t(\boldsymbol{\lambda}_0), \tag{2.6}$$

where

$$\lambda_i^* = \lambda_i \frac{\sum_{i=1}^N \lambda_{0,i}^2}{\sum_{i=1}^N \lambda_{0,i} \lambda_i}$$

is required to drop the factor from the model. Note that if $\lambda_{0,i} = \lambda_i$ for all $i$, then $\lambda_i^* = \lambda_i$ and the transformation is equivalent to (2.3). Furthermore, if $\lambda_{0,i} = 1$ then $\lambda_i^* = \lambda_i / \overline{\lambda}$, where $\overline{\lambda} = N^{-1} \sum_{i=1}^N \lambda_i$. By reorganizing (2.6), we obtain the cross-section augmented regression equation,

$$y_{it} = \boldsymbol{\beta}' \boldsymbol{x}_{it} + \lambda_i^* \overline{y}_t(\boldsymbol{\lambda}_0) + \boldsymbol{\gamma}_i' \overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}_0) + v_{it} \ , \tag{2.7}$$

where $\boldsymbol{\gamma}_i = -\lambda_i^* \boldsymbol{\beta}$ and $v_{it} = u_{it} - \lambda_i^* \overline{u}_t(\boldsymbol{\lambda}_0)$. In practice, the nonlinear restriction $\boldsymbol{\gamma}_i = -\lambda_i^* \boldsymbol{\beta}$ is ignored and, therefore, $\boldsymbol{\gamma}_i$ is treated as an additional parameter.[4]

### 2.2.3. The HNR and ALS approach

While the CCE and PC approach replace the unobserved *factor* by (weighted) averages of $y_{1t}, \dots, y_{Nt}$ and $\boldsymbol{x}_{1t}, \dots, \boldsymbol{x}_{Nt}$, the approaches suggested by Holtz-Eakin et al. (1988)

[3] This does not imply, however, that the CCE estimator is always inefficient whenever $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}_0$. As shown by Westerlund et al. (2019) the CCE estimator is asymptotically efficient if $r = k + 1$ and $u_{it}$ is i.i.d. across $i$ and $t$.

[4] The restricted version of the CCE estimator is considered in Everaert and De Groote (2016). In our experience, imposing the nonlinear restriction does not result in an important gain in efficiency. In the model with $r > 1$ the restriction cannot be imposed anyway.

(HNR) and Ahn et al. (2013) (ALS) replace the unknown factor *loadings* by linear combinations of $y_{i1}, \ldots, y_{iT}$ and $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}$:

$$\text{HNR:} \quad \frac{1}{f_{t-1}}(y_{i,t-1} - \boldsymbol{\beta}'\boldsymbol{x}_{i,t-1}) \;=\; \lambda_i + \frac{1}{f_{t-1}}u_{i,t-1} \tag{2.8}$$

$$\text{ALS:} \quad \frac{1}{f_T}(y_{iT} - \boldsymbol{\beta}'\boldsymbol{x}_{iT}) \;=\; \lambda_i + \frac{1}{f_T}u_{iT}\ . \tag{2.9}$$

The main difference between these two approaches is that in (2.8) the linear combination is time dependent, whereas in (2.9) the linear combination is the same for all time series. As we do not see any advantage in using the variant HNR (and in our simulations the HNR estimator tends to perform worse than the ALS estimator), we focus on the ALS variant in the following analysis.

Inserting (2.9) in the model (2.1) yields

$$\text{ALS:} \quad y_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \theta_t y_{iT} - \theta_t\boldsymbol{\beta}'\boldsymbol{x}_{iT} + \nu_{it} \qquad \text{for } t = 1, \ldots, T-1, \tag{2.10}$$

where $\theta_t = f_t/f_T$ and $\nu_{it} = u_{it} - \theta_t u_{iT}$. Note that this approach involves $T-1$ additional parameters $\theta_1, \ldots, \theta_{T-1}$, whereas the CCE approach involves $N(k+1)$ additional parameters, which may be a much larger number of parameters, in particular if $N$ is large relative to $T$.

Equation (2.10) can be estimated as a linear equation by ignoring the nonlinear relationship $\boldsymbol{\delta}_t = \theta_t\boldsymbol{\beta}$ and treating $\boldsymbol{\delta}_t$ as additional parameters, cf. Hayakawa (2012). Furthermore, as the regressor $y_{iT}$ is correlated with the errors, an instrumental variable approach is required for estimating the coefficients efficiently. Since it is assumed that $\boldsymbol{x}_{it}$ is strictly exogenous, we employ observations of all time periods to construct the $Tk \times 1$ instrumental variable vector $\boldsymbol{z}_i = (\boldsymbol{x}_{i1}', \boldsymbol{x}_{i2}', \ldots, \boldsymbol{x}_{iT}')'$. The first stage regression yields $\widehat{y}_{iT} = \widehat{\boldsymbol{\pi}}'\boldsymbol{z}_i$, where $\widehat{\boldsymbol{\pi}}'\boldsymbol{z}_i$ is the fitted value from a regression of $y_{iT}$ on $\boldsymbol{z}_i$. The second stage regression is

$$y_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \theta_t\widehat{y}_{iT} - \theta_t\boldsymbol{\beta}'\boldsymbol{x}_{iT} + \nu_{it}.$$

Estimating the latter equation by OLS yields the two-stage least squares (2SLS) estimator. Since the error term $\nu_{it}$ is autocorrelated (due to the common component $\theta_t u_{iT}$), a GMM estimator based on the moment condition $\mathbb{E}(\boldsymbol{\nu}_i \otimes \boldsymbol{z}_i) = \mathbf{0}$ with $\boldsymbol{\nu}_i = (\nu_{i1}, \ldots, \nu_{iT})'$ is more efficient, in general.

### 2.2.4. The RS estimator

The GMM estimator of Robertson and Sarafidis (2015) results from multiplying the original model by the vector of instruments $\boldsymbol{z}_i$ (e.g. the instruments of the ALS estimator) such that

$$\boldsymbol{z}_i y_{it} = \left(\boldsymbol{z}_i \boldsymbol{x}_{it}'\right) \boldsymbol{\beta} + \left(\boldsymbol{z}_i \lambda_i\right) f_t + \boldsymbol{z}_i u_{it} \ .$$

The respective moment condition is given by

$$\mathbb{E}\left(\boldsymbol{m}_{zy} - \boldsymbol{M}_{zx}\boldsymbol{\beta} - \boldsymbol{f} \otimes \boldsymbol{\gamma}\right) = 0,$$

where

$$\boldsymbol{m}_{zy} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i y_{i1} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i y_{iT} \end{pmatrix} \qquad \boldsymbol{M}_{zx} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{x}_{i1}' \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{x}_{iT}' \end{pmatrix}$$

$$\boldsymbol{\gamma} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i \lambda_i \qquad \mathbf{f} = \begin{pmatrix} f_1 & \cdots & f_T \end{pmatrix}'.$$

Note that in this model the $N$ factor loadings $\lambda_1, \ldots, \lambda_N$ enter in form of the $Tk$ dimensional vector $\boldsymbol{\gamma}$, resulting in a considerable dimensionality reduction whenever $N$ is much larger than $T$. The GMM estimator results from minimizing the criterion function

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{f}) = \left(\boldsymbol{m}_{zy} - \boldsymbol{M}_{zx}\boldsymbol{\beta} - \boldsymbol{f} \otimes \boldsymbol{\gamma}\right)' \boldsymbol{W}_N \left(\boldsymbol{m}_{zy} - \boldsymbol{M}_{zx}\boldsymbol{\beta} - \boldsymbol{f} \otimes \boldsymbol{\gamma}\right), \tag{2.11}$$

where $\boldsymbol{W}_N$ is a consistent estimator of the optimal weighting matrix

$$\boldsymbol{W} = \left[\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} \widetilde{\boldsymbol{u}}_i \widetilde{\boldsymbol{u}}_i'\right)\right]^{-1}$$

with $\widetilde{\boldsymbol{u}}_i = \boldsymbol{m}_{zy} - \boldsymbol{M}_{zx}\beta - \boldsymbol{f} \otimes \boldsymbol{\gamma}$. Robertson and Sarafidis (2015) propose to minimize the function $Q(\cdot)$ by applying a sequential GMM estimator. Let $f_t^0$ denote some starting value. Replacing $f_t$ by $f_t^0$, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are obtained by linear GMM. Replacing $\boldsymbol{\gamma}$ by the respective GMM estimator, we obtain an updated estimator for $f_t$ by another linear GMM estimation step. This sequential GMM estimator eventually converges to the minimum of (2.11). An alternative estimator based on linear GMM is

proposed by Juodis and Sarafidis (2020).

It is important to notice that the first order condition of the GMM estimator is invariant to some scaling factor $c$, such as $\boldsymbol{f}^* = c\boldsymbol{f}$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}/c$. The PC estimator implies $c = 1/\sqrt{\sum_{t=1}^{T} f_t^2}$ and the original ALS estimator imposes $c = 1/f_T$. The objective function of the least-squares estimator does not impose any normalization of the factors. There exists a unique minimum for the product $\boldsymbol{f} \otimes \boldsymbol{\gamma}$, but the decomposition into $\boldsymbol{\gamma}$ and $\boldsymbol{f}$ is somewhat arbitrary and depends on the starting value of the iterative algorithm.

## 2.3. Asymptotic properties for fixed $T$

The asymptotic properties of the PC and CCE estimators are typically derived by adopting a joint limit theory, where $T$ and $N$ tend to infinity (e.g. Pesaran 2006, Bai 2009, Greenaway-McGrevy et al. 2012 and Westerlund and Urbain 2015). The asymptotic analysis revealed that the PC and CCE estimators are $\sqrt{NT}$-consistent whenever $\sqrt{T}/N \to 0$ and $\sqrt{N}/T \to 0$. This requirement is fulfilled if for some fixed constant, $0 < a < \infty$, the paths of the sample sizes admit the inequality $aT^{0.5+\epsilon} < N < aT^{2-\epsilon}$ for some $\epsilon > 0$. Statistical inference based on these estimators suffers from an asymptotic bias whenever $T/N \to \kappa > 0$. This bias does not show up in the asymptotic analysis of Pesaran (2006), as he assumes that the coefficient vector $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{v}_i$ is individual specific, where $\boldsymbol{v}_i$ is a random error that prevents the estimator from achieving the usual $\sqrt{NT}$ convergence rate. In the literature cited above, bias-corrected estimators are suggested that remove the asymptotic bias from the limiting distribution.

For fixed $T$ and $N \to \infty$ the CCE estimator of the factors is consistent as $\overline{e}_t(\boldsymbol{\lambda}_0)$ converges in probability to $cf_t$, where $c$ is some scale factor that is different from zero. Therefore, the errors-in-variable problem vanishes for $N \to \infty$ and fixed $T$ (cf. Westerlund et al. 2019).

For the asymptotic analysis of the PC estimator, it is usually assumed that $\min(N, T) \to \infty$ (cf. Bai 2009) and, therefore, the PC estimator may be inconsistent if $T$ is fixed and $N \to \infty$ (see Remark 1 of Bai 2009). Under more restrictive assumptions it is, however, possible to show that the PC estimator of the factors is consistent if $T$ is fixed and $N \to \infty$. To focus on the main issues assume that $\boldsymbol{\beta}$ is known. Furthermore, we assume that the vectors $\boldsymbol{f} = (f_1, \ldots, f_T)'$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ are parameter vectors to be

estimated. The PC estimator solves the first order conditions:

$$\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{e}_i - \widehat{\boldsymbol{f}}\widehat{\lambda}_i)\widehat{\lambda}_i = \mathbf{0} \quad \text{where } \boldsymbol{e}_i = (e_{i1},\ldots,e_{iT})' \tag{2.12}$$

$$\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{e}_t - \widehat{f}_t\widehat{\boldsymbol{\lambda}})\widehat{f}_t = \mathbf{0} \quad \text{where } \boldsymbol{e}_t = (e_{1t},\ldots,e_{Nt})', \tag{2.13}$$

subject to $T^{-1}\sum_{t=1}^{T}\widehat{f}_t^2 = T^{-1}\widehat{\boldsymbol{f}}'\widehat{\boldsymbol{f}} = 1$. Since $\widehat{\lambda}_i = T^{-1}\widehat{\boldsymbol{f}}'\boldsymbol{e}_i$, we obtain

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{e}_i - \frac{1}{T}\widehat{\boldsymbol{f}}\widehat{\boldsymbol{f}}'\boldsymbol{e}_i\right)\boldsymbol{e}_i'\widehat{\boldsymbol{f}} \;=\; \boldsymbol{M}_{\widehat{\boldsymbol{f}}}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{e}_i\boldsymbol{e}_i'\right)\widehat{\boldsymbol{f}} \;=\; \mathbf{0}, \tag{2.14}$$

where $\boldsymbol{M}_{\widehat{\boldsymbol{f}}} = \boldsymbol{I}_T - T^{-1}\widehat{\boldsymbol{f}}\widehat{\boldsymbol{f}}'$ with $\boldsymbol{M}_{\widehat{\boldsymbol{f}}}\widehat{\boldsymbol{f}} = \mathbf{0}$. For $N \to \infty$ we have

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{e}_i\boldsymbol{e}_i' \xrightarrow{p} \sigma_\lambda^2\boldsymbol{f}\boldsymbol{f}' + \boldsymbol{\Sigma}_u,$$

where $\sigma_\lambda^2 = \operatorname*{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}\lambda_i^2$, $\boldsymbol{\Sigma}_u = \operatorname*{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}\boldsymbol{u}_i\boldsymbol{u}_i'$, and $\boldsymbol{u}_i = (u_{i1},\ldots,u_{iT})'$. Assume that $u_{it}$ is i.i.d. with $\boldsymbol{\Sigma}_u = \sigma_u^2\boldsymbol{I}_T$. As $N \to \infty$ the moment condition is solved by letting $\widehat{\boldsymbol{f}} = \boldsymbol{f}$ and, therefore, the PC estimator for $\boldsymbol{f}$ is consistent (up to a scaling factor). If $u_{it}$ is heteroskedastic or autocorrelated, then $\boldsymbol{M}_{\boldsymbol{f}}\boldsymbol{\Sigma}_u\boldsymbol{f} \neq 0$ in general and, therefore, the PC estimator is inconsistent as $N \to \infty$. On the other hand, if both $N$ and $T$ tend to infinity, the PC estimator is consistent no matter of a possible heteroskedasticity or (weak) autocorrelation (cf. Chamberlain and Rothschild 1983).

The asymptotic theory for the HNR and ALS estimators assumes that $T$ is fixed and $N$ tends to infinity. The GMM estimator is based on $kT(T-1)$ moment conditions with $k + T - 1$ unknown parameters. Therefore, no problem arises if $T$ is fixed and $N$ tends to infinity. Accordingly, the estimators are asymptotically normally distributed and centered around zero. Of course the problem of instrument proliferation arises if $T$ gets large and the asymptotic theory breaks down if $T^3/N \to \kappa > 0$ (cf. Bekker 1994 and Lee et al. 2017).[5]

[5] A practical solution is to reduce the set of instruments (cf. Juodis and Sarafidis 2018) or applying other methods of dimensionality reduction (Breitung 2015, Section 15.2.3).

## 2.4. Identification

Since the factor space is not identified without some normalization of the factors and factor loadings, the estimation approaches impose some normalization that may be problematical in empirical practice. The CCE and ALS approaches require the following conditions:

$$\text{CCE:} \qquad \frac{1}{N}\sum_{i=1}^{N}\lambda_{0,i}\lambda_i \neq 0, \tag{2.15}$$

$$\text{ALS:} \qquad f_T \neq 0\ , \tag{2.16}$$

whereas the requirement for the PC estimator $T^{-1}\sum_{t=1}^{T} f_t^2 > 0$ is unproblematic, as otherwise the factor does not exist. The violation of the restrictions (2.15) and (2.16) may result in poor distributional properties of the estimator. If, for example, $N^{-1}\sum \lambda_{0,i}\lambda_i = 0$, then the cross section mean $\overline{e}_t(\boldsymbol{\lambda}_0)$ does not depend on the factor and, therefore, the CCE estimator is biased whenever $\boldsymbol{x}_{it}$ and $\lambda_i f_t$ are correlated (cf. Westerlund and Urbain 2013). Similarly, if $f_T = 0$, then $y_{iT} = \boldsymbol{\beta}'\boldsymbol{x}_{iT} + u_{iT}$ and the instruments are not able to identify the parameters $\theta_t$ and $\boldsymbol{\delta}_t$.

One may argue that the chance that (2.15) or (2.16) is exactly zero is negligible, so that problems only occur in rare cases (if at all). Unfortunately, this is not true, as the problems already arise whenever $N^{-1}\sum \lambda_{0,i}\lambda_i = O_p(N^{-1/2})$. For illustration, let us assume $\lambda_{0,i} = 1$, such that $\overline{y}_t(\boldsymbol{\lambda}_0) = \overline{y}_t$ and $\overline{\lambda} = O_p(N^{-1/2})$. Including the cross-section averages $\overline{y}_t$ and $\overline{\boldsymbol{x}}_t$ is equivalent to augmenting with $\overline{e}_t$ and $\overline{\boldsymbol{x}}_t$. Furthermore,

$$\begin{aligned}\overline{e}_t &= \overline{\lambda} f_t + \overline{u}_t \\ &= \overline{\lambda} f_t^*\ ,\end{aligned}$$

where $f_t^* = f_t + (\overline{u}_t/\overline{\lambda})$. Since in our case $\overline{u}_t/\overline{\lambda} = O_p(1)$, it follows that the factor $f_t^*$ is different from $f_t$. In this case, $\overline{e}_t$ does not represent the true factor and the CCE estimator of $\boldsymbol{\beta}$ is inconsistent whenever the factor is correlated with the regressors.

To sidestep this difficulty, we follow the analysis of Mundlak (1978) and decompose the factor loadings into a systematic component related to the ordinary average $\overline{\boldsymbol{x}}_i$ and the projection error $\xi_i$:

$$\lambda_i = \gamma_0 + \boldsymbol{\gamma}_1'\overline{\boldsymbol{x}}_i + \xi_i, \tag{2.17}$$

where $\overline{\boldsymbol{x}}_i = T^{-1}\sum_{t=1}^{T}\boldsymbol{x}_{it}$ and $\xi_i$ is uncorrelated with $\overline{\boldsymbol{x}}_i$. In this specification $\boldsymbol{\gamma}_1'\overline{\boldsymbol{x}}_i$

represents a possible linear dependence of $\lambda_i$ on the regressors that gives rise to an endogeneity bias. Inserting (2.17) in (2.1) yields

$$y_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \lambda_i^* f_t + e_{it}^* \ ,$$

where $\lambda_i^* = \gamma_0 + \boldsymbol{\gamma}_1'\overline{\boldsymbol{x}}_i$, $e_{it}^* = \xi_i f_t + u_{it}$ and $\mathbb{E}(e_{it}^*|\boldsymbol{x}_{it}) = 0$. This estimation equation is related to the projection approach of Hayakawa (2012), who considers a projection of $\lambda_i$ on the vector $\boldsymbol{z}_i = vec(\boldsymbol{X}_i)$, also known as Chamberlain projection. A second difference to the Hayakawa (2012) approach is that he employs the projection for GMM estimation of ALS, whereas we employ the Mundlak projection in the context of CCE estimation.

The weighting scheme for the CCE estimator results as

$$\begin{aligned}
\overline{y}_t(\boldsymbol{\lambda}^*) &= \frac{1}{N\overline{\lambda_*^2}} \sum_{i=1}^{N} \lambda_i^* y_{it} \\
&= \widetilde{\gamma}_0 \left( \frac{1}{N} \sum_{i=1}^{N} y_{it} \right) + \widetilde{\boldsymbol{\gamma}}_1' \left( \frac{1}{N} \sum_{i=1}^{N} \overline{\boldsymbol{x}}_i y_{it} \right) \\
\text{where } \widetilde{\gamma}_0 &= \gamma_0/\overline{\lambda_*^2} \ \text{ and } \ \widetilde{\boldsymbol{\gamma}}_1 = \boldsymbol{\gamma}_1/\overline{\lambda_*^2}
\end{aligned}$$

and $\overline{\lambda_*^2} = \frac{1}{N}\sum_{i=1}^{N}(\lambda_i^*)^2$. Since $\widetilde{\gamma}_0$ and $\widetilde{\boldsymbol{\gamma}}_1$ are unknown, we augment the regression by the following $(k+1)^2$ cross section averages:

$$\begin{array}{ccccc}
\frac{1}{N}\sum_{i=1}^{N} y_{it} \ , & \frac{1}{N}\sum_{i=1}^{N} x_{1,it} \ , & \cdots \ , & \frac{1}{N}\sum_{i=1}^{N} x_{k,it} \ , \\
\frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} y_{it} \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{1,it} \ , & \cdots \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{k,it} \ , \\
\vdots & \vdots & & \vdots \\
\frac{1}{N}\sum_{i=1}^{N} \overline{x}_{k,i} y_{it} \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{k,i} x_{1,it} \ , & \cdots \ , & \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{k,i} x_{k,it} \ .
\end{array}$$

This estimator is referred to as CCE($M$).[6] It is important to note that this approach implies the inclusion of $(k+1)^2$ cross-section averages, attached with individual specific coefficients. It follows that $T$ needs to be larger than $(k+1)^2$ which may be a severe restriction in empirical practice. Furthermore, the small sample properties of the CCE($M$) estimator may suffer from a large number of auxiliary regressors.

[6] This estimator can be seen as a special case of the combination-CCE estimator proposed by Karabiyik et al. (2019b).

Similar normalization problems arise for the HNR and ALS approaches, but these estimators apply a normalization to the *factors*. For example, if $f_T$ is zero, then the linear combination of $y_{iT}$ and $\boldsymbol{x}_{iT}$ is not able to identify the factor and, therefore, the ALS approach is biased whenever $f_T = 0$ and $\boldsymbol{x}_{it}$ is correlated with $\lambda_i f_t$. If $T$ is small, then one may try out all possible time periods for normalization and select the normalization that minimizes the GMM objective function. For a large number of time series this approach is rather time consuming. In such cases the normalization may be selected by estimating the factor by the PC approach. Then, the normalization period with the largest factor (in absolute value) is selected as the normalization period.

In the appendix of Ahn et al. (2013) a more flexible approach is proposed, which we refer to as ALS*. Let $\boldsymbol{H}$ denote the $T \times (T-1)$ orthogonal complement of $\boldsymbol{f} = (f_1, \ldots, f_T)'$ such that $\boldsymbol{H}'\boldsymbol{f} = \boldsymbol{0}$. To obtain (2.10) we let

$$\boldsymbol{H}'_{\text{ALS}} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -\theta_1 \\ 0 & 1 & 0 & \cdots & 0 & -\theta_2 \\ \vdots & & \ddots & & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & -\theta_{T-1} \end{pmatrix}.$$

To avoid normalizing $T-1$ elements to unity, we transform the equations for unit $i$ by using a more general matrix with property $\boldsymbol{H}'\boldsymbol{f} = \boldsymbol{0}$, such that $\boldsymbol{H}'\boldsymbol{e}_i = \boldsymbol{H}'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$, $\widetilde{\boldsymbol{e}}_i = \boldsymbol{H}'\boldsymbol{e}_i$. Given $\boldsymbol{\beta}$, the estimator of $\boldsymbol{H}$ is based on the moment condition $\mathbb{E}(\boldsymbol{H}'\boldsymbol{e}_i\boldsymbol{z}_i') = \boldsymbol{0}$, where $\boldsymbol{z}_i = vec(\boldsymbol{X}_i)$. Accordingly, a GMM estimator for $\boldsymbol{H}$ can be obtained as

$$\widehat{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmin}} \left\{ tr\left(\boldsymbol{H}'\boldsymbol{\Omega}_{ez}\boldsymbol{\Omega}_{zz}^{-1}\boldsymbol{\Omega}_{ez}'\boldsymbol{H}\right)\right\} \quad \text{s.t.} \quad \boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_{T-1},$$

where $\boldsymbol{\Omega}_{ez} = N^{-1}\sum_{i=1}^N \boldsymbol{e}_i\boldsymbol{z}_i'$ and $\boldsymbol{\Omega}_{zz} = N^{-1}\sum_{i=1}^N \boldsymbol{z}_i\boldsymbol{z}_i'$. Accordingly, the estimator $\widehat{\boldsymbol{H}}$ is obtained as the matrix of eigenvectors corresponding to the smallest $T-1$ eigenvalues of the matrix $\boldsymbol{\Omega}_{ez}\boldsymbol{\Omega}_{zz}^{-1}\boldsymbol{\Omega}_{ez}'$. Given $\widehat{\boldsymbol{H}}$, the estimator for $\boldsymbol{\beta}$ is obtained from the OLS regression

$$\widehat{\boldsymbol{H}}'\boldsymbol{y}_i = \widehat{\boldsymbol{H}}'\boldsymbol{X}_i\boldsymbol{\beta} + \widetilde{\boldsymbol{e}}_i.$$

This estimation step yields an updated estimator for $\boldsymbol{\beta}$ that can be used to obtain a new estimator of $\boldsymbol{H}$, until convergence. A drawback of this variant of the ALS estimator is that no standard errors for $\boldsymbol{\beta}$ are readily available, as the respective estimation step is affected by the estimation error in $\widehat{\boldsymbol{H}}$.

It is interesting to compare this approach to the PC estimator of Bai (2009), which can be obtained by solving the problem

$$\widetilde{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmin}} \left\{ tr\left(\boldsymbol{H}'\boldsymbol{\Omega}_{ee}\boldsymbol{H}\right)\right\} \quad \text{s.t.} \quad \boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_{T-1},$$

where $\boldsymbol{\Omega}_{ee} = N^{-1}\sum_{i=1}^{N} \boldsymbol{e}_i\boldsymbol{e}_i'$. Accordingly, the difference between the PC and ALS/RS approaches is that the former extracts the factors from the residual vector $\boldsymbol{e}_i$, whereas the ALS/RS approach first projects the residuals on the space spanned by the vector of instruments $\boldsymbol{z}_i$. Accordingly, the latter approach requires that the factors are correlated with the regressors, whereas the PC approach does not.

Robertson and Sarafidis (2015) show that their estimator considered in Section 2.2.4 is asymptotically equivalent to ALS$^*$ if the error $u_{it}$ is i.i.d. If $u_{it}$ is heteroskedastic and/or serially correlated, then the weighting matrix $\boldsymbol{W}_n$ results in an asymptotic efficiency gain.

## 2.5. Multiple factors

So far we assumed that there is only a single factor. It is not difficult to see that for a panel data model with a vector of $r \geq 1$ factors $\boldsymbol{f}_t$ and the conformable $r \times 1$ loading vector $\boldsymbol{\lambda}_i$, the estimation equation (2.3) is given by

$$y_{it} - \boldsymbol{\lambda}_i'\overline{\boldsymbol{y}}_t^*(\boldsymbol{\Lambda}) = \left[\boldsymbol{x}_{it}' - \boldsymbol{\lambda}_i'\overline{\boldsymbol{X}}_t^*(\boldsymbol{\Lambda})\right]\boldsymbol{\beta} + u_{it} - \boldsymbol{\lambda}_i'\overline{\boldsymbol{u}}_t(\boldsymbol{\Lambda}), \tag{2.18}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)'$ and

$$\overline{\boldsymbol{y}}_t^*(\boldsymbol{\Lambda}) = \left(\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'\right)^{-1}\sum_{i=1}^{N}\boldsymbol{\lambda}_i y_{it}$$

$$\text{and} \quad \overline{\boldsymbol{X}}_t^*(\boldsymbol{\Lambda}) = \left(\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'\right)^{-1}\sum_{i=1}^{N}\boldsymbol{\lambda}_i\boldsymbol{x}_{it}'$$

and the $r \times 1$ vector $\overline{\boldsymbol{u}}_t(\boldsymbol{\Lambda})$ is constructed in a similar manner. This shows that efficient estimation requires $r$ linear independent weighting schemes applied to $\boldsymbol{y}_t = (x_{1t}, \ldots, y_{Nt})'$ and $\boldsymbol{X}_t = (\boldsymbol{x}_{1t}', \ldots, \boldsymbol{x}_{Nt}')'$.

To show consistency of the modified CCE estimator, CCE($M$), a different reasoning is required. For the ease of exposition assume $k = 2$ regressors and $r = 2$ factors. We

obtain 2 different weighting schemes:

$$\overline{y}_t^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} y_{it} \qquad \overline{x}_{1,t}^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{1,it} \qquad \overline{x}_{2,t}^{(1)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{1,i} x_{2,it}$$
$$\overline{y}_t^{(2)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{2,i} y_{it} \qquad \overline{x}_{1,t}^{(2)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{2,i} x_{1,it} \qquad \overline{x}_{2,t}^{(2)} = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{2,i} x_{2,it}$$

that are used to obtain the following relationships:

$$\begin{pmatrix} \overline{y}_t^{(1)} \\ \overline{y}_t^{(2)} \end{pmatrix} - \begin{pmatrix} \overline{x}_{1,t}^{(1)} & \overline{x}_{2,t}^{(1)} \\ \overline{x}_{1,t}^{(2)} & \overline{x}_{2,t}^{(2)} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \xi_1^{(1)} & \xi_2^{(1)} \\ \xi_1^{(2)} & \xi_2^{(2)} \end{pmatrix} \begin{pmatrix} f_{1,t} \\ f_{2,t} \end{pmatrix} + O_p(N^{-1/2})$$

where $\xi_k^{(\ell)} = N^{-1}\sum_{i=1}^{N} \overline{x}_{\ell,i}\lambda_{k,i}$. Accordingly, if the matrix

$$\boldsymbol{\Xi} = \begin{pmatrix} \xi_1^{(1)} & \xi_2^{(1)} \\ \xi_1^{(2)} & \xi_2^{(2)} \end{pmatrix}$$

is invertible,[7] we can obtain the linear combinations that represent the factors as

$$\begin{pmatrix} f_{1,t} \\ f_{2,t} \end{pmatrix} = \boldsymbol{\Xi}^{-1} \begin{pmatrix} \overline{y}_t^{(1)} \\ \overline{y}_t^{(2)} \end{pmatrix} - \boldsymbol{\Xi}^{-1} \begin{pmatrix} \overline{x}_{1,t}^{(1)} & \overline{x}_{2,t}^{(1)} \\ \overline{x}_{1,t}^{(2)} & \overline{x}_{2,t}^{(2)} \end{pmatrix} \boldsymbol{\beta} + O_p(N^{-1/2}).$$

Thus, asymptotically the space spanned by $(f_{1,t}, f_{2,t})$ is contained in the space spanned by the corresponding 6 cross-sectional averages $\overline{y}_t^{(1)}$, $\overline{y}_t^{(2)}$, $\overline{x}_{1,t}^{(1)}$, $\overline{x}_{2,t}^{(1)}$, $\overline{x}_{1,t}^{(2)}$, and, $\overline{x}_{2,t}^{(2)}$.[8]

## 2.6. Determining the number of factors

As argued by Pesaran (2006), the CCE estimator is consistent if the actual number of factors $r$ is not larger than $k+1$. This requires however that $r-1$ factors are correlated with the $k$ regressors. This is due to the fact that one factor can be identified by the cross-section average $\overline{e}_t(\boldsymbol{\lambda}_0) = \overline{y}_t(\boldsymbol{\lambda}_0) - \boldsymbol{\beta}'\overline{\boldsymbol{x}}_t(\boldsymbol{\lambda}_0)$, whereas the identification of the other factors requires some relationship to the cross-section averages of the regressors $\overline{\boldsymbol{x}}_t$.

[7] Note that for finite $N$ the matrix $\boldsymbol{\Xi}$ is almost surely invertible, even if $\lambda_i$ and $\boldsymbol{x}_{it}$ are uncorrelated for all $i$ and $t$. To establish consistency, we require that the probability limit of $\boldsymbol{\Xi}$ is invertible as $N \to \infty$.

[8] The alert reader may have noticed that the linear combination does not involve the ordinary cross-section averages $N^{-1}\sum_i y_{it}$, $N^{-1}\sum_i x_{1,it}$ and $N^{-1}\sum_i x_{2,it}$ that are employed in the CCE estimator. These additional averages are not required for identification but often improve the statistical properties of the estimator. They may also help to escape the problems resulting from a (nearly) singular matrix $\boldsymbol{\Xi}$.

Furthermore, the correlation pattern needs to be sufficiently informative for identifying the factors.

It is often argued that the CCE approach is attractive, as we do not need to select the number of factors, whereas for all other approaches, the number of factors needs to be known (or determined from the data). If the number of factors is smaller than $k+1$ and the normalization requirements are satisfied, then the CCE estimator is consistent, but the small sample properties may suffer from including many cross-section averages. This is comparable to applying the PC estimator with $r = k+1$ factors. As shown by Moon and Weidner (2015), under some additional assumptions,[9] the PC estimator is robust against over-specifying the number of factors. A similar result is obtained by Westerlund et al. (2019) for the CCE estimator. Since under certain conditions the CCE estimator for $\boldsymbol{\beta}$ is as efficient as the OLS estimator using the true factors, there is no gain in (asymptotic) efficiency by changing the weighting scheme or imposing nonlinear restrictions to the auxiliary parameters that are implied by knowing the number of factors. It is, however, not clear whether this result provides a good guidance for empirical applications in finite samples.

In practice, it may therefore be interesting to estimate the number of factors. To this end, we may invoke the criteria proposed by Bai and Ng (2002) and Ahn and Horenstein (2013). Both approaches are based on the eigenvalues of the residual covariance matrix. Denote by $\widehat{\mu}_1 \geq \cdots \geq \widehat{\mu}_T$ the ordered eigenvalues of the $T \times T$ sample covariance matrix $\widehat{\boldsymbol{\Omega}}_{ee} = N^{-1}\sum_{i=1}^{N} \widehat{\boldsymbol{e}}_i \widehat{\boldsymbol{e}}_i'$, where the residual vector $\widehat{\boldsymbol{e}}_i$ is obtained by estimating the model with maximum number of factors $r^*$. Furthermore, let

$$\widehat{\sigma}_u^2(r) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{u}_{it}^2 \;=\; \frac{1}{T}\sum_{j=r+1}^{T}\widehat{\mu}_j$$

where $\widehat{u}_{it}$ denotes the residual from estimating the model with $r$ factors. Bai and Ng's (2002) criterion $IC_{p2}$ minimizes

$$\mathrm{BN}(r) = \log\big(\widehat{\sigma}_u^2(r)\big) + r\,\frac{N+T}{NT}\log(min[N,T]),$$

for $r \in \{0,1,\ldots,r^*\}$, whereas the criterion proposed by Ahn and Horenstein (2013) maximizes the eigenvalue ratios

$$\mathrm{AH}(r) = \widehat{\mu}_j/\widehat{\mu}_{j+1} \quad \text{for } r \in \{1,2,\ldots,r^*\}$$

[9] The proof of Moon and Weidner (2015) requires $T \to \infty$ and is based on the i.i.d. assumption but they note that it appears that their results extend to a less restrictive setting.

and the mock eigenvalue $\widehat{\mu}_0 = \left(\sum_{j=1}^{T} \widehat{\mu}_j\right) / \log(T)$. Let $r_0$ denote the true number of factors. If $\widehat{\boldsymbol{\beta}}_* - \boldsymbol{\beta} = O_p(1/\sqrt{NT})$, we have

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \widehat{\boldsymbol{\beta}}_*'\boldsymbol{x}_{it})^2 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 - 2\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\boldsymbol{x}_{it}'(\widehat{\boldsymbol{\beta}}_* - \boldsymbol{\beta}) + O_p\left(\frac{1}{NT}\right) \\ &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 + O_p\left(\frac{1}{\sqrt{NT}}\right).\end{aligned}$$

Accordingly, the BN and AH criteria include an additional term of order $O_p((NT)^{-1/2})$ that does not affect the asymptotic properties as $N$ and $T$ tend to infinity.

Let us consider the asymptotic properties of the respective estimators $\widehat{r}$ if $T$ is fixed and $N \to \infty$. The condition $\lim_{N\to\infty} P(\widehat{r} < r_0) = 0$ implies (cf. Bai and Ng 2002)

$$c(N,T) = \frac{N+T}{NT}\log(min[N,T]) \to 0. \tag{2.19}$$

As condition (2.19) is not satisfied for fixed $T$, the BN criterion may select some $\widehat{r} < r_0$, even if $N \to \infty$. The requirement $\lim_{N\to\infty} P(\widehat{r} > r_0) = 0$ implies

$$\lim_{N\to\infty} P\left((r - r_0)c(N,T) + \log\big(\widehat{\sigma}_u^2(r)\big) - \log\big(\widehat{\sigma}_u^2(r_0)\big) > 0\right) = 1 \quad \text{for all } r > r_0. \tag{2.20}$$

Since $\log\big(\widehat{\sigma}_u^2(r_0)\big) - \log\big(\widehat{\sigma}_u^2(r)\big) = O_p(N^{-1}) + O_p(T^{-1})$ for $r > r_0$ (cf. Lemma 4 of Bai and Ng 2002), it may happen that for small $T$, condition (2.20) is violated as well. Hence, the BN criterion may not be consistent for fixed $T$. In practice, it is nevertheless possible that the BN criterion selects the number of factors consistently, if the eigenvalues $\widehat{\mu}_1, \ldots, \widehat{\mu}_{r_0-1}$ are sufficiently large and $\widehat{\mu}_{r_0+1}, \ldots, \widehat{\mu}_{r^*}$ are sufficiently small relative to $\widehat{\mu}_{r_0}$.

Since for fixed $T$, $\widehat{\mu}_r$ is $O_p(1)$ for all $r = 1, \ldots, T$, it follows that the eigenvalue ratio AH($r$) is $O_p(1)$ for fixed $T$ and all $r \in \{1, \ldots, r^*\}$. Therefore, the AH criterion cannot be shown to be a consistent selection rule for fixed $T$. It may nevertheless perform well, if the slope of the eigenvalue function is sufficiently steep at $r = r_0$.

A possibility to sidestep these problems is to adopt the BIC selection criteria of Ahn et al. (2013) and Robertson and Sarafidis (2015). These criteria are based on the Sargan-Hansen specification test for GMM estimators. If the number of factors is too small, then the remaining cross-correlation among the residuals results in a large value of the test statistic. The penalty function is constructed such that the sum of the test statistic and the penalty function obtains a minimum at the correct number of factors as $N$ tends to infinity.

## 2.7. Monte Carlo Simulations

In this section we assess the performance of alternative estimation methods in various settings and highlight some favorable and problematic aspects of alternative estimation methods. The simulation results in Sections 2.7.1 – 2.7.2 are based on the following simple data-generating process

$$y_{it} = \beta x_{it} + \lambda_i f_t + u_{it} \tag{2.21}$$

$$x_{it} = \mu + \lambda_i f_t + \lambda_i + f_t + \varepsilon_{it} \tag{2.22}$$

with $\beta = 0.5$ and $r = 1$. Hence, the regressor is correlated with the loadings, the factor and the product of both. The regression error $u_{it}$ and the idiosyncratic component of the regressor, $\varepsilon_{it}$, are independent standard normal random variables. The constant $\mu$ is drawn from a $U[0, 1]$ distribution. The DGPs in Sections 2.7.1 to 2.7.2 differ with respect to the distributional assumptions on the factors and their loadings.

The (near) violation of the normalization restrictions for the CCE and ALS estimators are examined in Section 2.7.1. In Section 2.7.2, we compare the PC and CCE estimator with regard to their different weighting schemes. In Section 2.7.3 we address the estimation of the number of factors, $r$, for the PC, ALS* and RS approaches. There, we consider a similar DGP as in (2.21) and (2.22) for $r = 1$ and $r = 2$. The last subsection 2.7.4 considers the relative performance of the CCE, PC, ALS* and RS estimation approaches in more general settings that are based on the DGPs considered by Bai (2009), Chudik et al. (2011) and Ahn et al. (2013).

### 2.7.1. Normalization failure

As argued in Section 2.4, the CCE and ALS/HRN approaches may suffer from a violation of their normalization conditions. The performance already deteriorates if the parameters approach the $\sqrt{N}$-vicinity of the problematic subspace. In a model with a single factor, the normalization of the equally weighted CCE estimator ($\lambda_{0,i} = 1$) requires that $\overline{\lambda} = N^{-1} \sum_{i=1}^{N} \lambda_i \neq 0$. We have argued that whenever $\overline{\lambda} = c/\sqrt{N}$, the factor cannot be represented by a linear combination of $\overline{y}_i$ and $\overline{x}_i$ as $N \to \infty$.

Sarafidis and Wansbeek (2012) and Westerlund and Urbain (2013) analyze the performance of the CCE estimator when the normalization condition is violated. In order to study the performance of the CCE estimator when $\overline{\lambda}$ is different but close to zero, we

consider the model in (2.21) and (2.22), where we generate the factor loadings as

$$\text{DGP1:} \quad \lambda_i \sim \mathcal{N}(\mu_\lambda, 1) \text{ for } \mu_\lambda \in [0,1] \quad \text{and} \quad f_t \sim \mathcal{N}(0,1).$$

Hence, the loadings are normally distributed with expectation that ranges from 0 to 1.

Figures 2.1 (a) – (d) present the absolute bias for the original CCE, the Mundlak type CCE(M) estimator suggested in Section 2.4, and the PC estimator for $N = 100$ and $N = 500$ with a small ($T = 10$) and moderate ($T = 50$) number of time periods. The PC estimator of Bai (2009) is obtained by a sequential estimation procedure using the pooled OLS estimator as starting value for $\beta$ (see Section 2.2.1). It turns out that the CCE estimator is severely biased even if the mean of $\lambda_i$ is substantially different from zero. This is due to the fact that a bias already occurs whenever $\mu_\lambda = O(N^{-1/2})$. This reasoning predicts that for fixed $\mu_\lambda$ the bias gets smaller if $N$ increases. Indeed, this is what we observe when comparing panel (a) and (c) as well as (b) and (d). Note that $\sqrt{100}/\sqrt{500} \approx 0.44$ and, therefore, we expect that the bias reduces to a value less than one half which is a good approximation for $\mu_\lambda > 0.1$. The other two estimators, PC and CCE(M), are virtually unbiased, which is expected as the estimators do not rely on the assumption $\mu_\lambda \neq 0$.

In a similar manner, the normalization of the ALS estimator may be problematic if the factors approach the problematic subspace. The ALS estimator requires $f_T \neq 0$. To examine the consequences of an (approximate) violation of this normalization condition, we consider the model in (2.21) and (2.22) where the factors are generated as:

$$\text{DGP2: } f_t \sim \mathcal{N}(0,1) \text{ for } t = 1, ..., T-1 \text{ and } f_T \sim \mathcal{N}(\mu_T, 0.5) \text{ for } \mu_T \in [0,1]$$

and the factor loadings are standard normally distributed. As the final value of the factor is crucial, we generate it by a distribution with expectation ranging from 0 to 1.

Figures 2.1 (e) – (f) present the bias for the ALS estimator when $T = 5$ and $N = 100$ or $N = 500$, respectively. As expected, the ALS estimator is severely biased whenever $\mu_T = \mathbb{E}(f_T)$ is small. But even for moderate values of $\mu_T$ the bias remains substantial and decreases only gradually for larger values of $\mu_T$. It should be noted that if the regression includes an individual specific intercept, then the factors are demeaned and, therefore, assuming a nonzero mean appears inappropriate.

Figures 2.1 (e) – (f) also present the bias of two estimators that circumvent the problems with the normalization of the original ALS estimator. The estimator ALS$^*$ refers to the GMM estimator that estimates the matrix $\boldsymbol{H}$ that is used to remove the

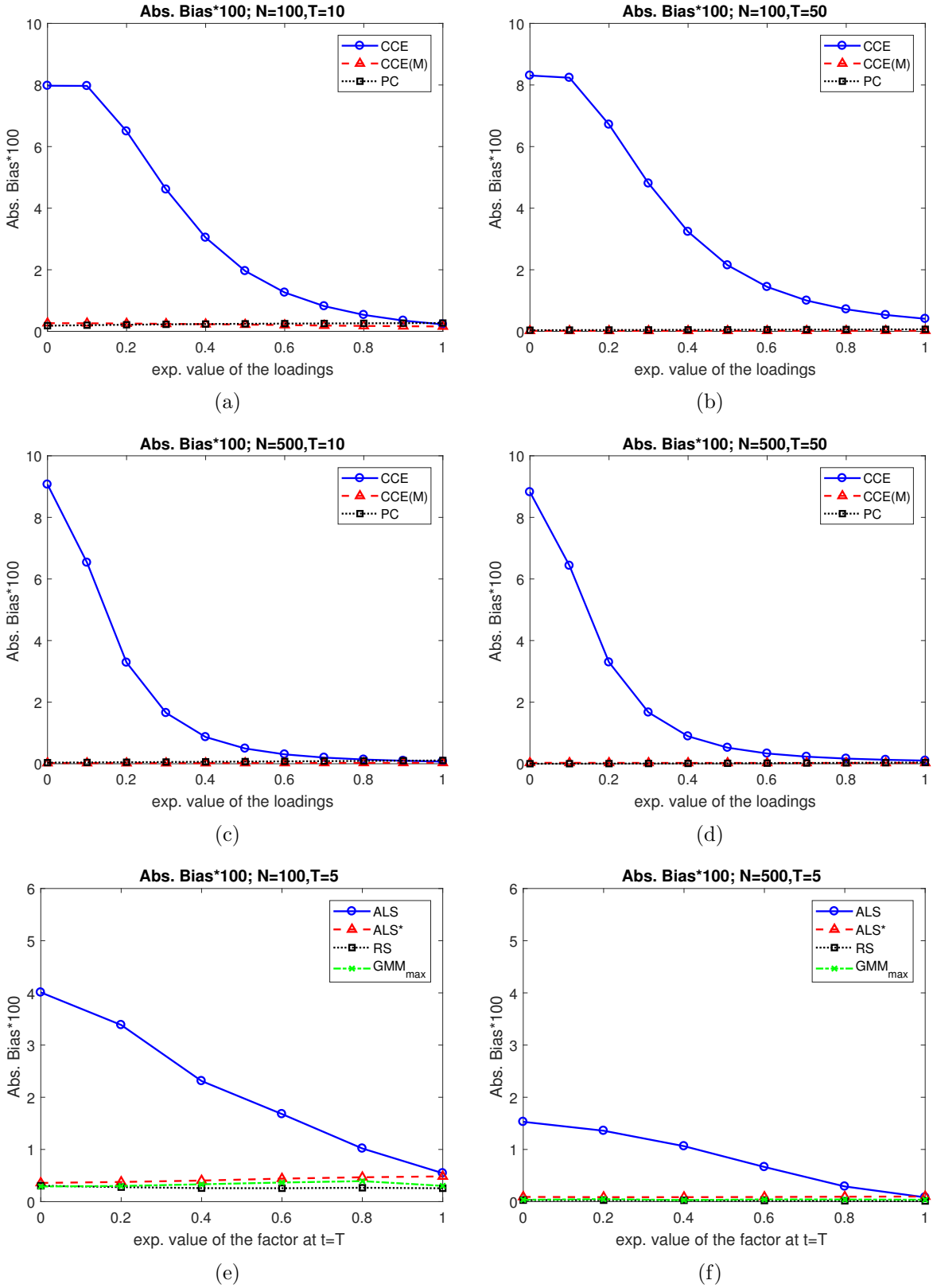



Figure 2.1.: Normalization failure for CCE (DGP1) and ALS (DGP2)

factors (see Section 2.4).[10] Our simulation results suggest that this estimator performs quite well in terms of bias, as it is virtually unbiased for all values of $\mu_T$. Another approach to escape the normalization problem is the GMM$_{max}$ estimator, where in a first step the factor is estimated using the PC approach. In the second step, the time period for the normalization is chosen according to the maximum absolute value of the estimated factor and the original ALS estimator is adapted, where the time period with the largest factor is shifted to the end of the sample. Both estimators are able to reduce the bias dramatically.

The figures also include the RS estimator, which corresponds to the FIVU estimator of Robertson and Sarafidis (2015). This estimator does not require $f_T \neq 0$ for normalization (see Section 2.2.4) and thus the bias does not depend on the value of $\mu_T$. The RS estimator has a slight advantage in terms of bias when $N = 100$. With $N = 500$, the bias of the ALS$^*$, GMM$_{max}$ and RS estimators is nearly zero.

To summarize, our findings confirm earlier evidence that the normalization applied for the original CCE or ALS/HNR estimators may be problematical, whenever the factors or loadings approach a normalization failure. It is, however, easy to adjust the estimators such that they perform well for all values of the parameter space. Our Monte Carlo exercise indicates that the PC and CCE(M) estimators as well as ALS$^*$, GMM$_{max}$ and RS are very robust against a possible normalization failure.

### 2.7.2. Fixed versus data driven weights

From the reasoning of Section 2.2, it turns out that the CCE estimator is expected to outperform the PC estimator whenever the weighting scheme $\boldsymbol{\lambda}_0$ comes close to the actual set of loadings $\boldsymbol{\lambda}$, see also Westerlund and Urbain (2015). For equal weights with $\lambda_{0,i} = 1$ for all $i$, the CCE estimator performs well, whenever (i) the absolute value of the mean of the loadings is large (to avoid the normalization failure) and (ii) the variance of the loadings is small. Our DGP3 represents such a scenario, whereas the DGP4 favors the PC estimator by generating factor loadings with large variance,

$$\text{DGP3: } \lambda_i \sim \mathcal{N}(1, 0.1), \;\; f_t \sim \mathcal{N}(0, 1)$$
$$\text{DGP4: } \lambda_i \sim \mathcal{N}(1, 3), \;\; f_t \sim \mathcal{N}(0, 1).$$

The remaining details of the simulation setup are identical to the model in (2.21) and (2.22).

[10] Following Ahn et al. (2013), we use $\beta = 0$ as starting value for the iterative ALS$^*$ procedure.

Table 2.1.: Fixed versus data driven weights

| | | Bias*100 | | | RMSE*100 | | |
|---|---|---|---|---|---|---|---|
| N | T | PC | CCE | CCE($M$) | PC | CCE | CCE($M$) |
| | | | | DGP3 | | | |
| 50 | 10 | 1.23 | 0.00 | 0.19 | 6.43 | 5.12 | 5.93 |
| 100 | 10 | 0.56 | 0.06 | 0.21 | 3.94 | 3.56 | 4.04 |
| 100 | 20 | 0.10 | −0.14 | −0.09 | 2.43 | 2.33 | 2.42 |
| 100 | 50 | 0.09 | −0.04 | 0.02 | 1.49 | 1.48 | 1.51 |
| 100 | 100 | 0.08 | −0.03 | 0.02 | 1.06 | 1.06 | 1.08 |
| 500 | 500 | 0.05 | −0.01 | −0.01 | 0.20 | 0.20 | 0.20 |
| | | | | DGP4 | | | |
| 50 | 10 | 0.18 | −2.31 | 0.19 | 4.65 | 6.62 | 5.97 |
| 100 | 10 | 0.24 | −1.09 | 0.22 | 3.26 | 4.05 | 4.17 |
| 100 | 20 | 0.01 | −1.30 | −0.08 | 2.15 | 3.05 | 2.45 |
| 100 | 50 | 0.08 | −1.22 | 0.01 | 1.34 | 2.36 | 1.51 |
| 100 | 100 | 0.10 | −1.20 | 0.01 | 0.97 | 2.00 | 1.08 |
| 500 | 500 | 0.08 | −0.24 | −0.01 | 0.20 | 0.36 | 0.20 |

This table reports the simulation results generated with DGPs 3 and 4. The results are based on 1000 replications.

The results reported in Table 2.1 clearly confirm our assertion that the CCE estimator outperforms the PC estimator in DGP3, whereas the PC estimator performs better for DGP4. This finding suggests to find a weighting scheme that comes close to the actual distribution of the loadings. This is the notion behind the Mundlak type CCE variant that employs the individual specific means $\overline{y}_i$ and $\overline{x}_i$, since a linear combination of these averages can be seen as (CCE type) estimates of the loadings $\lambda_i$. Therefore, we hope to improve the original CCE estimator by applying weights that are correlated with the loadings. Our results from the simple Monte Carlo experiment suggest that the CCE(M) approach of choosing a data driven weighting scheme performs similar to the best estimator in the respective situation. Furthermore, as shown in the previous subsection, the CCE(M) estimator sidesteps the risk of a normalization failure. Provided that this estimator is similarly easy to compute as the original CCE estimator, it appears as if this estimator is a robust variant of the original CCE estimator.

### 2.7.3. Selecting the number of factors

In practice, it is necessary to select the number of factors for the PC and GMM estimation procedures. The choice is important, since misspecifying the number of factors can have severe consequences: Overspecifying the number of factors can have adverse effects on the sampling properties of the estimators, while an underspecification may lead to inconsistent estimates if the ignored factors are correlated with the regressors. One possibility for selecting the number of factors is simply to specify the number according to some ad hoc rule, for instance $r = k + 1$, as usually advocated for the CCE approach. Another option is to use a consistent criterion for the number of factors, such as the ones proposed by Bai and Ng (2002) (hereafter: BN) and Ahn and Horenstein (2013) (AH). Note that these selection criteria were developed for the pure factor model without regressors. Furthermore, the asymptotic theory underlying these approaches requires $T \to \infty$ (see Section 2.6). It is therefore interesting to investigate the performance of these criteria that were not initially developed for a small number of time periods. For the GMM estimators, the number of factors can be estimated using model information criteria, such as the Schwarz Criterion (BIC) considered by Ahn et al. (2013) and Robertson and Sarafidis (2015).

Table 2.2.: Hit rates for selection criteria

| | | r=1 | | | | r=2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| N | T | $BN_{PC}$ | $AH_{PC}$ | $BIC_{ALS*}$ | $BIC_{RS}$ | $BN_{PC}$ | $AH_{PC}$ | $BIC_{ALS*}$ | $BIC_{RS}$ |
| 100 | 5 | 0.0 | 94.6 | 91.7 | 83.0 | 0.0 | 46.8 | 86.4 | 76.6 |
| 250 | 5 | 0.0 | 96.2 | 96.9 | 96.7 | 0.0 | 50.8 | 93.2 | 89.5 |
| 500 | 5 | 0.0 | 96.9 | 98.8 | 98.3 | 0.0 | 52.1 | 96.3 | 94.1 |
| 250 | 10 | 100.0 | 99.9 | 90.6 | 97.0 | 99.6 | 86.4 | 89.7 | 92.9 |
| 500 | 10 | 99.9 | 99.9 | 96.7 | 98.4 | 99.4 | 89.8 | 95.9 | 96.9 |
| 500 | 15 | 100.0 | 100.0 | 92.3 | 99.6 | 100.0 | 97.9 | 92.9 | 98.8 |

In order to study the performance of these selection criteria, we consider a similar model as in (2.21) and (2.22) with $r = 1$ and $r = 2$. For the loadings and factors, we assume the following DGP,

$$\text{DGP5: } \lambda_{j,i} \sim \mathcal{N}(0,1),\ f_{j,t} \sim \mathcal{N}(0,1) \ \text{ for } \ j = 1, 2.$$

As reported in Table 2.2, the hit rates for a single factor, $r = 1$, are nearly 100% for the BN and AH criteria whenever $T \geq 10$. For $T = 5$ the BN criterium does not work and nearly always picks the maximum number of factors. On the other hand, the AH

criterion works remarkably well, even for a number of time periods as small as $T = 5$.[11] The hit rates for the BIC criteria exceed 90% in all but one case. For $r = 2$ the hit rates for the AH criterion are substantially lower, but the estimators are still quite accurate, even if $T = 10$ and $N$ is large. For the BIC criteria, the hit rates decrease by only a small amount and do not seem to be very sensitive to the number of factors, in particular if $N > 100$.

In Table 2.3, we report bias and RMSE for the PC, ALS$^*$ and RS estimators based on the true number of factors ($r = 1$ and $r = 2$) as a benchmark. In addition we assess the performance of the estimators, when the number of factors is estimated based on selection criteria.[12] As expected, using the AH method for $r = 1$ in order to estimate the number of factors for the PC estimator produces bias and RMSE results that are of similar magnitude as the true number of factors. Applying the BIC criterion to estimate the number of factors for the GMM estimators produces very accurate estimates when $N > 100$, accordingly.

For $r = 2$, the performance of the PC estimator using the AH criterion shows a considerable bias, in particular if $T$ is as small as 5. In contrast, bias and RMSE of the GMM estimators applying the BIC criterion are similar to the estimators based on the true number of factors when $N > 100$. When $T$ increases to 10, there is still a substantial performance gap between the PC estimator using the AH method and the PC estimator based on the true number of factors, whereas the GMM estimators based on the BIC criterion perform much better. This is surprising as Table 2.2 suggests that the hit rates of the BIC criterion are only slightly better in these cases. The reason is that the AH criterion tends to underestimate the number of factors, whereas the BIC criterion overestimates the number of factors in case the correct number of factors is not found.

Consider, for instance, $T = 10$ and $N = 500$. The BIC estimator finds the correct number of factors ($r = 2$) in more than 95% of the cases and overestimates the number in the other ($< 5\%$) cases. The AH estimator finds the correct value of $r = 2$ in 89.8% of the cases, however underestimates the number in all other cases. Since the estimator is biased if the number of factors is too small, the AH criterion tends to produce a large negative bias in some cases, whereas the BIC criterion tends to produce unbiased

[11] The performance is similar to the case where $\beta$ is known (not shown). Therefore, the estimation of $\beta$ does not seem to have an important effect on the performance of the BN and AH selection criteria. Furthermore, the growth ratio statistic of Ahn and Horenstein (2013) performs similar to the eigenvalue ratio statistic. For reasons of space we do not show the respective results.

[12] To save space, we do not show results for the estimators based on the BN criterion, since the hit rates are either 0% or (close to) 100%.

Table 2.3.: Selecting the number of factors

| | | $r=1$ | | | | $r=2$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Bias*100 | | RMSE*100 | | Bias*100 | | RMSE*100 | |
| N | T | $PC_r$ | $PC_{AH}$ | $PC_r$ | $PC_{AH}$ | $PC_r$ | $PC_{AH}$ | $PC_r$ | $PC_{AH}$ |
| 100 | 5 | 0.20 | 0.31 | 5.10 | 5.30 | 0.67 | 5.47 | 6.83 | 11.06 |
| 250 | 5 | 0.12 | 0.24 | 3.22 | 3.51 | 0.29 | 4.89 | 4.14 | 10.13 |
| 500 | 5 | 0.14 | 0.30 | 2.25 | 2.66 | 0.22 | 4.54 | 3.04 | 9.13 |
| 250 | 10 | 0.07 | 0.08 | 2.04 | 2.04 | 0.17 | 1.51 | 2.20 | 4.79 |
| 500 | 10 | 0.09 | 0.10 | 1.42 | 1.44 | 0.06 | 1.07 | 1.55 | 3.93 |
| 500 | 15 | 0.07 | 0.07 | 1.06 | 1.06 | 0.11 | 0.28 | 1.18 | 1.81 |
| | | $ALS^*_r$ | $ALS^*_{BIC}$ | $ALS^*_r$ | $ALS^*_{BIC}$ | $ALS^*_r$ | $ALS^*_{BIC}$ | $ALS^*_r$ | $ALS^*_{BIC}$ |
| 100 | 5 | 0.14 | 0.15 | 6.22 | 6.84 | −0.33 | −0.65 | 7.46 | 8.15 |
| 250 | 5 | −0.01 | 0.04 | 3.69 | 3.83 | 0.08 | 0.10 | 4.33 | 4.53 |
| 500 | 5 | 0.23 | 0.23 | 2.62 | 2.64 | 0.04 | −0.01 | 3.08 | 3.26 |
| 250 | 10 | −0.02 | −0.02 | 2.25 | 2.36 | 0.00 | −0.02 | 2.23 | 2.32 |
| 500 | 10 | 0.10 | 0.10 | 1.59 | 1.61 | −0.06 | −0.06 | 1.58 | 1.59 |
| 500 | 15 | 0.04 | 0.03 | 1.20 | 1.22 | 0.03 | 0.03 | 1.18 | 1.20 |
| | | $RS_r$ | $RS_{BIC}$ | $RS_r$ | $RS_{BIC}$ | $RS_r$ | $RS_{BIC}$ | $RS_r$ | $RS_{BIC}$ |
| 100 | 5 | −0.58 | 0.74 | 6.01 | 7.93 | −0.92 | −0.26 | 7.88 | 9.00 |
| 250 | 5 | −0.17 | −0.12 | 3.65 | 3.76 | −0.21 | −0.14 | 4.72 | 4.99 |
| 500 | 5 | 0.11 | 0.10 | 2.60 | 2.66 | −0.07 | −0.10 | 3.58 | 3.59 |
| 250 | 10 | −0.40 | −0.29 | 2.42 | 2.68 | −0.86 | −0.73 | 3.20 | 3.21 |
| 500 | 10 | −0.10 | −0.10 | 1.66 | 1.66 | −0.44 | −0.41 | 2.16 | 2.11 |
| 500 | 15 | −0.17 | −0.17 | 1.29 | 1.29 | −0.65 | −0.62 | 2.05 | 2.00 |

This table reports bias and RMSE results for DGP5 with $r = 1$ and $r = 2$ for the PC, $ALS^*$ and RS estimators with the true number of factors and estimated number of factors based on selection criteria. The results are based on 1000 replications.

estimators with a slightly larger variance than estimating with the correct number of factors in some very rare cases.

### 2.7.4. Performance in more general setups

So far the DGPs considered in this paper were simplified versions of the ones considered in the literature and focus on the particular features of these models. In the following, we study the relative performance of the CCE, PC, ALS* and RS approaches in more sophisticated simulation setups, similar to the simulation experiments of Bai (2009), Chudik et al. (2011) and Ahn et al. (2013). The details of these data generating processes are presented in Appendix A. The Monte Carlo design of Bai (2009) employs two regressors that are correlated with two factors, their loadings and the product of both. The idiosyncratic error is i.i.d. across individuals and time periods. We refer to this model as DGP6. DGP7 refers to the factor model of Chudik et al. (2011) that includes two regressors and three factors. A special feature of this DGP is that the factor loadings of the regressors are independent of the loadings in the errors $e_{it}$ and, therefore, the regressors are not correlated with the errors. The factors are generated by independent AR(1) processes and the idiosyncratic component $u_{it}$ is heteroskedastic but mutually and serially uncorrelated. DGP8 corresponds to the Monte Carlo design of Ahn et al. (2013), which includes two regressors and two factors. The first regressor is correlated with the first factor and the second regressor is correlated with the second factor. The idiosyncratic error is autocorrelated but the variances are identical across panel units and time periods.

The results in Table 2.4 indicate that the relative performance of the estimators depends quite sensitively on the DGP considered. The first panel of Table 2.4 presents the results for DGP6. The CCE estimator is not consistent in this setting, since the rank condition is violated and both factor and loading vectors are correlated with both regressors. The other three estimators are consistent in this setting, where the RS estimator is the least biased when $T = 5$ and the ALS* exhibits the lowest bias for $T \geq 10$. The latter performs best in terms of RMSE with only slight advantages over the PC estimator when $T \geq 10$.

The second panel of Table 2.4 reports the results for DGP7. The CCE estimator is the favored one in this setting. It has a very small bias and exhibits the lowest RMSE for nearly all considered $(N, T)$ combinations, in particular if $T$ is as small as 5. Comparing the PC and GMM estimators, the results slightly favor the PC estimator in terms of RMSE. The difference between the PC and the CCE estimator is negligible when $T = 15$ and $N = 500$. With regard to the GMM estimators, the RS estimator has a marginally

Table 2.4.: Performance in more general setups

| | | Bias*100 | | | | | | | | RMSE*100 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CCE | | PC | | ALS* | | RS | | CCE | | PC | | ALS* | | RS | |
| N | T | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ | $\beta_1$ | $\beta_2$ |
| DGP Bai (2009) | | | | | | | | | | | | | | | | | |
| 100 | 5 | 10.27 | 10.27 | 3.54 | 3.90 | −2.14 | −1.92 | −0.75 | −0.07 | 24.40 | 24.19 | 13.65 | 13.94 | 13.34 | 13.95 | 15.71 | 16.21 |
| 250 | 5 | 10.71 | 10.45 | 1.61 | 1.27 | −0.69 | −1.51 | −0.61 | −1.21 | 21.98 | 22.02 | 8.77 | 8.34 | 7.74 | 8.10 | 9.23 | 9.37 |
| 500 | 5 | 10.78 | 11.51 | 0.36 | 1.01 | −0.69 | −0.07 | −0.63 | −0.04 | 21.31 | 22.02 | 6.02 | 6.17 | 5.23 | 5.35 | 6.23 | 6.39 |
| 250 | 10 | 13.43 | 13.98 | 0.35 | 0.69 | −0.21 | 0.14 | −1.43 | −1.02 | 17.78 | 18.24 | 4.27 | 4.20 | 4.17 | 4.03 | 5.28 | 4.97 |
| 500 | 10 | 13.08 | 13.14 | 0.19 | 0.26 | −0.13 | −0.05 | −0.58 | −0.50 | 17.36 | 17.54 | 3.05 | 2.97 | 2.74 | 2.67 | 3.42 | 3.23 |
| 500 | 15 | 13.75 | 13.72 | 0.17 | 0.14 | −0.03 | −0.05 | −0.65 | −0.72 | 16.54 | 16.50 | 2.18 | 2.13 | 2.14 | 2.12 | 2.54 | 2.56 |
| DGP Chudik et al. (2011) | | | | | | | | | | | | | | | | | |
| 100 | 5 | −0.28 | −0.42 | 1.71 | 0.94 | −1.48 | −2.01 | 0.70 | 0.18 | 8.60 | 8.89 | 10.82 | 10.21 | 11.15 | 11.47 | 12.04 | 11.77 |
| 250 | 5 | −0.04 | −0.10 | 0.37 | 0.51 | −1.07 | −0.99 | −0.01 | −0.14 | 5.36 | 5.19 | 5.73 | 5.80 | 6.83 | 6.49 | 6.67 | 6.64 |
| 500 | 5 | −0.01 | −0.04 | 0.14 | 0.24 | −0.22 | −0.07 | 0.12 | 0.20 | 3.69 | 3.71 | 4.18 | 4.08 | 5.08 | 4.71 | 4.95 | 4.44 |
| 250 | 10 | −0.01 | −0.10 | 0.10 | 0.01 | −0.07 | −0.06 | 0.03 | −0.08 | 2.73 | 2.72 | 2.80 | 2.74 | 3.17 | 3.11 | 3.97 | 3.99 |
| 500 | 10 | 0.15 | 0.03 | 0.17 | 0.08 | 0.11 | 0.09 | 0.00 | −0.02 | 1.93 | 1.90 | 2.06 | 1.92 | 2.24 | 2.14 | 2.60 | 2.68 |
| 500 | 15 | 0.07 | −0.04 | 0.13 | 0.04 | 0.03 | −0.03 | 0.18 | −0.05 | 1.50 | 1.44 | 1.49 | 1.40 | 1.59 | 1.53 | 2.23 | 2.12 |
| DGP Ahn et al. (2013) | | | | | | | | | | | | | | | | | |
| 100 | 5 | 3.68 | 3.75 | 0.11 | 0.56 | −0.23 | 0.08 | 0.00 | 0.06 | 10.50 | 10.98 | 9.54 | 9.66 | 5.73 | 5.78 | 6.75 | 6.42 |
| 250 | 5 | 1.63 | 1.84 | −0.05 | 0.02 | 0.06 | 0.09 | −0.01 | 0.07 | 6.93 | 6.89 | 6.81 | 7.12 | 3.41 | 3.25 | 3.59 | 3.43 |
| 500 | 5 | 0.75 | 0.85 | 0.14 | −0.50 | 0.12 | −0.03 | 0.13 | −0.06 | 4.89 | 5.03 | 6.26 | 6.08 | 2.27 | 2.36 | 2.34 | 2.41 |
| 250 | 10 | 2.14 | 1.88 | −0.59 | −0.79 | −0.03 | −0.07 | −0.11 | −0.19 | 5.81 | 5.77 | 2.99 | 2.90 | 2.06 | 2.05 | 2.49 | 2.51 |
| 500 | 10 | 1.10 | 0.90 | −0.67 | −0.69 | 0.00 | −0.10 | 0.00 | −0.09 | 4.04 | 4.01 | 2.28 | 2.49 | 1.42 | 1.46 | 1.61 | 1.70 |
| 500 | 15 | 1.13 | 0.89 | −0.47 | −0.52 | 0.05 | 0.02 | −0.05 | −0.07 | 4.02 | 3.75 | 1.56 | 1.67 | 1.09 | 1.09 | 1.50 | 1.55 |

This table reports bias and RMSE results for the CCE, PC, ALS* and RS estimators generated by the DGPs 6-8. The results are based on 1000 replications.

lower RMSE when $T = 5$ and $N$ is large, while the results indicate small advantages for the ALS$^*$ estimator when $T \geq 10$.

The third panel of Table 2.4 presents the results for DGP8. The GMM estimators are the least biased estimators in this setting. The ALS$^*$ estimator exhibits the smallest RMSE for all $(N, T)$ combinations with only slight advantages over the RS estimator. For example, for $T = 10$ and $N = 500$, the RMSE of the ALS$^*$ estimator is about 40% lower than the RMSE of the PC estimator and more than 60% lower than the RMSE of the CCE estimator. The CCE estimator is problematic in this setting, since the expectation of the loadings is equal to zero. The PC estimator is problematic in this small $T$ setting. However, the RMSE is lower for larger samples with $T = 15$ and $N = 500$.

## 2.8. Conclusion

In this paper we compare three existing approaches for estimating factor augmented panel data models. We argue that the PC estimator can be seen as an estimated analog of the optimal transformation for eliminating the common factors from the data. The CCE estimator applies a data transformation that has the important advantage that the weighting scheme is fixed and does not involve any sampling error. This ensures that the estimator is consistent even if $T$ is fixed, whereas the PC estimator requires much more restrictive assumptions (such as i.i.d. errors) whenever $T$ is fixed. The third estimation approach corresponds to the nonlinear GMM estimators of Ahn et al. (2013) and Robertson and Sarafidis (2015). In contrast to the PC and CCE estimators, the number of parameters does not depend on $N$, which makes these estimators particularly attractive for models with large $N$ and small $T$.

In this paper we focus on the typical micro panel data setup where $T$ is small compared to $N$. Since for an approximate factor model the consistency of the PC estimator requires $T \to \infty$, it is interesting to investigate how large $T$ needs to be for ensuring the PC estimator to be approximately unbiased. Our Monte Carlo experiments indicate that for all data generating mechanisms considered in this paper $T = 10$ is already sufficient to achieve reasonable small sample properties of the PC estimator.

Some versions of the estimators impose normalization conditions that may be problematical in practice. For the original CCE estimator, we propose a simple weighting scheme based on a decomposition similar to Mundlak (1978). The resulting CCE(M) estimator is able to escape the endogeneity bias that may occur in the $\sqrt{N}$ vicinity of the normalization failure at the cost of introducing a larger number of additional auxiliary

parameters. The PC, ALS* and RS estimators sidestep the possibility of a normalization failure and perform well in all our Monte Carlo experiments. Sometimes the CCE and ALS* estimators perform slightly better than the PC estimator, but in other Monte Carlo setups the PC estimator tends to outperform all other competitors. Furthermore, we show that for small $T$ the selection criteria for the number of factors proposed by Bai and Ng (2002) and Ahn and Horenstein (2013) may be inconsistent, whereas the BIC criteria of Ahn et al. (2013) and Robertson and Sarafidis (2015) perform well.

# Chapter 3.

# Empirical Challenges for Optimal Portfolio Selection

## 3.1. Introduction

The prominent maximum Sharpe ratio (MSR) portfolio of Markowitz (1952) requires reliable estimates of expected returns and the covariance matrix of stock returns. When these moments are estimated using the sample analogs from historical data (the so-called plug-in method), this often leads to extreme portfolio weights that excessively fluctuate over time and typically exhibit poor out-of-sample performance (see, e.g., Michaud, 1989; Best and Grauer, 1991; Chopra and Ziemba, 1993; DeMiguel et al., 2009b). Estimating expected returns from historical data is particularly prone to errors. Additionally, errors in estimating the first moment have a larger impact on the estimation of portfolio weights than errors in estimating the second moment (see, e.g. Chopra and Ziemba, 1993; Jagannathan and Ma, 2003; DeMiguel et al., 2009a). Therefore, in practice and research, information regarding the first moment is often ignored, leading to the estimation of the global minimum variance (GMV) portfolio.

However, obtaining reliable estimates of the covariance matrix becomes particularly challenging in high-dimensional applications, where the number of investable assets ($N$) is roughly equivalent to the number of available time series ($T$). In this case, the sample covariance matrix suffers from the curse of dimensionality. This becomes most evident when the number of assets exceeds the number of time periods, as the covariance matrix becomes singular. Nevertheless, considering that the number of parameters to be estimated in an $N \times N$ covariance matrix is $N(N+1)/2$, the sample size $T$ should be substantially larger than $N$ to avoid unreliable parameter estimates. However, this is often not the case in modern portfolios with numerous assets.

To mitigate the impact of estimation errors in the covariance matrix, various approaches have been introduced in the literature. These include, for instance, regulariza-

tion methods such as the well known linear shrinkage estimators (Ledoit and Wolf, 2003, 2004a,b) or more recently, nonlinear shrinkage estimators (e.g., Ledoit and Wolf, 2020), as well as dimensionality reduction techniques such as factor models (e.g., De Nard et al., 2019). Particularly in high dimensional applications, these approaches typically improve the out-of-sample performance of estimated GMV portfolios.

Another prominent issue is that the unrestricted MSR or GMV optimization often results in a large number of negative weights. From a conventional perspective, this implies that an investor is taking short positions on the corresponding stocks with negative weights. Besides the fact that the feasibility of short-selling is severely limited in practice, the presence of negative weights in a portfolio leads to high risk exposure and turnover rates. To address this concern, various strategies have been proposed in the literature. These strategies either completely exclude short sales as in Jagannathan and Ma (2003) or impose constraints on the weight vector as in DeMiguel et al. (2009a) and Fan et al. (2012).

Both portfolio models can be represented as regression models. This opens up the opportunity, for instance, to restrict the 1-norm of the weights by employing LASSO techniques in the regression estimation. While the unconstrained regression of Kempf and Memmel (2006) provides the normalized weights of the GMV portfolio (which sum to unity), the well-known regression of Britten-Jones (1999) for the MSR portfolio yields weights that require normalization to obtain the plug-in MSR weights. This subsequent normalization is, however, not without difficulties when applying L1-regularization to the Britten-Jones regression.

This paper examines a variety of aspects related to empirical challenges in portfolio selection, which are further explored in the subsequent sections. In Section 3.2, we consider the MSR and GMV portfolios and their representation as regression models. We establish a relationship between the Kempf-Memmel regression for the GMV weights and the Britten-Jones regression for the MSR portfolio. This relationship is used to derive a variant of the Britten-Jones regression that directly provides the normalized weights of the MSR portfolio.

In Section 3.3, we examine the interpretation of negative weights and illustrate how the classical short-sale strategy can lead to excessive leverage and exposure. We propose an alternative strategy involving put options, which results in considerably more moderate portfolio weights and constraints exposure. Mathematically, both strategies result in different normalization schemes, as the short-selling strategy normalizes the weights such that they sum to unity, while the put option strategy requires the sum of the absolute weights to be one.

In Section 3.4, we focus on three major challenges for estimating portfolio weights. Using simulation studies, we first investigate the impact of estimation uncertainties in the covariance matrix on portfolio performance under various concentration ratios ($N/T$). Second, we emphasize the detrimental effects of substantial estimation uncertainties in expected returns on the accuracy of portfolio weight estimation. Third, we consider the less prominent effect of the challenging weight normalization in the classical short-selling strategy.

In the empirical application in Section 3.5, we analyze the out-of-sample performance of 15 models using two datasets characterized by different concentration ratios ($N/T$). In addition to the plug-in estimators for the MSR and GMV portfolios, we consider alternative regularization methods for the covariance matrix and portfolio weights, including LASSO versions of the Kempf-Memmel and (normalized) Britten-Jones regressions. Section 3.6 concludes.

## 3.2. The MSR and GMV portfolios

Following the literature on optimal portfolio selection, we assume that the $N \times 1$ vector of returns $\boldsymbol{R}_t = (R_{1,t}, \ldots, R_{N,t})'$ is an independent and identically distributed random vector with $\mathbb{E}(\boldsymbol{R}_t) = \boldsymbol{\mu}$ and $\mathbb{E}(\boldsymbol{R}_t - \boldsymbol{\mu})(\boldsymbol{R}_t - \boldsymbol{\mu})' = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is positive definite.[1] The efficient frontier is obtained by minimizing the Lagrangian objective function

$$\min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - \lambda(\boldsymbol{w}'\boldsymbol{\mu} - \mu_p), \tag{3.1}$$

where $\boldsymbol{w}$ is the $N \times 1$ vector of portfolio weights, $\lambda$ is the Lagrangian multiplier and $\mu_p$ denotes the target portfolio return.[2] The vector of optimal portfolio weights $\boldsymbol{w}^*$ is proportional to the "raw" (non-normalized) weights $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Normalizing the weights such that they sum up to unity yields the weights that maximize the Sharpe ratio $SR(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{\mu}/\sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}}$ for $\boldsymbol{w} \in \mathbb{R}^n$, given by

$$\text{MSR:} \quad \boldsymbol{w}^*_{MSR} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{1}_N'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}, \tag{3.2}$$

[1] Note that this assumption rules out riskless assets. If the portfolio includes riskless assets, we subtract the (maximum of the) riskless yield from the returns of the risky assets. The weights of the riskless asset can be obtained in a second step by optimizing the utility of a convex combination of the tangential portfolio and the riskless asset.

[2] In many textbooks the optimization problem includes a second constraint that the weights sum up to unity. In this case the solution depends on $\mu_p$. The combination of $\mu_p$ and the associated minimum volatility yields the efficient frontier. The tangential portfolio that maximizes the Sharpe ratio is identical to a portfolio with weight vector $\boldsymbol{w}^*$.

where $\mathbf{1}_N$ denotes an $N \times 1$ vector of ones.

The (global) *minimum-variance* portfolio (GMV) is obtained by minimizing the portfolio variance subject to the normalization $\sum_{i=1}^{N} w_i = 1$. The resulting portfolio is equivalent to a MSR portfolio assuming that the means of all assets are identical:

$$\text{GMV:} \quad \boldsymbol{w}_{GMV}^{*} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_N}{\mathbf{1}_N'\boldsymbol{\Sigma}^{-1}\mathbf{1}_N}.$$

In practice, the moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and must be replaced by estimated versions, such as the sample moments $\widehat{\boldsymbol{\mu}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{R}_t$ and $\widehat{\boldsymbol{\Sigma}} = T^{-1}\sum_{t=1}^{T}(\boldsymbol{R}_t - \widehat{\boldsymbol{\mu}})(\boldsymbol{R}_t - \widehat{\boldsymbol{\mu}})'$. The reliable estimation of the expected return vector $\boldsymbol{\mu}$ is often considered challenging in practical contexts. Therefore, it seems more appealing to avoid the estimation of the mean vector and instead focus on the GMV portfolio (e.g. DeMiguel et al., 2009b). As noted by Kempf and Memmel (2006), the weights of the GMV portfolio can be obtained by minimizing the objective function

$$Q(\boldsymbol{w}, \alpha) = (\boldsymbol{w}'\boldsymbol{R}_t - \alpha)^2$$

subject to the constraint $\sum w_i = 1$. Inserting $w_1 = 1 - \sum_{i=2}^{N} w_i$, the solution can be obtained by running the regression

$$R_{1,t} = \alpha + w_2(R_{1,t} - R_{2,t}) + \cdots + w_N(R_{1,t} - R_{N,t}) + u_t. \tag{3.3}$$

Let $\widehat{w}_i$ denote the OLS estimator of $w_i$. The minimum variance portfolio weights result as

$$\boldsymbol{w}_{GMV}^{*} = \begin{pmatrix} 1 - \sum_{i=2}^{N}\widehat{w}_i \\ \widehat{w}_2 \\ \vdots \\ \widehat{w}_N \end{pmatrix}.$$

A similar approach can be adopted for obtaining the MSR solution (3.2). Let $\boldsymbol{w}'\boldsymbol{R}_t = \mu_p + u_t$ where $\mu_p = \boldsymbol{w}'\boldsymbol{\mu}$ denotes the portfolio return. Dividing the equation by $\mu_p$ yields

$$1 = \tilde{\boldsymbol{w}}'\boldsymbol{R}_t + \tilde{u}_t, \tag{3.4}$$

where $\tilde{\boldsymbol{w}} = \boldsymbol{w}/\mu_p$ and $\tilde{u}_t = -u_t/\mu_p$. Note that $\mathbb{E}(\tilde{u}_t^2) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}/\mu_p^2$ is the inverse of the squared Sharpe ratio of the portfolio. Hence, minimizing the residual sum of squares $\sum \widetilde{u}_t^2$ maximizes the (absolute) Sharpe ratio of the portfolio. This motivates the regres-

sion approach proposed by Britten-Jones (1999), where the raw weights are obtained from a regression of ones on the asset returns $R_{1,t}, \ldots, R_{N,t}$.

A drawback of the Britten-Jones regression is that the subsequent normalization of the weights affects statistical inference. To impose the normalization $\sum w_i = 1$, we adopt the Kempf-Memmel regression (3.3). Due to the idempotency of the projection matrix that is used to transform data to deviations from their mean, the first order condition of the empirical analog of (3.1),

$$\begin{aligned} \boldsymbol{w} &= \lambda \left[ \sum_{t=1}^{T} (\boldsymbol{R}_t - \overline{\boldsymbol{R}})(\boldsymbol{R}_t - \overline{\boldsymbol{R}})' \right]^{-1} \sum_{t=1}^{T} \boldsymbol{R}_t \\ &= \lambda \left[ \sum_{t=1}^{T} \boldsymbol{R}_t (\boldsymbol{R}_t - \overline{\boldsymbol{R}})' \right]^{-1} \sum_{t=1}^{T} \boldsymbol{R}_t \cdot 1, \end{aligned}$$

implies that the normalized weights are proportional to the IV estimator of $\boldsymbol{w}$ in the regression (3.4) by using $\boldsymbol{R}_t$ as a vector of instrument variables (IV). The two-stage least-squares interpretation of the IV regression entails that the IV regression is equivalent to running an ordinary regression of the form

$$R_{1,t} = \gamma \widehat{c}_t + w_2(R_{1,t} - R_{2,t}) + \cdots + w_N(R_{1,t} - R_{N,t}) + u_t^* \ , \tag{3.5}$$

where $\widehat{c}_t = \boldsymbol{R}_t' \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}$ denotes the fitted value from a regression of ones on $\boldsymbol{R}_t$. Accordingly, the normalized Britten-Jones regression just replaces the constant in the Kempf-Memmel regression by the variable $\widehat{c}_t$.

These results can be used to construct regularized versions of the estimators in order to improve the small sample properties. For the LASSO variant of regression (3.3), the returns are mean-adjusted in order to remove the (unrestricted) constant from the regression. Note that the usual standardization of variables would disregard information about the assets' risk and is hence deactivated when conducting the LASSO regression. For the normalized version of the Britten-Jones regression (3.4), we first remove the term $\widehat{c}_t$ from the regression in (3.5) by applying the theorem of Frisch and Waugh (1933) and Lovell (1963): First, a regression of the dependent variable on $\widehat{c}_t$ is conducted. Second, regressions of the independent variables on $\widehat{c}_t$ are performed. To obtain the normalized coefficients $w_2, ..., w_N$, the residuals resulting from the first-step regression are regressed on the residuals originating from the second-step regressions. When applying LASSO, we therefore avoid to shrink the parameter $\gamma$ towards zero. Since this regression does not include a constant term, the variables are not mean-adjusted. Additionally, it is

important to note that the weight of the return used as dependent variable results as one minus the sum of all other weights, e.g. $w_1 = 1 - \sum_{i=2}^{N} w_i$, which means that this weight is exempt from shrinkage towards zero. However, the regressions (3.3) and (3.5) are invariant to the choice of the reference asset and therefore, an asset return expected to exhibit a large weight is selected as dependent variable, see Section 3.5.2 for details.

## 3.3. On the interpretation of negative weights

It is a matter of fact that the optimal portfolio regularly exhibits negative weights. For illustration let us consider a portfolio of just two assets. The covariance matrix of the two returns $\boldsymbol{R} = (R_1, R_2)'$ is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad \text{with} \quad \boldsymbol{\Sigma}^{-1} = \varphi \begin{pmatrix} \frac{\sigma_2}{\sigma_1} & -\varrho \\ -\varrho & \frac{\sigma_1}{\sigma_2} \end{pmatrix},$$

where $\varphi = [\sigma_1\sigma_2(1-\rho^2)]^{-1}$ and $\varrho$ denotes the correlation between the two returns. Accordingly, for the GMV portfolio, one weight becomes negative whenever

$$\varrho > \min\left(\frac{\sigma_2}{\sigma_1}, \frac{\sigma_1}{\sigma_2}\right),$$

whereas for the tangential (MSR) portfolio, the condition for a negative weight is

$$\varrho > \min\left(\frac{\mu_1\sigma_2}{\mu_2\sigma_1}, \frac{\mu_2\sigma_1}{\mu_1\sigma_2}\right).$$

Since in practice the correlation among stock price returns is typically large, the occurrence of negative weights is highly probable in empirical applications. For example, concerning the dataset used for the empirical application in Section 3.5, we observe that 45% of the weights of the sample GMV portfolio and nearly 50% of the sample MSR weights are negative.

What does a negative weight mean? Since a positive weight tells the investor how many stocks to *buy*, a negative weight suggests that the investor should *sell* a certain amount of stocks and invest the resulting cash flow in other stocks. Accordingly, a negative weight suggests that the investor holds a short position in the respective asset, which is opened by borrowing shares (or other assets) corresponding to the negative weight. The investor sells the borrowed shares and employs the cash-flow for purchasing shares with positive weights. This investment plan will be called *short-selling strategy*.

Another possibility is to adopt a *put option strategy*. For this purpose, the investor

buys a put option, which we assume to have a $1:1$ inverse relationship with the underlying asset.[3] Essentially, this strategy shifts the minus sign of the weight to the asset's return, as the put option strategy involves a positive investment in an asset with an inverse return. While these two strategies may appear similar at first glance, they differ substantially in some important aspects.

For concreteness let us consider a simple example, where $\mathbb{E}(\boldsymbol{R}) = \boldsymbol{\mu} = (0.18, 0.08)'$ and the covariance matrix is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1.2 \\ 1.2 & 1 \end{pmatrix}. \tag{3.6}$$

The MSR weights are given by $w_1^* = 3$ and $w_2^* = -2$. While this scenario does not seem unrealistic, no investor would pursue the implied short-selling strategy. Note that according to this investment plan, the investor invests 3 times the wealth in asset 1, resulting in an exposure of 3. Consequently, the portfolio variance is as large as 7.6, which even exceeds the variances of the individual assets. In contrast, the put option strategy shifts the minus sign from the weight to the corresponding return yielding the return vector $\widetilde{\boldsymbol{R}} = (R_1, -R_2)'$ with covariance matrix

$$\begin{pmatrix} 2 & -1.2 \\ -1.2 & 1 \end{pmatrix} \tag{3.7}$$

and the MSR weights $\widetilde{\boldsymbol{w}}^* = (0.6, 0.4)$. These weights make much more sense as there is no need to borrow a huge amount of asset 2 in order to invest it in asset 1. From a mathematical point of view, the main difference between the two shorting strategies is that the short-selling strategy implies the normalization $\sum w_i^* = 1$, whereas the put option strategy adopts the normalization $\sum |w_i^*| = 1$. Accordingly, the put option MSR weights are obtained by dividing the raw weights $v_i$ by $\sum |v_i|$. A negative sign of the resulting weight implies the investment in a put option position.

It is important to notice that the Sharpe ratios of both shorting strategies with weight vectors $(3, -2)$ and $(0.6, -0.4)$ are identical. This is due to the fact that the short-selling strategy inflates means and standard deviations by the same factor. Specifically, the

[3] In practice, the (absolute) leverage of an option often exceeds 1. In such cases, the returns (and their moments) are multiplied by the leverage, leading to the division of the raw weights by the same leverage factor.

relationship between the weights of the two shorting strategies is

$$\widetilde{w}_i^* = \frac{\sum_{j=1}^N w_j^*}{\sum_{j=1}^N |w_j^*|} w_i^*.$$

The adjustment factor in front of $w_i^*$ is smaller or equal to one (in our example this factor is equal to 0.2). Hence, the standard deviation and (absolute) return of the put option strategy are typically (much) smaller. Furthermore, in most applications, the weights implied by the put option strategy are far more appealing than the excessively fluctuating weights from the short-selling strategy.

Another problem with the short-selling approach occurs if the sum of the raw weights $v_1, \ldots, v_N$ is negative. In such cases, the standard (short-selling) normalization $w_1^* + ... + w_N^* = 1$ switches the sign of the raw weights. In our simple example, this situation can occur if both mean returns are negative. Assume that $\boldsymbol{\mu} = (-0.6, -0.4)$. In this scenario, the MSR portfolio with the standard normalization yields $\boldsymbol{w^*} = (0.6, 0.4)$. However, the conventional interpretation suggests that the investor should go long in both assets, although both returns are negative. Such a portfolio is clearly not optimal. If we adopt the put option normalization with $|w_1^*| + |w_2^*| = 1$, the optimal portfolio results as $(-0.6, -0.4)$, implying to buy put options for both assets. This approach yields a positive and, indeed, a *maximum* Sharpe ratio.

It is important to note that the two different shorting strategies result in different GMV portfolios. For our simple example above, the put option strategy implies that the negative sign of the second weight shifts to the return of the second asset. Accordingly the covariance matrix of this portfolio with $\widetilde{\boldsymbol{R}} = (R_1, -R_2)'$ is given in (3.7). The GMV weights are obtained as $\widetilde{\boldsymbol{w}} = (0.407, 0.593)'$. The corresponding portfolio variance is 0.104. Notice that this variance is substantially smaller than the variance of the standard GMV portfolio. It might seem counterintuitive to observe a portfolio with positive weights in long/short positions that possesses a much lower variance than the "gobal minimum variance portfolio allowing for short-selling". However, notice that the covariance matrix (3.7) is much better suited for risk hedging than (3.6), as it includes two negatively correlated assets.

## 3.4. The statistical properties of estimated weights

In this section, we examine the statistical properties of the GMV and MSR estimators based on the model assumptions underlying the standard framework considered in Section 3.2. Specifically, we assume that returns are distributed as $\boldsymbol{R}_t \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It is well

known that in empirical practice, returns are neither normally distributed nor independent. Furthermore, the moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ typically change over time. Notwithstanding these limitations, it is interesting to study the performance of the standard approaches within such admittedly unrealistic laboratory conditions.

In order to mimic the statistical properties of actual return series, we compute the sample mean vector $\widehat{\boldsymbol{\mu}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}$ from the daily returns of 760 stocks from major stock price indices that are also employed in the empirical analysis of Section 3.5.[4] We generate artificial data by drawing $T$ independent observations from a $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ distribution, where $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$ denote the sample mean and covariance matrix of the stock returns. We construct portfolios of size $N$ by randomly selecting $N$ columns without replacement from the $T \times 760$ matrix of realizations. In order to assess the performance of the estimated portfolio weights, we compute the Sharpe ratio (SR) of the resulting portfolios, evaluated using 100 out-of-sample realizations of the returns.[5] Furthermore, we report the correlations between the empirical weights and the optimal weights, considering the hypothetical scenario of known moments.

We focus on three major challenges for estimating the optimal weights in the GMV and MSR approach:

1. the covariance matrix is ill conditioned,
2. the uncertainty about the mean is substantial,
3. the normalization has a crucial effect on the statistical properties of the portfolio.

In the following, we first examine each of the three problems separately before the overall performance of the GMV and MSR approaches is considered in Section 3.5.

### 3.4.1. Estimating the covariance matrix

First, we study the effect of uncertainty about the covariance matrix. Since the estimated GMV portfolio is a function of the sample covariance matrix, the properties of the estimated covariance matrix are crucial for the performance of the GMV portfolio. The weights of the MSR portfolio depend in addition on the estimated mean vector. If the returns are normally distributed, the estimation errors $\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}$ and $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$ are independent. Therefore, it makes sense to study the effect of the estimation errors for the two moments separately. To this end, we initially assume that the mean vector is known for the MSR portfolio. Accordingly, the weights are computed as $\boldsymbol{w}^*(\widehat{\boldsymbol{\Sigma}}) = \widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu} / \mathbf{1}_n' \widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}$.

[4] See Section 3.5.1 for the details of the dataset.

[5] The results are not sensitive to the number of out-of-sample periods. The Sharpe ratios are computed as the average of 1000 Sharpe ratios computed from 100 out-of-sample returns.

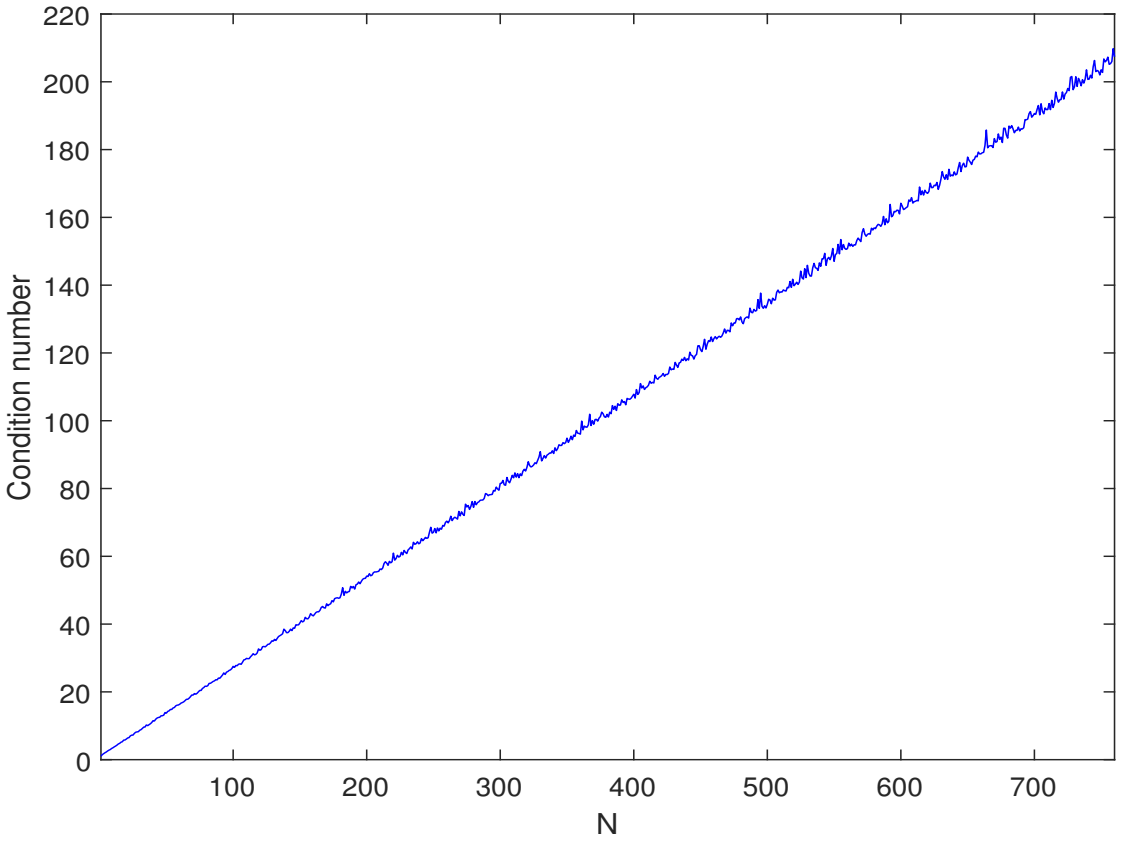



Figure 3.1.: Condition number as a function of $N$

The figure displays the mean of the condition numbers of the empirical covariance matrix based on 1000 random samples of $N$ assets, drawn from the set of all 760 assets contained in our CRSP dataset. See Section 3.5.1 for the details of the dataset.

As shown by Okhrin and Schmid (2006, Corollary 1), the estimation error of the GMV weights is $O_p(N/T)$. This finding extends to the estimator $\boldsymbol{w}^*(\widehat{\boldsymbol{\Sigma}})$. Accordingly, we expect a large estimation error as $N/T$ approaches unity. For $N/T > 1$, the inverse of the sample covariance matrix must be substituted by the (Moore-Penrose) pseudo inverse. A key concern for the small sample properties of the inverse $\widehat{\boldsymbol{\Sigma}}^{-1}$ is that the condition number (e.g. the ratio of the largest to the smallest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$) typically increases with $N$. This is due to the fact that if the assets are driven by common risk factors, the largest eigenvalue of the covariance matrix is $O(N)$, whereas the smallest is $O(1)$. Figure 3.1 presents the condition number as a function of the $N$. The covariance matrices of smaller portfolios are constructed by drawing randomly $N$ assets from the set of all 760 assets without replacement. This experiment is repeated 1000 times and the average condition number is plotted in Figure 3.1. Clearly, the rate of the largest

eigenvalue is well characterized as a linear function of $N$. Since the condition number is a measure for the sensitivity of the output of a linear system with respect to small input changes, the sampling variability tends to increase with the condition number.

This reasoning suggests that some regularization is required for estimating the inverse of a large covariance matrix. A popular method is the shrinkage estimator proposed by Ledoit and Wolf (2003). This estimator is an optimally weighted average of the usual sample covariance matrix and the estimated covariance matrix of a factor model. Some other regularization methods (such as the LASSO variant of the Kempf-Memmel regression) will be considered in Section 3.5. In Table 3.1, we compare the performance of the estimated weights for the MSR and GMV portfolios using (i) the unrestricted sample covariance (sample) and (ii) the regularized covariance matrix of Ledoit and Wolf (2003), referred to as LW1F. Furthermore, we include (iii) an extremely simple "regularization", denoted as "correlation-neglect" (CN). This approach essentially ignores the correlation among the returns yielding weights proportional to $\mu_i/\widehat{\sigma}_i^2$ (resp. $1/\widehat{\sigma}_i^2$) for the MSR (GMV) portfolio. This simple variant reflects the investment strategy of less sophisticated investors (e.g., Benartzi and Thaler, 2001; Eyster and Weizsäcker, 2016).

The results of our Monte Carlo experiment indicate a clear improvement of the estimated weights by adopting the regularization strategy LW1F, in particular, as the number of assets ($N$) approaches the sample size. As illustrated in Table 3.1, the performance of the estimated weights for the GMV portfolio based on the sample covariance deteriorates considerably when $N = T$, whereas the estimated weights perform reasonable (although worse than using the LW1F regularization) when $N$ substantially differs from $T$. The reason is that the inverse of the sample covariance matrix for the case $N = T$ is based on all $N$ eigenvalues, where some of the eigenvalues may happen to be very small. This results in huge and volatile portfolio weights. If $N > T$, the pseudo inverse employs only the $T$ largest eigenvalues, which results in a much more stable behavior of the respective portfolio weights. However, the correlation between the estimated and optimal weights may be substantially diminished. Notably, the simple CN strategy, which neglects the correlation among the assets, performs nearly on par with the GMV estimator using the full sample covariance matrix for moderate portfolio sizes.

A similar pattern emerges for the MSR strategy, when assuming known mean returns. As expected, this additional information improves the SR substantially. However, the relative performance of the estimated weights remains similar (albeit more pronounced) compared to the GMV strategy. In particular, the results suggest that the information contained in the covariances is more important for the MSR strategy, as the performance for the CN estimator, which sets all covariances to zero, performs much worse.

Table 3.1.: Performance of weights with estimated covariance matrix

| | MSR | | | | | | GMV | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sharpe Ratio | | | $\text{corr}(\widehat{w}_i, w_i)$ | | | Sharpe Ratio | | | $\text{corr}(\widehat{w}_i, w_i)$ | | |
| Strategy | sample | LW1F | CN | sample | LW1F | CN | sample | LW1F | CN | sample | LW1F | CN |
| $T$ | | | | | | $N = 50$ | | | | | | |
| 150 | 1.38 | 1.54 | 0.91 | 0.78 | 0.91 | 0.70 | 0.77 | 0.82 | 0.77 | 0.74 | 0.88 | 0.75 |
| 250 | 1.47 | 1.58 | 0.85 | 0.87 | 0.94 | 0.71 | 0.74 | 0.75 | 0.69 | 0.84 | 0.91 | 0.75 |
| 500 | 1.54 | 1.57 | 0.87 | 0.94 | 0.96 | 0.71 | 0.83 | 0.83 | 0.73 | 0.92 | 0.94 | 0.76 |
| 750 | 1.53 | 1.54 | 0.86 | 0.96 | 0.97 | 0.72 | 0.70 | 0.70 | 0.71 | 0.95 | 0.96 | 0.76 |
| 1000 | 1.60 | 1.60 | 0.90 | 0.97 | 0.97 | 0.71 | 0.85 | 0.84 | 0.76 | 0.96 | 0.97 | 0.76 |
| | | | | | | $N = 250$ | | | | | | |
| 150 | 1.64 | 2.61 | 0.94 | 0.33 | 0.80 | 0.56 | 0.73 | 0.97 | 0.80 | 0.25 | 0.69 | 0.48 |
| 250 | 0.15 | 2.71 | 0.87 | 0.04 | 0.83 | 0.57 | 0.07 | 0.93 | 0.71 | 0.08 | 0.74 | 0.49 |
| 500 | 2.23 | 2.83 | 0.87 | 0.67 | 0.87 | 0.57 | 0.76 | 0.94 | 0.72 | 0.63 | 0.81 | 0.49 |
| 750 | 2.54 | 2.86 | 0.97 | 0.79 | 0.89 | 0.57 | 0.93 | 1.05 | 0.83 | 0.76 | 0.85 | 0.49 |
| 1000 | 2.81 | 3.00 | 0.88 | 0.84 | 0.91 | 0.57 | 0.95 | 1.01 | 0.73 | 0.81 | 0.87 | 0.49 |
| | | | | | | $N = 500$ | | | | | | |
| 150 | 2.02 | 3.08 | 0.85 | 0.22 | 0.72 | 0.49 | 0.89 | 1.00 | 0.70 | 0.15 | 0.59 | 0.37 |
| 250 | 2.20 | 3.40 | 0.90 | 0.28 | 0.75 | 0.49 | 0.92 | 1.16 | 0.76 | 0.21 | 0.65 | 0.37 |
| 500 | 0.03 | 3.53 | 0.80 | 0.03 | 0.79 | 0.49 | 0.03 | 0.99 | 0.65 | 0.06 | 0.72 | 0.37 |
| 750 | 2.41 | 3.64 | 0.84 | 0.55 | 0.82 | 0.49 | 0.67 | 1.06 | 0.68 | 0.53 | 0.77 | 0.37 |
| 1000 | 2.95 | 3.61 | 0.90 | 0.68 | 0.84 | 0.49 | 0.94 | 1.13 | 0.75 | 0.66 | 0.80 | 0.37 |
| | | | | | | $N = 760$ | | | | | | |
| 150 | 2.24 | 3.43 | 0.91 | 0.15 | 0.65 | 0.43 | 0.97 | 1.04 | 0.76 | 0.10 | 0.52 | 0.30 |
| 250 | 2.53 | 3.70 | 0.85 | 0.20 | 0.68 | 0.44 | 0.81 | 0.96 | 0.70 | 0.15 | 0.58 | 0.31 |
| 500 | 2.34 | 3.93 | 0.79 | 0.29 | 0.73 | 0.44 | 0.71 | 1.12 | 0.64 | 0.24 | 0.67 | 0.31 |
| 750 | 0.43 | 4.03 | 0.85 | 0.09 | 0.76 | 0.44 | 0.17 | 1.16 | 0.70 | 0.11 | 0.72 | 0.31 |
| 1000 | 2.51 | 4.22 | 0.92 | 0.48 | 0.79 | 0.44 | 0.76 | 1.21 | 0.77 | 0.46 | 0.75 | 0.30 |

The table reports the annualized Sharpe ratios and correlation coefficients with the true weights for the MSR and GMV portfolios, assuming a known mean vector. For estimating the covariance matrix three alternatives are presented: sample and the regularization techniques LW1F and CN. The returns are simulated as a normal distribution employing sample moments from the CRSP dataset. The results are based on 1000 replications.

### 3.4.2. Estimating the mean returns

In this section, we focus on the estimation of the mean vector and assume that the covariance matrix is known. An important empirical challenge is that the portfolio returns are typically small. For the stock return dataset, 90 percent of the 760 daily mean returns are in the interval $[0.0001, 0.18]$ percent and only 7 percent of the mean returns are statistically significant at the 5% significance level. The average $t$-statistic is 1.013 indicating that estimation error is typically of the same magnitude as the mean itself. Since the estimation error is $O_p(T^{-1/2})$, our asymptotic reasoning assumes that the mean return is of the same order of magnitude, that is, $\boldsymbol{\mu} = \boldsymbol{\eta}/\sqrt{T}$, where $\boldsymbol{\eta}$ is some given vector $\boldsymbol{\eta} \in \mathbb{R}^N$. Let us assume that

$$\sqrt{T}\,\widehat{\boldsymbol{\mu}} \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Sigma}).$$

Note that in this setup, the sample means are unbiased but do not converge to their counterparts $\boldsymbol{\mu}$ as $T \to \infty$. Thus, the vector of weights can be represented as

$$\widehat{\boldsymbol{w}}^* = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}}{\mathbf{1}_N'\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta} + \mathbf{1}_N'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}},$$

where $\boldsymbol{\varepsilon} = \sqrt{T}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. The distribution of the $i$-th weight is obtained as

$$\begin{aligned} P\left(\widehat{\boldsymbol{w}}_i^* \leq c\right) &= P\left(\boldsymbol{\delta}_c'\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta} \leq -\boldsymbol{\delta}_c'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}\right) \\ &= P\left(z < \psi_i(c)\right) \;=\; \Phi\left(\psi_i(c)\right), \end{aligned} \tag{3.8}$$

where $z$ is a standard normally distributed random variable with c.d.f. $\Phi(\cdot)$ and

$$\psi_i(c) = \frac{\boldsymbol{\delta}_c'\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}}{\sqrt{\boldsymbol{\delta}_c'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_c}} \quad \text{and} \quad \boldsymbol{\delta}_c = c\mathbf{1} - \boldsymbol{e}_i.$$

The density of $w_i^*$ as a function of $c$ is obtained as $\phi\left(\psi_i(c)\right)\partial\psi_i(c)/\partial c$, where $\phi(\cdot)$ denotes the density of the standard normal distribution. Hence, the distribution is Gaussian only if $\psi_i(c)$ is a linear function. Figure 3.2 depicts the function $\psi_i(c)$ for the first weight. The graph indicates that $\psi_i(c)$ considerably deviates from a linear function. While the implied skewness is relatively small, the kurtosis is substantially larger than that of a normal distribution, as the function tends to become flat for large absolute values of $c$.

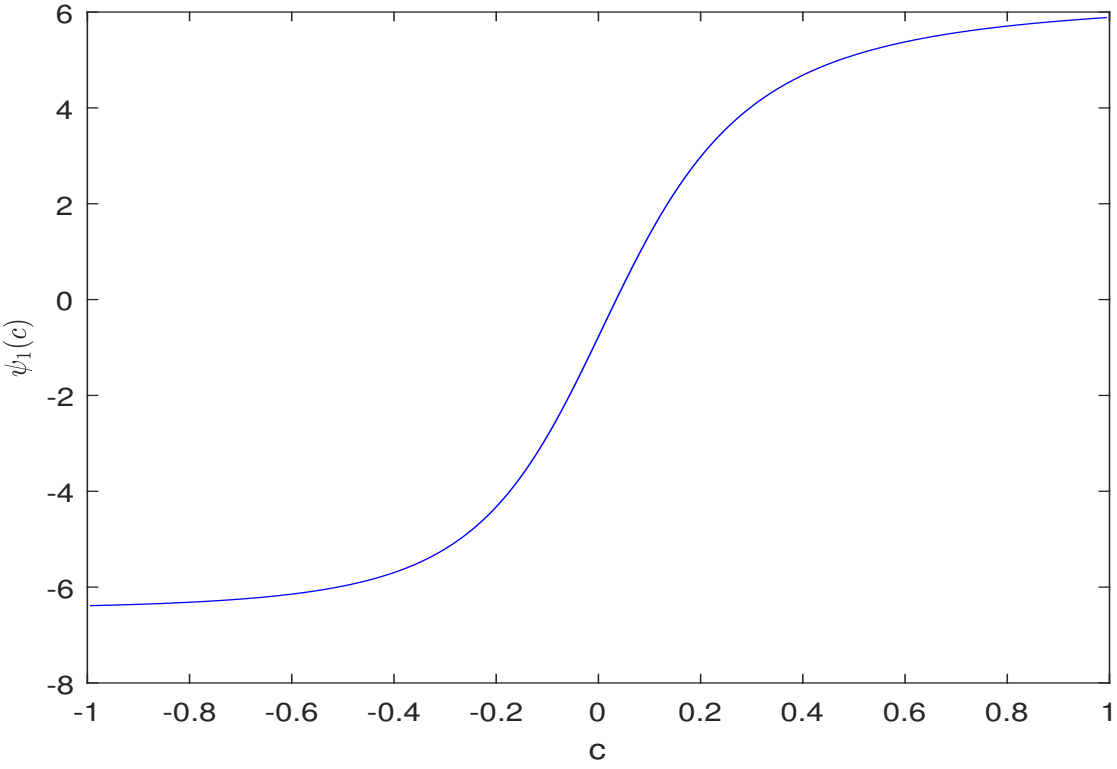



Figure 3.2.: Function $\psi_1(c)$ for the first asset

In our simulations, the asymptotic distribution in (3.8) approximates well the actual distribution of the weights when simulating the data as $\boldsymbol{R}_t \overset{iid}{\sim} \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$.[6] Figure 3.3 compares the asymptotic and the estimated density of the simulated returns for the first stock by using a bandwidth of 0.03. Furthermore, Figure 3.4 depicts the asymptotic densities of the first five weights.

It becomes apparent that the potential to learn about the optimal weights from the data is quite limited. The 90 percent confidence interval is $w_1^* \in [-0.0232, 0.0555]$. This indicates that there is a possibility of the weight to be negative, while the true value is clearly positive. Table 3.2 presents some further information on the distribution of the estimated weights. The mean absolute value of the weights is 0.030 and the mean of their standard deviations is 0.033. As the standard deviation is of similar magnitude, the uncertainty about the weights is substantial. In many cases, the probability of the estimated weight having the wrong sign exceeds 20 percent (see Table 3.2). Therefore, even in an ideal scenario with i.i.d. normally distributed returns and a constant mean,

[6] For normally distributed errors and fixed covariance matrix, the distribution is exact for all $T > N$. If $\boldsymbol{\Sigma}$ is replaced by the sample covariance matrix, the sample size $T$ must be large enough (as in our example) such that the estimation error in $\widehat{\boldsymbol{\Sigma}}$ is negligible.

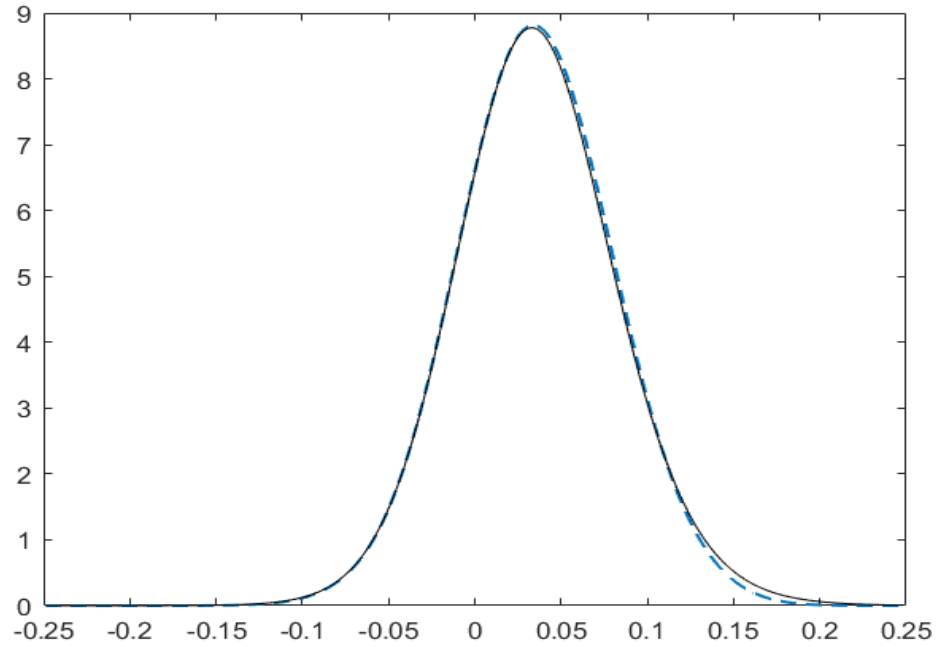



Figure 3.3.: Comparison of the asymptotic density (broken blue) and the empirical density (solid black)

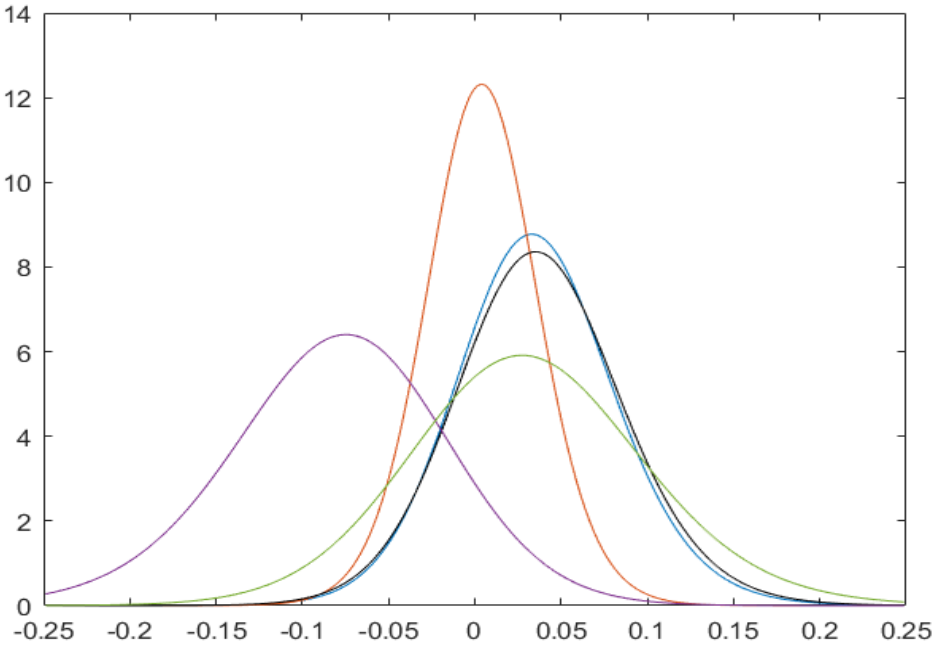



Figure 3.4.: Asymptotic densities for the first 5 estimated weights $\widehat{w}_1^*, \ldots, \widehat{w}_5^*$.

Table 3.2.: Descriptive statistics for the first 10 estimated weights

| $i$ | $w_i^*$ | sample mean | theoretical | std.dev. | $P(\widehat{w}_i^*) < 0$ |
|---|---|---|---|---|---|
| 1 | 0.0349 | 0.0351 | 0.0360 | 0.0342 | 0.15 |
| 2 | 0.0042 | 0.0039 | 0.0043 | 0.0239 | 0.43 |
| 3 | 0.0369 | 0.0380 | 0.0378 | 0.0358 | 0.14 |
| 4 | −0.0782 | −0.0790 | −0.0802 | 0.0445 | 0.97 |
| 5 | 0.0299 | 0.0308 | 0.0314 | 0.0479 | 0.25 |
| 6 | −0.0240 | −0.0270 | −0.0244 | 0.0448 | 0.73 |
| 7 | 0.0216 | 0.0219 | 0.0219 | 0.0302 | 0.24 |
| 8 | 0.0193 | 0.0194 | 0.0198 | 0.0153 | 0.10 |
| 9 | 0.0424 | 0.0438 | 0.0437 | 0.0407 | 0.13 |
| 10 | −0.0085 | −0.0087 | −0.0086 | 0.0118 | 0.78 |

The table reports the sample mean of 10.000 replications of the estimated MSR weights, where the returns are distributed as $\boldsymbol{R}_t \overset{iid}{\sim} \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ and the covariance matrix is treated as known. The column "theoretical" reports the mean of the asymptotic distribution, the column "std.dev." presents the sample standard deviations of the estimated weights $\widehat{w}_i^*$.

the ability to obtain reliable information about the implied optimal weights is severely constrained.

In the previous subsection, we observed that shrinkage towards a simple structure helps to reduce the sampling variability of the estimates. It is therefore interesting to investigate whether a similar shrinkage approach for the mean improves the performance of the estimated weights. To this end, we consider a convex combination of the unrestricted mean and the mean resulting from a factor model with a single factor. The factor model is given by

$$\boldsymbol{R}_t = \boldsymbol{\beta} f_t + \boldsymbol{v}_t,$$

where $\boldsymbol{\beta}$ is an $N \times 1$ vector of loading coefficients. The factor is estimated by the first principal component and $\boldsymbol{\beta}$ is estimated by regressing the returns $R_{1,t}, \ldots, R_{N,t}$ on the estimated factor. The restricted estimator for the means results as

$$\widehat{\boldsymbol{\mu}}_\theta = \theta\, \widehat{\boldsymbol{\mu}} + (1-\theta)\, \widehat{\boldsymbol{\beta}}\, \overline{f} \quad \text{with } 0 \leq \theta \leq 1,$$

where $\widehat{\boldsymbol{\beta}}$ denotes the least-squares estimator of $\boldsymbol{\beta}$ and $\overline{f}$ indicates the ordinary mean of the estimated factor. Note that for $\theta = 1$, the standard MSR weights are obtained, while $\theta = 0$ yields the GMV weights. The shrinkage parameter is varied between $\theta = 1$ and $\theta = 0$ with step size 0.1. It should be noted that, in practice, both estimators $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\mu}}_\theta$

Table 3.3.: Shrinkage estimation of the mean vector (MSR)

| $\theta$ | 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | 0.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $T$ | | | | | | $N = 50$ | | | | | |
| 250 | 0.25 | 0.25 | 0.26 | 0.26 | 0.26 | 0.28 | 0.29 | 0.31 | 0.35 | 0.39 | 0.55 |
| 500 | 0.41 | 0.43 | 0.44 | 0.46 | 0.48 | 0.52 | 0.55 | 0.58 | 0.60 | 0.61 | 0.57 |
| 750 | 0.56 | 0.56 | 0.57 | 0.59 | 0.63 | 0.66 | 0.68 | 0.71 | 0.72 | 0.73 | 0.61 |
| 1000 | 0.62 | 0.62 | 0.63 | 0.64 | 0.66 | 0.67 | 0.68 | 0.71 | 0.74 | 0.74 | 0.64 |
| $T$ | | | | | | $N = 250$ | | | | | |
| 250 | 0.46 | 0.47 | 0.49 | 0.48 | 0.50 | 0.51 | 0.52 | 0.51 | 0.52 | 0.53 | 0.45 |
| 500 | 0.79 | 0.79 | 0.79 | 0.77 | 0.79 | 0.79 | 0.80 | 0.81 | 0.80 | 0.76 | 0.50 |
| 750 | 0.91 | 0.92 | 0.92 | 0.92 | 0.92 | 0.93 | 0.93 | 0.93 | 0.95 | 0.94 | 0.61 |
| 1000 | 1.13 | 1.14 | 1.14 | 1.14 | 1.15 | 1.16 | 1.17 | 1.17 | 1.15 | 1.06 | 0.56 |
| $T$ | | | | | | $N = 500$ | | | | | |
| 250 | 0.57 | 0.58 | 0.60 | 0.60 | 0.61 | 0.62 | 0.62 | 0.60 | 0.57 | 0.49 | 0.33 |
| 500 | 1.09 | 1.10 | 1.09 | 1.10 | 1.10 | 1.10 | 1.10 | 1.07 | 1.03 | 0.95 | 0.42 |
| 750 | 1.26 | 1.26 | 1.27 | 1.28 | 1.29 | 1.29 | 1.29 | 1.28 | 1.26 | 1.11 | 0.45 |
| 1000 | 1.40 | 1.40 | 1.41 | 1.41 | 1.42 | 1.43 | 1.44 | 1.44 | 1.43 | 1.36 | 0.62 |
| $T$ | | | | | | $N = 760$ | | | | | |
| 250 | 0.75 | 0.75 | 0.75 | 0.76 | 0.77 | 0.78 | 0.78 | 0.81 | 0.79 | 0.69 | 0.40 |
| 500 | 1.25 | 1.25 | 1.26 | 1.27 | 1.28 | 1.28 | 1.28 | 1.27 | 1.22 | 1.09 | 0.37 |
| 750 | 1.55 | 1.54 | 1.55 | 1.55 | 1.56 | 1.56 | 1.55 | 1.55 | 1.54 | 1.39 | 0.50 |
| 1000 | 1.69 | 1.69 | 1.69 | 1.69 | 1.70 | 1.70 | 1.70 | 1.69 | 1.66 | 1.56 | 0.58 |

The table reports annualized Sharpe ratios for the MSR portfolio using the restricted mean estimator with a known covariance matrix. The returns are simulated as a normal distribution employing sample moments from the CRSP dataset. The results are based on 1000 replications.

might yield rather different values. Since the mean return of the market factor and the loadings are typically positive, all elements of $\widehat{\boldsymbol{\mu}}_\theta$ are positive as well, whereas the mean of some assets may be negative.

Table 3.3 shows that some shrinkage of the mean improves the performance of the MSR portfolio, at least in our laboratory conditions of i.i.d. returns with known covariance matrix and constant mean and variances. The optimal shrinkage parameter is around $\theta = 0.1$ for a smaller set of assets and increases slightly as $N$ increases. The relative gain relative to the original MSR portfolio ($\theta = 1$) is sizable but the gain deteriorates slightly as $N$ gets larger.

### 3.4.3. The effect of the normalization

As argued in Section 3.3, the treatment of negative weights is important for the performance of the GMV and MSR strategies. The short-selling strategy corresponds to the standard solution that allows for short-selling and imposes the normalization that the weights sum up to unity. Conversely, the put option strategy adopts an alternative normalization, requiring that the sum of the *absolute* weights is equal to one. We argued that the short-selling strategy typically involves a much higher risk due to the leverage effect implied by substantial asset borrowing.

In this section, we perform experiments to examine the effect of the normalization. In addition to the short-selling and put option strategies, we also consider the optimal weights under the no short-selling constraint $w_i > 0$ for all $i = 1, \ldots, N$. The return data is generated as before by using the historical means and covariances of the dataset described in Section 3.5.1. The sample size of the estimation sample is $T = 1000$ and LW1F regularization is applied to the covariance matrix.

The standard approach with short-selling implies that the investor borrows the shares with negative weights and invests the cash flow in shares with positive weights. As a result, such a portfolio allocates an amount that exceeds the initially investable capital. This amount is proportional to the sum of the positive weights $\xi = \sum_{i=1}^{N}(w_i + |w_i|)/2$. The factor $\xi \geq 1$ is labeled as the "exposure factor". It is related to the leverage factor $\ell = \sum |w_i| / \sum w_i$ via $\ell = 2\xi - 1$. Note that $\xi$ is larger than one whenever there is at least one negative weight and, therefore, $\xi > 1$ implies $\ell > \xi$.

Table 3.4 presents the exposure factor ($\xi$) associated with the short-selling strategy (indicated by $\sum w_i = 1$). It turns out that the exposure tends to increase with the number of assets ($N$). This is due to the fact that probability for negative weights increases with $N$. For $N \geq 250$, the exposure factor of the MSR portfolio exceeds 30, implying that the total investment related to the short-selling portfolio is more than 30 times as large as for the put option or no short-selling strategies. For all available 760 stocks, the exposure factor is even larger than 50. The GMV portfolio exhibits considerably lower exposure factors, as it assumes that the mean returns of all stocks are identical and positive, resulting in a lower probability for negative weights.

As noted in Section 3.3, the absolute value of the SR is identical no matter whether the portfolio weights are normalized as $\sum w_i = 1$ or $\sum |w_i| = 1$. From Table 3.4 it turns out, however, that the average SR is slightly smaller when applying the former normalization. This is due to some few cases where the sum of the raw weights, $\sum v_i$, is negative. In this case, the SR of the former normalization is typically negative, whereas the SR with the latter normalization is positive. These (rather few) cases lead to the

Table 3.4.: Alternative normalizations

| norm. | MSR | | | | | GMV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SR | $\mu$ | $\sigma$ | turn | $\xi$ | SR | $\mu$ | $\sigma$ | turn | $\xi$ |
| | | | | | $N = 50$ | | | | | |
| $\sum w_i = 1$ | 0.59 | 18.00 | 208.52 | 134.06 | 19.11 | 0.84 | 11.55 | 13.93 | 5.23 | 1.38 |
| $\sum \lvert w_i \rvert = 1$ | 0.68 | 3.87 | 5.76 | 2.91 | 1.00 | 0.84 | 6.60 | 7.94 | 2.97 | 1.00 |
| $w_i \geq 0$ | 0.81 | 18.04 | 22.56 | 3.54 | 1.00 | 0.82 | 11.95 | 14.80 | 3.41 | 1.00 |
| | | | | | $N = 250$ | | | | | |
| $\sum w_i = 1$ | 1.10 | 130.44 | 143.34 | 39.85 | 30.02 | 1.02 | 10.53 | 10.32 | 1.82 | 2.04 |
| $\sum \lvert w_i \rvert = 1$ | 1.16 | 2.82 | 2.46 | 0.58 | 1.00 | 1.02 | 3.44 | 3.36 | 0.59 | 1.00 |
| $w_i \geq 0$ | 0.97 | 18.74 | 19.75 | 0.77 | 1.00 | 0.85 | 10.55 | 12.53 | 0.75 | 1.00 |
| | | | | | $N = 500$ | | | | | |
| $\sum w_i = 1$ | 1.29 | 132.38 | 122.42 | 23.69 | 36.26 | 1.13 | 10.22 | 9.05 | 1.20 | 2.53 |
| $\sum \lvert w_i \rvert = 1$ | 1.34 | 2.29 | 1.71 | 0.29 | 1.00 | 1.13 | 2.52 | 2.24 | 0.30 | 1.00 |
| $w_i \geq 0$ | 1.00 | 18.68 | 18.85 | 0.39 | 1.00 | 0.88 | 10.50 | 11.86 | 0.38 | 1.00 |
| | | | | | $N = 760$ | | | | | |
| $\sum w_i = 1$ | 1.48 | 220.03 | 140.84 | 23.38 | 52.51 | 1.21 | 10.08 | 8.37 | 0.92 | 2.88 |
| $\sum \lvert w_i \rvert = 1$ | 1.51 | 2.08 | 1.39 | 0.19 | 1.00 | 1.21 | 2.12 | 1.76 | 0.19 | 1.00 |
| $w_i \geq 0$ | 1.04 | 19.01 | 18.30 | 0.26 | 1.00 | 0.94 | 10.73 | 11.47 | 0.25 | 1.00 |

The table reports the annualized Sharpe ratio, return, standard deviation, turnover and exposure for the MSR and GMV portfolios. Both portfolios are evaluated under the the short-selling, put option and no short sale strategies. The returns are simulated as normal distribution employing sample moments from the CRSP dataset. The results are based on estimation samples of $T = 1000$ with 1000 replications. The covariance matrix is subject to LW1F regularization.

reported small differences in the average SR. Additionally, notice that the SR is identical for both normalizations of the GMV portfolio, since in this case, we observe no instance where the sum of the raw weights is negative.

For $N = 50$, the no short-selling portfolio performs slightly better than the other portfolios in terms of the SR for the MSR approach. However, for larger portfolios, the no short-selling portfolio exhibits a considerably lower SR. With respect to the means and variances of the portfolios, the differences are much more pronounced. Due to the high leverage and exposure, both the average mean and volatility of the short-selling portfolio are substantially higher compared to both the put option and no short-selling portfolios.

Table 3.4 also presents the turnover rate (turn) of the portfolios. This indicator measures the extent the assets are reallocated after the holding period ($m = 100$). The turnover is a rough measure of the transaction costs due to the restructuring of the portfolio, see Section 3.5. The short-selling portfolios exhibit a markedly higher turnover (and thus transaction costs) than the put option and no short-selling portfolios.

## 3.5. Empirical Analysis

In this section, we evaluate the out-of-sample performance of alternative portfolio strategies on American stock data. In particular, we focus on the major challenges for estimating portfolio weights that are emphasized in Section 3.4. As mentioned before, the estimation of portfolio weights is highly prone to estimation errors in the moments leading to extreme weights and poor out-of-sample performance. Hence, finding sound estimates for the moments is crucial but not trivial, in particular when when the concentration ratio $N/T$ comes close to one. As noted in Section 3.4.2, the estimation of expected returns via the sample means is particularly challenging, since the estimation error is typically of the same magnitude as the mean itself. In addition, errors in the estimation of the mean returns are more harmful for estimating portfolio weights than errors in the estimation of covariance matrices. This led many authors of the recent academic literature to ignore the data on mean returns and instead focus their efforts on enhancing the estimation of the GMV portfolio. In this section, we compare the performance of several popular estimators of the covariance matrix and furthermore consider investment strategies that focus on the direct estimation of the portfolio weights via LASSO regressions. In addition, we show empirically that the put option strategy is effective in reducing the portfolio variance and typically performs much more desirable with regard to the portfolio weight properties, most notably, when taking into account the mean returns to estimate MSR portfolios.

### 3.5.1. Data

We extracted daily stock data from the Center for Research in Security Prices (CRSP) for the time period 3 January 2000 until 14 November 2019 yielding 5000 daily returns (WRDS, 2020). Our primary dataset comprises 760 stocks for which we have an almost complete return history for the considered time period.[7] To ensure a certain quality standard of the stocks included in our investment universe, we restrict attention to stocks that were constituents of the S&P 500 index on the last trading day of the years between 2010 and 2019 plus constituents of the NASDAQ composite index. All included stocks are common shares that are traded on the NYSE or NASDAQ stock exchanges with no more than ten recorded trading days without trading volume. We chose these requirements to avoid the inclusion of illiquid stocks with high spreads.

In addition, we consider the 49 industry portfolios from Kenneth French's website

[7] We allow for <1% of missing returns during the considered time period. The few missing values are replaced by zeros.

(French, 2020). For this dataset, all stocks from NYSE, AMEX and NASDAQ are assigned to one of 49 industries. We consider the dataset for the time period 1 July 1969 until 15 January 2020 yielding 12750 daily returns. In Appendix B.1, we also examine the same time period as for the CRSP dataset.

### 3.5.2. Alternative Portfolio Selection Strategies

This section provides short descriptions of the alternative estimators we consider. Besides the standard plug-in estimators, we include some very simple regularization strategies for estimating the covariance matrix and compare them to more complex and sophisticated models from the recent literature. In addition, we consider LASSO variants of the regression approaches to portfolio allocation described in Section 3.2. Finally, we examine the performance of an investment strategy, in which the concentration ratio is artificially reduced by subdividing the $N$ assets into smaller blocks. For each of the portfolio estimators, we report results for the short-selling and put option strategies. The weights for an estimation strategy are obtained by applying the following normalizations to the raw weights $\boldsymbol{v}$, respectively:

$$\text{short-selling: } \boldsymbol{w} = \frac{\boldsymbol{v}}{\mathbf{1}_N'\boldsymbol{v}}, \qquad \text{put option: } \boldsymbol{w} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_1}.$$

We include the following estimators in our comparison.

#### Standard models

1. **Plug-in MSR portfolio (MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$))**
   We consider the plug-in MSR portfolio with the expected return and covariance matrix estimated via their sample analogous. Thus, the vector of raw weights is determined as $\hat{\boldsymbol{v}} = \widehat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}$.

2. **Plug-in GMV portfolio (GMV($\widehat{\boldsymbol{\Sigma}}$))**
   For the plug-in GMV portfolio, we employ the sample covariance matrix, assuming that the expected returns are identical for all assets. That is, we ignore the data on expected returns and compute the raw weights as $\hat{\boldsymbol{v}} = \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_N$.

3. **Naive portfolio ($\mathbf{1/N}$)**
   This strategy omits the estimation of both moments and simply invests $1/N$ in all assets. DeMiguel et al. (2009b) claim that it is difficult to outperform the $1/N$ portfolio out-of-sample, as allocation mistakes caused by using this naive strategy

may be less severe than the consequences of estimation errors. Hence, this simple portfolio serves as a benchmark strategy.

**Regularization strategies for the covariance matrix**

4. **Correlation-neglect (CN)**
   This simple regularization strategy solely considers the estimated variances, ignoring the potentially noisy estimates of the correlations among the assets. The estimated covariance matrix results as $\widehat{\mathbf{\Sigma}} = \text{diag}(\hat{\sigma}_1^2, ..., \hat{\sigma}_N^2)$ so that $\hat{v}_i \geq 0$ for all $i$. Thus, the weights for both normalization approaches are identical for the CN portfolio.

5. **Single-factor (1F)**
   This strategy uses the estimated covariance matrix of an exact single factor model, $\widehat{\mathbf{\Sigma}}_{\mathbf{1f}}$, as input for $\mathbf{\Sigma}$ to compute the raw weights. The factor is estimated as the first principal component of the sample covariance matrix.

6. **Linear shrinkage (LW1F)**
   Ledoit and Wolf (2003) propose a shrinkage estimator in the style of Stein (1956) to regularize the covariance matrix, which is a convex linear combination of the market-factor covariance matrix and the sample covariance matrix,

$$\widehat{\mathbf{\Sigma}}_{LW1F} = \omega \widehat{\mathbf{\Sigma}}_{1F} + (1 - \omega)\widehat{\mathbf{\Sigma}},$$

   where $\widehat{\mathbf{\Sigma}}_{1F}$ is the shrinkage target and $\omega$ denotes the shrinkage intensity.[8] The sample covariance matrix is asymptotically unbiased but suffers from substantial estimation error when the concentration ratio comes close to one. The market-factor covariance matrix, on the other hand, has considerably fewer parameters to estimate and contains much less estimation error. We consider $\widehat{\mathbf{\Sigma}}_{LW1F}$ as linear shrinkage estimator in our empirical application, since it is a particularly suitable shrinkage target in finance applications. Ledoit and Wolf (2004a,b), however, suggest the constant correlation matrix and (a multiple of) the identity matrix as alternative shrinkage targets.

7. **Nonlinear shrinkage (LWNL)**
   We consider the (analytical) nonlinear shrinkage estimator of Ledoit and Wolf (2020). The nonlinear shrinkage approach generalizes linear shrinkage to a multiple of the

[8] In order to obtain the optimal shrinkage intensity, Ledoit and Wolf (2003) consider a loss function based on the Frobenius norm, which is a quadratic measure of distance between the shrinkage estimator $\mathbf{\Sigma}_{LW1F}$ and the true covariance matrix. The optimal shrinkage intensity depends on population quantities, which can be estimated from the data to obtain a feasible version. We use the Matlab package covShrinkage, available on Michael Wolf's website.

identity matrix proposed by Ledoit and Wolf (2004b). While the latter approach is equivalent to shrinking the sample eigenvalues to their grand mean with equal intensity, nonlinear shrinkage uses individual (positive or negative) shrinkage intensities for each sample eigenvalue. To overcome the problem that the number of parameters increases with $N$, the sample eigenvalues are transformed with a shrinkage function. Ledoit and Wolf (2020) present an analytical solution to directly estimate the oracle shrinkage function.[9]

8. **Dynamic Model (DCCNL)**
   All previously considered models rely on the assumption that the return data are i.i.d., which may be overly restrictive. However, the accurate estimation of dynamic covariance matrices (via multivariate GARCH models) is very challenging in large-dimensional asset applications due to their complexity and the large number of parameters to estimate. Recently, Engle et al. (2019) managed to robustify the dynamic conditional correlation (DCC) model of Engle (2002) for large $N$ applications by using nonlinear shrinkage of Ledoit and Wolf (2012, 2015) in the estimation of the (static) correlation targeting matrix and using the composite likelihood method of Pakel et al. (2020) to estimate the DCC parameters. For details on the implementation of this DCCNL model, consider the original paper Engle et al. (2019).[10]

#### Regularization strategies for the weights

9. **Short-sale-constrained MSR ($\mathbf{MSR}_+$)**
   A prominent method to improve the stability of portfolio weights is to extend the MSR and GMV optimization problems with short sale constraints on the portfolio weights. In the context of the MSR portfolio, DeMiguel et al. (2009b) emphasize that constraining short sales corresponds to shrinking the mean returns towards the average of the mean returns. In this study, we consider the classical no-short-selling portfolio imposing a lower bound of zero on the portfolio weights.[11]

10. **Short-sale-constrained GMV ($\mathbf{GMV}_+$)**
    For this strategy, the GMV optimization problem is extended with short sale con-

[9] In earlier papers, Ledoit and Wolf (2012, 2015, 2017b) propose numerical strategies to estimate the shrinkage function. As mentioned in Ledoit and Wolf (2020), the analytical solution is much faster and similarly accurate. We use the Matlab routine analytical_shrinkage, available on Michael Wolf's website.

[10] We use the Matlab routine DCC_NL06, available on Michael Wolf's website

[11] Instead of imposing a nonnegativity constraint on the portfolio weights, short-selling can also be limited by constraining the 1-norm of the weight vector when solving the optimization problem, cf. DeMiguel et al. (2009a) and Fan et al. (2012).

straints. Jagannathan and Ma (2003) demonstrate that constraining short sales when estimating the GMV portfolio corresponds to shrinking the extreme elements of the sample covariance matrix.

11. **Kempf-Memmel-LASSO (KML)**

We examine an alternative shrinkage method and consider LASSO variants of the regressions depicted in Section 3.2. The LASSO regression encourages sparsity with regard to the number of model parameters by constraining the L1-norm of the regression coefficients (i.e. the portfolio weights).

We estimate a LASSO variant of the Kempf-Memmel (KM) regression and obtain the coefficients by solving the following minimization problem,

$$\min_{w_2,\dots,w_N} \sum_{t=1}^{T} u_t^2 + \lambda \sum_{i=2}^{N} |w_i|, \tag{3.9}$$

where the variables are mean-adjusted to exclude the parameter of the constant from the shrinkage estimation. However, the variables are not standardized in order to retain the information about the risk of the assets. This type of regularization can eliminate weights from the portfolio (i.e. $w_i = 0$) and thus focuses on the most important assets. The amount of zero-weights depends on the size of the tuning parameter $\lambda$, which controls the shrinkage intensity. The larger $\lambda$, the more weights are set to zero and the corresponding assets are eliminated from the portfolio. A standard approach to select the tuning parameter is cross-validation.

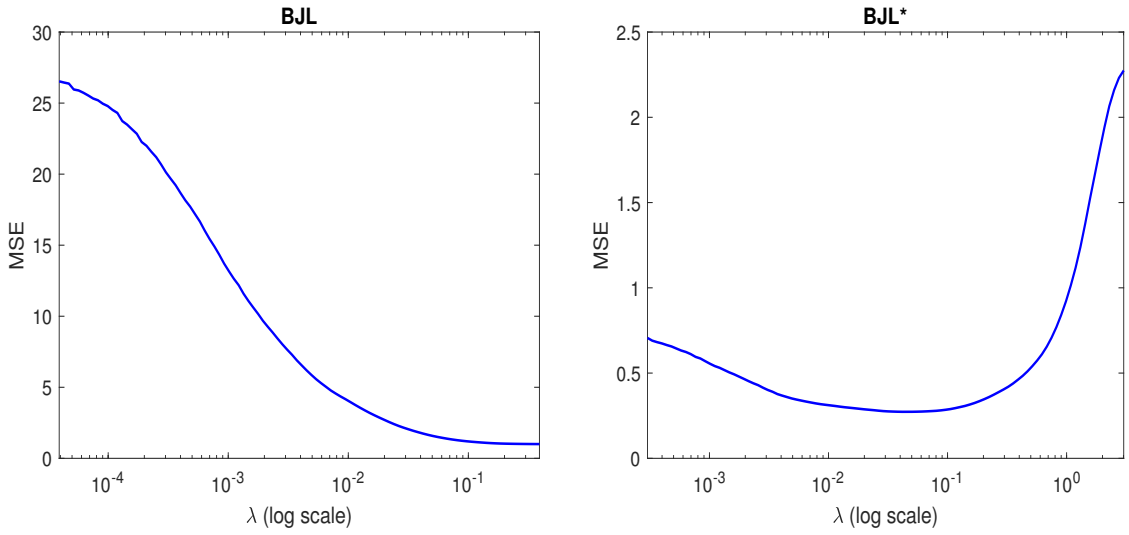
In contrast to the original KM regression, the LASSO variant is sensitive to the choice of the reference asset as a dependent variable, since the corresponding weight is not shrunk but determined by the adding-up constraint $w_1 = 1 - \sum_{i=2}^{N} w_i$. To choose a plausible reference asset, we make an educated guess: We estimate the original KM regression (hence the plug-in GMV portfolio) in a first stage and use the asset with the largest (positive) weight as a reference asset in the LASSO regression. This data-driven procedure increases the probability that the reference asset would not be removed by the L1-regularization.[12]

12. **(Normalized) Britten-Jones-LASSO (BJL$^*$)**

For the LASSO variant of the Britten-Jones (BJ) regression, we consider the second

[12] Frey and Pohlmeier (2016) propose to augment the asset universe by a reference portfolio (e.g. the $1/N$ portfolio) in order to impose shrinkage on all weights. Using their approach yields a quite similar portfolio performance. In general, the choice of the reference asset does not seem to have a major impact on the portfolio performance in large-dimensional applications.

stage of the 2SLS approach depicted in Section 3.2. In contrast to the original BJ regression, this version does not require a challenging normalization such that the weights sum up to unity after estimating the LASSO regression. In addition, the original BJ-regression features a constant variable on the left side (typically a vector of ones), which causes problems when applying L1-regularization with classical cross-validation. Figure 3.5 presents the MSE paths for LASSO versions of the original BJ (BJL) and the normalized BJ (BJL*) regressions for an exemplary estimation period of the CRSP dataset using 5-fold cross-validation and a grid of 100 $\lambda$-values. The MSE of BJL is steadily decreasing and exhibits a minimum at the highest $\lambda$ value suggesting to set all weights equal to zero. Besides the fact that this result is unreasonable, obtaining normalized weights is impossible in this scenario. On the other hand, the MSE of the BJL* is minimized at $\lambda = 0.0418$, which eliminates 500 out of 760 assets in this case.




(a) (b)

Figure 3.5.: Example of the MSE paths for LASSO versions of the original Britten-Jones and the normalized Britten-Jones (BJL*) regressions. The $x$-axis is represented on a logarithmic scale.

In contrast to the MAXSER approach proposed by Ao et al. (2019), our approach does neither require an estimate for the response variable nor to specify a risk constraint.[13]

[13] Ao et al. (2019) propose an unconstrained regression representation that is equivalent to the mean-variance portfolio for a given risk constraint. In contrast to the approach of Britten-Jones (1999), their dependent variable is estimated from the data in order to approximate the correct scaling of the mean-variance portfolio. As a result, the BJ-regression is multiplied by a constant leading to a similar LASSO solution path for the MAXSER and BJL regressions (see Ao et al., 2019, Figure 4).

Instead, applying the L1-regularization to the second stage of the 2SLS approach yields the normalized weights in a direct way. As mentioned in Section 3.2, we remove the constant $\hat{c}_t$ from equation (3.5) by applying the theorem of Frisch and Waugh (1933) and Lovell (1963). We find the reference asset with a similar approach as for the KML regression, using the asset with the largest weight from the original BJ-regression (plug-in MSR) as dependent variable.

### Blocking strategies

Since we expect the estimation error of the covariance matrix to be large whenever the concentration ratio $N/T$ comes close to one, we propose a very simple strategy that reduces the concentration ratio by subdividing the assets into blocks. For this strategy, we utilize the information on $\boldsymbol{\mu}$ by sorting the assets according to their sample means, $\hat{\boldsymbol{\mu}}$, in the respective estimation period and separate the assets into $B$ (approximately) equal-sized blocks. In a second step, the plug-in GMV weights are determined for each of the $b = 1, ..., B$ blocks such that the weights within each block sum up to unity. In a third step, the weights of all blocks are merged back into one vector and the portfolio weights for the blocking strategy are computed as linear combination of the within-block weights, i.e.

$$\hat{w}^*_{block} = \theta_b w^b_{\text{GMV}(\widehat{\boldsymbol{\Sigma}}),\text{i}} \quad \text{for } i = 1, ..., N, \text{ and } b = 1, ..., B,$$

where $w^b_{\text{GMV}(\widehat{\boldsymbol{\Sigma}}),\text{i}}$ denotes the within-block weight of asset $i$ in block $b$ and $\theta_b$ indicates the weight of block $b$ with $\sum_{b=1}^{B} \theta_b = 1$. Finally, the original order of the assets is restored. We consider the following alternative weighting schemes for $\theta_b$.

13. **Equal-weighted (block(eq))**
    The blocks are equally-weighted, hence $\theta_b = 1/B$ for $b = 1, ..., B$.

14. **Ordinal-weighted (block(ord))**
    The blocks are ordinally weighted, hence $\theta_b = b/B!$ for $b = 1, ..., B$, such that the block comprising the assets with the largest historical mean returns receives the largest weight.

15. **In-sample Sharpe ratio (block(SR))**
    We also consider a data-driven weighting scheme, where the weights of the blocks are scaled according to their in-sample Sharpe ratio in the current estimation period. To

The authors select a tuning parameter through a cross-validation procedure that leads to a portfolio attaining a higher Sharpe Ratio than BJL (see their Section 1.5 for details).

ensure that the block weights are positive and sum up to unity, the weight for block $b$ is determined as

$$\theta_b = \frac{\exp(\widehat{\mathrm{SR}}_b)}{\sum_{b=1}^{B} \exp(\widehat{\mathrm{SR}}_b)} \quad \text{for } b = 1, ..., B,$$

where $\widehat{\mathrm{SR}}_b$ denotes the in-sample Sharpe ratio of the assets contained in block $b$ in an estimation period.

### 3.5.3. Methodology for Evaluating the Performance

To compare the performance of alternative portfolio strategies, we use the following *rolling window* procedure to generate a time series of out-of-sample returns for each strategy. Let $T$ denote the total number of available daily returns. We choose an estimation window length of $M$ returns (e.g. 1000 returns corresponding to roughly four years of daily data) to estimate the vector of portfolio weights for each strategy $k$. Next, we choose an investment period of length $h$ (e.g. 50 trading days) to generate out-of-sample returns for each strategy. During an investment period, the number of shares are held constant. In order to determine the out-of-sample returns for the subsequent investment period, we move the estimation window by $h$ and update the portfolio weights, accordingly. This rolling window procedure is continued until the end of the dataset is reached. For instance, if $M = 1000$ and $h = 50$, the first estimation window goes from $t = 1$ to $t = 1000$ and the first investment period from $t = 1001$ to $t = 1050$. The second estimation window goes from $t = 51$ to $t = 1050$ and the second investment period from $t = 1051$ to $t = 1100$ and so on. Eventually, we obtain $(T - M)/h$ weight vectors for $H$ investment periods and $T - M$ out-of-sample returns for each strategy.

As mentioned above, we hold the assets (instead of the weights) fixed during an investment period and rebalance according to the updated weights at the beginning of the subsequent investment period.[14] Let $\tau$ denote an investment period, such that $\tau = 1, ..., H$. We determine for each strategy $k$ the amount of stocks for asset $i$ that are acquired on the rebalancing date of $\tau$ as

$$\hat{s}_{k,i,\tau} = \frac{\hat{w}_{k,i,\tau}}{p_i^\tau}, \tag{3.10}$$

where $\hat{w}_{k,i,\tau}$ denotes the corresponding estimated weight and $p_i^\tau$ indicates the buying

[14] The other way around, thus holding the weights constant, would imply excessive turnover. The weights of stocks in a portfolio constantly change when prices vary, hence maintaining a specific weight vector requires constant trading.

price of asset $i$ in investment period $\tau$.[15] The out-of-sample portfolio return for time period $t$ and strategy $k$ is obtained as

$$\widehat{Rp}_{k,t} = \frac{(\boldsymbol{p}_t - \boldsymbol{p}_{t-1})' \hat{\boldsymbol{s}}_{k,\tau}}{\boldsymbol{p}'_{t-1} \hat{\boldsymbol{s}}_{k,\tau}} \quad \text{for } t = M+1, ..., T \text{ and } \tau = 1, ..., H, \tag{3.11}$$

where $\boldsymbol{p}_t$ and $\boldsymbol{s}_\tau$ are $N$-dimensional vectors.

In order to examine the performance of the alternative portfolio strategies, we use the out-of-sample returns for each strategy to compute the average annualized portfolio return

$$\hat{\mu}_k = \left( \frac{1}{T-M} \sum_{t=M+1}^{T} \widehat{Rp}_{k,t} \right) \times 250, \tag{3.12}$$

the annualized portfolio standard deviation

$$\hat{\sigma}_k = \sqrt{\left( \frac{1}{T-M-1} \sum_{t=M+1}^{T} (\widehat{Rp}_{k,t} - \hat{\mu}_k)^2 \right) \times 250} \tag{3.13}$$

and the annualized portfolio Sharpe ratio

$$\widehat{\mathrm{SR}}_k = \frac{\hat{\mu}_k}{\hat{\sigma}_k} \times 250. \tag{3.14}$$

We do not prioritize these three performance criteria and consider them to be equally important, even if we compare alternative estimators of the GMV portfolio. As pointed out by De Nard et al. (2019), estimators of the GMV portfolio should be primarily evaluated according to the extent to which they minimize the variance while a high return and Sharpe ratio are of secondary importance. However, we want to compare portfolio performances in general and consider the GMV portfolio as an expedient to deal with the estimation error in the sample means.

Since we are particularly interested in the effect of the normalization, we examine the performance of the short-selling and put option strategies that are considered in Section 3.3. The short-selling strategy is implemented by employing a negative weight in Equation (3.10) in case the investor holds a short position in the corresponding asset. The put option strategy, on the other hand, essentially involves only positive investments in either stocks or put options, such that the sum of the absolute weights is equal to one. Hence, we use the absolute weights in Equation (3.10) for this strategy. In order

[15] For instance, if the first investment period ($\tau = 1$) starts in $t = 1001$ then $p_i^\tau = p_{i,1000}$. This assumes that prices are constant after-hours, so that the (daily) return of stock $i$ is determined by the price difference of $p_{i,t+1}$ and $p_{i,t}$.

to mimic the inverse returns of put options, the prices in Equations (3.10) and (3.11) are replaced by shadow prices in case the (original) weights from the optimization are negative. This shadow price for put options is computed as

$$p_{i,t}^{\text{put}} = 2 \times p_i^{\tau} - p_{i,t}, \tag{3.15}$$

such that the price of the put option is identical to the share price at the time of investment in $\tau$.[16]

As noted in Section 3.3, the standard deviation and absolute return of a portfolio applying the normalization $\sum |w_i| = 1$ are typically much smaller than for the classical short-selling strategy, while the Sharpe ratio remains identical, since the mean and standard deviation of the short-selling strategy are inflated by the same factor. However, in our empirical application, we observe occasionally large deviations between the Sharpe ratios, depending on the normalization scheme. There are two reasons for this. Firstly, as noted in Section 3.3, applying the standard normalization $\sum w_i = 1$ reverses the signs of the weights whenever the sum of the raw weights is negative. This implies that mean and standard deviation of the short-selling strategy are inflated by the same (absolute) factor, but with opposite signs. Secondly, and more importantly, the deviations in Sharpe ratios result from the manner in which we conduct our out-of-sample evaluation. The inflation factor for mean and standard deviation is constant throughout an investment period, but changes from one investment period to the next. Hence, following the common practice of computing the portfolio mean and standard deviation for the entire out-of-sample period leads to different Sharpe ratios.[17]

In addition to the performance criteria formulated in (3.12) - (3.14), we are interested in the properties of the weights for each investment strategy. It is well known that implementing investment strategies based on sample estimates of means and the covariance matrix produces excessively fluctuating weights with extreme long and short positions (cf. DeMiguel et al., 2009b). Another, less noticed reason for the extreme weights results from the classical normalization scheme, such that the weights sum up to one. This normalization can have a crucial effect on the statistical properties of the weights, since the

[16] As an example, consider a share that costs 100\$ at the time of investment. The price of the put option results as $2 \times 100\$ - 100\$ = 100\$$ at this time. If the price of the share drops to 90\$ in the subsequent time period, investors owning the put option gained 10%, since the new price of the put option results as $2 \times 100\$ - 90\$ = 110\$$. In the very few cases where $p_{i,t} > 2 \times p_i^{\tau}$, we set the price of the put option equal to zero such that the put option is out of the money.

[17] We follow this rule, as it is common practice in the literature to compute the three performance criteria in equations (3.12) - (3.14) for the entire out-of-sample period. However, it is important to note that considering the average for all investment periods would indeed yield identical Sharpe ratios for both normalization strategies.

raw weights are divided by a number that is typically close to zero.

In order to gain insights in the properties of the weight vectors, we report for each strategy and both normalization variants the average minimum and maximum weights for all investment periods as well as the average exposure factor,

$$\bar{\xi_k} = \frac{1}{H} \sum_{\tau=1}^{H} \|\hat{\boldsymbol{w}}_\tau\|_1, \tag{3.16}$$

where $H$ denotes the number of investment periods. We are additionally interested in the fluctuation of weights over the investment periods. Hence, we assess the amount of trading required and compute the following turnover formula for strategies applying the classical short-selling normalization (cf. DeMiguel et al. (2009a) and De Nard et al. (2019),

$$\text{Turnover} = \frac{1}{H-1} \sum_{\tau=1}^{H-1} \sum_{i=1}^{N} (|\hat{w}_{k,i,\tau+1} - \hat{w}_{k,i,\tau^+}|), \tag{3.17}$$

where $w_{k,i,\tau}$ denotes the weight in asset $i$ in the investment period $\tau$ and $w_{k,i,\tau+1}$ is the desired weight in the subsequent investment period $\tau + 1$. The weight $w_{k,i,\tau^+}$ denotes the percentage share of asset $i$ in the portfolio just before rebalancing at $\tau + 1$ and is obtained as[18]

$$\hat{w}_{k,i,\tau^+} = \frac{\hat{s}_{i,\tau} p_i^{\tau^+}}{\sum_{i=1}^{N} \hat{s}_{i,\tau} p_i^{\tau^+}}.$$

The computation of the turnover for put option portfolios requires some modifications. Since the put option strategy essentially involves only positive investments in either stocks or put options, we compute the turnover for $i = 1, ..., 2N$ assets, where the first $N$ assets are stocks and the remaining $N$ assets constitute put options with positive weights. For the $N$ stocks, we apply the turnover formula in (3.17) as set out above. For the put options, we compute the weights $w_{i,\tau^+}$ using the pricing formula in Equation (3.15) and apply the turnover formula in (3.17) for $i = N + 1, ...2N$, accordingly.

### 3.5.4. Performance

In this section, we compare the out-of-sample performance of the alternative portfolio strategies listed in Section 3.5.2 for the 760 stocks of the CRSP dataset and for the 49 industry portfolios. We report Sharpe ratio, portfolio return and standard deviation for each strategy.

[18] This implies that also the naive strategy exhibits a turnover larger than 0. For this strategy, $w_{i,\tau} = w_{i,\tau+1} = 1/N$, but $w_{i,\tau^+}$ may be different due to changes in asset prices between $\tau$ and $\tau + 1$.

### CRSP dataset

Table 3.5 reports for both shorting strategies the out-of-sample Sharpe ratios, portfolio returns and portfolio standard deviations for the CRSP dataset with an estimation window of 1000 trading days and an investment period of 50 trading days. Hence, for this dataset, we examine an investment environment, in which the concentration ratio, $N/T$, amounts to 0.76 and is relatively large. As mentioned before, the poor out-of-sample performance of the plug-in mean-variance estimator (MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$)) is well documented in the academic literature. The results in Table 3.5 confirm this observation and show the devastating effects on performance when considering the sample means in the estimation of the plug-in estimator, resulting from estimation error. The performance improves considerably when omitting the information on the sample means and restricting the expected returns to be identical across assets (GMV($\widehat{\boldsymbol{\Sigma}}$) portfolio) or even when ignoring all information from the data ($1/N$ portfolio). Although the Sharpe ratios of the GMV($\widehat{\boldsymbol{\Sigma}}$) and $1/N$ portfolios are comparable for the short-selling strategy, the GMV($\widehat{\boldsymbol{\Sigma}}$) portfolio is more effective in reducing the portfolio standard deviation. In this scenario, the out-of-sample performance of the GMV($\widehat{\boldsymbol{\Sigma}}$) portfolio can be considerably improved by using effective regularization strategies for the covariance matrix.

The results for the short-selling strategy in Table 3.5 show that the simple regularization strategies CN and 1F attain slightly higher Sharpe ratios than the GMV($\widehat{\boldsymbol{\Sigma}}$) and 1/N portfolios. While the 1F portfolio yields marginal improvements in terms of portfolio return and standard deviation compared to the GMV($\widehat{\boldsymbol{\Sigma}}$) portfolio, the CN strategy exhibits a return and standard deviation that are in the range of the $1/N$ portfolio. The latter is not surprising, since CN ignores the information on the covariances between assets and thus involves only long positions. Put otherwise, the strategy does not exploit differences between assets by taking long and short positions and additionally invests some (small) fraction in each asset. Hence, the CN strategy can be considered as an optimized version of the naive strategy that optimizes with regard to the assets' variances without leveraging.

All other covariance regularization strategies outperform the CN and 1F portfolios by a substantial margin. Interestingly, the LW1F portfolio attains the highest Sharpe ratio and lowest standard deviation among the regularization portfolios for the short-selling strategy. That is, the LW1F portfolio also outperforms the LWNL portfolio, which ranks on the third place in terms of standard deviation. While Ledoit and Wolf (2017a) state that nonlinear shrinkage clearly outperforms linear shrinkage in their out-of-sample analysis, it should be noted that they compare nonlinear shrinkage to linear shrinkage towards (a multiple of) the identity matrix (LWID). However, we examine

the performance of the estimator that shrinks towards the single-factor model, which is a particularly suitable shrinkage target in finance applications. The LWID portfolio, on the other hand, exhibits a Sharpe ratio of 0.70 and a standard deviation of 9.99 in our out-of-sample evaluation (not shown in Table 3.5) and is thus outperformed by the LW1F and LWNL portfolios.

Table 3.5.: Performance measures for the CRSP dataset

| | short-selling | | | put option | | |
|---|---|---|---|---|---|---|
| Strategy | SR | $\mu$ | $\sigma$ | SR | $\mu$ | $\sigma$ |
| MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) | $-0.31$ | -3739 | 12223 | 0.18 | 0.21 | 1.19 |
| GMV($\widehat{\boldsymbol{\Sigma}}$) | 0.55 | 7.38 | 13.32 | 0.67 | 0.79 | 1.18 |
| 1/N | 0.56 | 11.28 | 20.31 | 0.56 | 11.28 | 20.31 |
| | covariance regularization (only GMV) | | | | | |
| CN | 0.61 | 10.65 | 17.53 | 0.61 | 10.65 | 17.53 |
| 1F | 0.59 | 7.71 | 13.07 | 0.69 | 3.90 | 5.64 |
| LW1F | 0.81 | 7.21 | 8.86 | 0.87 | 1.82 | 2.10 |
| LWNL | 0.76 | 7.19 | 9.42 | 0.84 | 1.99 | 2.37 |
| DCCNL | 0.70 | 6.69 | 9.53 | 0.59 | 1.81 | 3.08 |
| | weight constraints | | | | | |
| $\text{GMV}_+$ | 0.79 | 9.06 | 11.53 | 0.79 | 9.06 | 11.53 |
| $\text{MSR}_+$ | 0.65 | 12.10 | 18.60 | 0.65 | 12.10 | 18.60 |
| KML | 0.80 | 7.65 | 9.52 | 0.86 | 3.51 | 4.09 |
| BJL | 0.79 | 7.51 | 9.54 | 0.85 | 3.44 | 4.04 |
| | blocking strategies | | | | | |
| block(eq) | 0.93 | 8.72 | 9.33 | 0.95 | 2.16 | 2.27 |
| block(ord) | 1.02 | 9.93 | 9.69 | 1.01 | 2.40 | 2.37 |
| block(SR) | 1.03 | 10.40 | 10.06 | 1.02 | 2.54 | 2.49 |

The table reports the annualized out-of-sample Sharpe ratios, returns and standard deviations of 15 portfolio strategies applied to the CRSP dataset comprising 760 stocks. The estimation period spans 1000 trading days with rebalancing every 50 days. The out-of-sample period ranges from December 26th. 2003 to November 14th. 2019.

Table 3.5 further shows that the Sharpe ratio of the dynamic conditional correlation model with nonlinear shrinkage (DCCNL) is lower than the Sharpe ratios of the static shrinkage and LASSO portfolios, which mainly stems from the lower portfolio return of this strategy. The portfolio standard deviation is, however, close behind that of the LWNL portfolio and ranks on the fifth place. While it may be restrictive to assume

an unconditional coviariance matrix, it is important to note that the DCCNL model introduces additional complexity and uncertainty with regard to parameter estimation, choice of the GARCH model and distributional assumptions. On the other hand, our study is more suited for a static setting, since we assume 50-day holding periods, which is at odds with the assumption of a constantly changing covariance matrix. Hence, one should take caution when interpreting these results. Furthermore, the combination of the DCC model and nonlinear shrinkage in Engle et al. (2019) enables the estimation of dynamic models for large-dimensional applications in the first place.

Table 3.5 reports results for the estimators involving weight constraints for both, the GMV and MSR portfolio. While the $\text{MSR}_+$ portfolio yields much more sensible results than the plug-in estimator of the MSR portfolio, the Sharpe ratio is only in the midfield due to a comparatively high standard deviation. The Sharpe ratios of the $\text{GMV}_+$ and LASSO portfolios are of a similar magnitude to that of the LW1F strategy and range between 0.79 and 0.81. However, the standard deviation of the $\text{GMV}_+$ portfolio clearly exceeds that of the LASSO and LW-shrinkage strategies. Hence, the portfolio risk can be further reduced by allowing short sales while using some effective regularization strategy.

While the performance of the LASSO and static LW-shrinkage strategies is quite similar, the LASSO portfolios invest in a considerably lower amount of assets. For instance, the KML strategy invests on average in 291 of the 760 shares (see Section 3.5.5 for more details). As mentioned in Section 3.5.2, the amount of eliminated weights depends on the size of the tuning parameter in the L1 penalty term of the LASSO optimization problem. For the results reported in Table 3.5, we used the popular tool of 5-fold cross-validation in order choose the tuning parameter $\lambda$. We examined several other methods to choose $\lambda$ for the LASSO regressions and (the classical) cross-validation belongs to the best-performing approaches. The results of this comparison can be found in Appendix B.2.

The last panel of Table 3.5 reports the outcomes for the blocking strategies using three blocks. The results show that the blocking strategies attain the highest Sharpe ratios among all considered portfolios. Although these strategies simply apply the minimum-variance optimization within the subdivided blocks of assets, the portfolio standard deviations are well below the standard deviation of the plug-in GMV portfolio. The equal-weighted blocking strategy (block(eq)) attains a standard deviation of 9.33, which is second lowest among all considered portfolios for the short-selling strategy. Hence, reducing the concentration ratio via subdividing the assets into blocks seems to be an effective tool for improving portfolio performance. As mentioned in Section 3.5.2, the three blocking strategies differ with regard to the weighting of the individual

blocks. Since the assets are sorted according to their sample means before the blocks are formed, the ordinal-weighted blocking strategy (block(ord)) and typically the Sharpe ratio weighted strategy (block(SR)) attach the largest weight to the block comprising the best-performing assets. Interestingly, this increases the portfolio return and Sharpe ratio of the blocking strategy while the data-driven weighting scheme yields the best performance with regard to these two criteria.

We argued in Section 3.4.3 that the normalization of the (raw) weights has a major influence on the statistical properties of the portfolios. The short-selling strategy typically involves high leverage, in particular in high-dimensional asset applications, which results in larger portfolio means and variances. The three columns on the right-hand side of Table 3.5 report the performance results for the put option strategy. Obviously, there are no differences for portfolios involving only long positions (e.g. 1/N, CN, $\text{GMV}_+$ and $\text{MSR}_+$). On the other hand, (absolute) return and volatility are considerably reduced for all other portfolios (e.g. portfolios that exhibit negative weights for the short-selling strategy). This is most striking for the $\text{MSR}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ portfolio, which now yields a small, but positive, portfolio return. The standard deviation of the $\text{GMV}(\widehat{\boldsymbol{\Sigma}})$ portfolio is reduced by a factor of 11, while the portfolio return is roughly 9 times smaller. For the other strategies involving put options, the reduction is less extreme but still sizable.

In most cases, the standard deviation decreases by a larger factor than the portfolio return when applying the put option strategy, which results in an increased Sharpe ratio. For instance, the Sharpe ratio increases by almost 21% for the $\text{GMV}(\widehat{\boldsymbol{\Sigma}})$ portfolio, by nearly 10% for the LWNL strategy and by roughly 7% and 8% for the KML and BJL* strategies, respectively. It turns from negative to positive for the $\text{MSR}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ portfolio. The only exceptions are the block(ord) and block(SR) portfolios, for which the Sharpe ratios are slightly reduced and the DCCNL strategy where it decreases by almost 17%. As mentioned in Section 3.5.3, the Sharpe ratios of both normalization schemes differ in case the sum of the raw weights is negative for the short-selling strategy, which occurs, however, rarely for GMV portfolios. More importantly, the adjustment factor of the portfolio mean and standard deviation changes from one investment period to the next and thus differs when the performance measures are computed for the entire out-of-sample period.[19]

With regard to the Sharpe ratio, there are only a few differences in the ranking of the portfolios between the two shorting strategies. Most notably, the $\text{GMV}(\widehat{\boldsymbol{\Sigma}})$ portfolio improves and the DCCNL deteriorates. There are, however, important differences when

[19] In addition, the maximal loss of investing in a put option is 100% (the investment) while the potential loss is infinity for short-selling a share (if the price increases to infinity).

considering the portfolio return and standard deviation individually: While these measures are in a similar range for the LASSO and LW-regularization approaches applying the short-selling strategy, they are nearly twice as high for the LASSO-approaches applying the put option strategy. In the latter case, the blocking portfolios feature returns and standard deviations that are above LW1F but below the LASSO strategies. Interestingly, the plug-in portfolios exhibit the lowest volatility but also very low returns. Hence, for portfolios with excessive leverage and exposure rates, the put option strategy tends to overly hedge risk, leading to a substantial reduction in return. Conversely, less-leveraged strategies, such as LASSO, yield a considerably higher return and standard deviation.

#### 49 industry portfolios

In this section, we examine the out-of-sample performance for the 49 industry portfolios. As for the CRSP dataset, we consider an estimation period of 1000 trading days and an investment period of 50 trading days, which leads to a completely altered estimation setting for this dataset, in which the concentration ratio amounts to approximately 0.05. Table 3.6 reports the results for the two alternative normalization schemes.

The first thing to notice is that the absolute performance significantly improves in comparison to the CRSP dataset, since Sharpe ratios and portfolio returns are (much) higher and standard deviations are lower for all considered portfolios. There are two reasons for this: Firstly, we consider the industry portfolios as individual assets in our optimization and, as pointed out by Ledoit and Wolf (2017a), portfolios bear a lower risk than individual shares that we considered in the CRSP dataset. Secondly, the out-of-sample period we observe for the industry portfolios is considerably longer, starting in January 1973, which makes a clear difference. When adjusting the out-of-sample period for both datasets, the absolute performance is much more similar (see Appendix B.1). In addition to the enhanced overall performance for the industry portfolio dataset, there are also some differences with regard to the relative performance of the various strategies.

Table 3.6.: Performance measures for the 49 industry portfolios

| | short-selling | | | put option | | |
|---|---|---|---|---|---|---|
| Strategy | SR | $\mu$ | $\sigma$ | SR | $\mu$ | $\sigma$ |
| MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) | −0.13 | −63.31 | 497.45 | 2.30 | 4.72 | 2.05 |
| GMV($\widehat{\boldsymbol{\Sigma}}$) | 2.05 | 16.20 | 7.90 | 2.29 | 5.55 | 2.42 |
| 1/N | 1.44 | 20.89 | 14.52 | 1.44 | 20.89 | 14.52 |
| | covariance regularization (only GMV) | | | | | |
| CN | 1.53 | 20.62 | 13.45 | 1.53 | 20.62 | 13.45 |
| 1F | 1.72 | 16.03 | 9.30 | 1.92 | 6.40 | 3.33 |
| LW1F | 2.06 | 16.22 | 7.88 | 2.30 | 5.75 | 2.50 |
| LWNL | 2.08 | 16.38 | 7.88 | 2.33 | 5.83 | 2.51 |
| DCCNL | 2.14 | 16.93 | 7.91 | 2.46 | 6.18 | 2.52 |
| | weight constraints | | | | | |
| $\text{GMV}_+$ | 1.71 | 17.61 | 10.27 | 1.71 | 17.61 | 10.27 |
| $\text{MSR}_+$ | 1.73 | 21.89 | 12.68 | 1.73 | 21.89 | 12.68 |
| KML | 2.05 | 16.24 | 7.90 | 2.24 | 6.28 | 2.80 |
| BJL* | 2.26 | 20.29 | 8.97 | 2.49 | 9.80 | 3.94 |
| | blocking strategies | | | | | |
| block(eq) | 2.15 | 19.11 | 8.89 | 2.33 | 8.98 | 3.85 |
| block(ord) | 2.23 | 19.93 | 8.93 | 2.44 | 9.37 | 3.84 |
| block(SR) | 2.27 | 20.07 | 8.85 | 2.49 | 9.59 | 3.84 |

The table reports the annualized out-of-sample Sharpe ratios, returns and standard deviations of 15 portfolio strategies applied to the 49 industry portfolios. The estimation period spans 1000 trading days with rebalancing every 50 days. The out-of-sample period ranges from June 18th. 1973 - January 15th. 2020.

For the short-selling strategy, the out-of-sample performance of the MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) portfolio is poor with a negative Sharpe ratio. As for the CRSP dataset, the performance improves considerably when excluding the information on the sample means (GMV($\widehat{\boldsymbol{\Sigma}}$)) or when the information from the data is completely ignored (1/N). In contrast to the results for the CRSP dataset, the GMV($\widehat{\boldsymbol{\Sigma}}$) portfolio clearly dominates the 1/N, simple regularization (CN and 1F) and short sale constrained ($\text{GMV}_+$ and $\text{MSR}_+$) strategies in terms of Sharpe ratio and is comparable to the KML and LW1F portfolios in this regard.

Its standard deviation ranks on the second place together with the KML strategy, closely behind the LW1F and LWNL portfolios. Hence, the volatility of the plug-in GMV estimator can hardly be reduced by using regularization strategies in this low concentration ratio setting.

The ranking of the (L1-)regularization strategies according to the standard deviation largely corresponds to that for the CRSP dataset, although the distances between the low-volatility strategies (LW1F, LWNL, DCCNL, KML) are considerably smaller. These sophisticated shrinkage strategies clearly outperform the simple regularization strategies CN and 1F. There are, however, marked differences concerning the Sharpe ratio ranking. Most notably, the BJL$^*$ strategy attains the highest Sharpe ratio among the regularization strategies. Considering the fact that BJL$^*$ applies L1-regularization to the weights of the MSR portfolio, the strategy successfully enhances Sharpe ratio and return as compared to the best performing GMV-estimators in this scenario. A second difference is that the Sharpe ratio of the $\text{GMV}_+$ portfolio clearly ranks behind the LW and LASSO approaches and even falls behind the simple 1F strategy. Lastly, the DCCNL and LWNL portfolios yield somewhat higher Sharpe ratios than the LW1F strategy, which is, however, driven by (slightly) larger portfolio returns.

Together with BJL$^*$, the blocking portfolios attain the highest Sharpe ratios among all considered strategies. As for the CRSP dataset, attaching the largest weight to the block comprising the assets with the highest historical sample means improves the Sharpe ratio and portfolio return for the blocking strategy. All three performance measures are quite similar for the block(SR) and BJL$^*$ portfolios. Only the naive and CN strategies achieve a somewhat larger return, which is accompanied by a considerably higher volatility.[20] Even though the blocking strategies are among the best in terms of Sharpe ratio and return, their standard deviations exceed that of the $\text{GMV}(\widehat{\boldsymbol{\Sigma}})$, LW and LASSO strategies.

The three columns on the right-hand side of Table 3.6 report the performance results for the industry portfolios when the put option strategy is applied. As expected, the portfolios involving put options yield returns and standard deviations that are substantially lower in comparison to the short-selling portfolios. The Sharpe ratios of all portfolios investing in put options improves, since the standard deviations decrease by a larger factor than the portfolio returns. In comparison to the CRSP dataset, the inflation factors for the alternative strategies are more similar and range between two and three. The $\text{MSR}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ portfolio stands out as an exception, since it attains a competitive Sharpe ratio of 2.30 for the put option strategy. It exhibits the lowest standard

[20] As for the CRSP dataset, the 1/N and CN strategies show a similar performance with slight advantages for the CN portfolio.

deviation among all considered portfolios, reduced by over 240 times as compared to the short-selling version. The GMV($\widehat{\boldsymbol{\Sigma}}$) portfolio yields a similar Sharpe ratio and exhibits the second-lowest standard deviation.

In this setting involving a low concentration ratio, none of the considered covariance regularization approaches for estimating the GMV portfolio is able to reduce the volatility for the put option strategy as compared to the plug in estimator and only the DCCNL portfolio achieves a noticeably higher Sharpe ratio of 2.46. In line with the results for the short-selling strategy, the BJL* and block(SR) strategies attain the highest Sharpe ratios among all considered portfolios and exhibit a comparable performance. The BJL* and blocking strategies have higher standard deviations than the LW and KML approaches, which are, however, accompanied by larger portfolio returns.

### 3.5.5. Analysis of weights

In this section, we examine the out-of-sample weight properties for both shorting strategies across the two datasets. As mentioned in Section 3.5.3, we report the turnover, average minimum and maximum weights as well as the average exposure factor for the alternative portfolios.

#### CRSP dataset

Table 3.7 reports the results for the 760 stocks of the CRSP dataset. For the short-selling strategy, the MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) portfolio is extreme by any measure and suggests to invest 882 times of one's (investible) wealth into stocks. The widely divergent weights result from the substantial estimation error that is involved when including the sample means in the optimization without imposing any restrictions, since it produces large absolute weights in order to optimally exploit the assets' differences. However, as noted in Section 3.4.2, it is not uncommon to obtain weights with the wrong sign. Ignoring the sample means effectively reduces turnover, exposure and the weight dispersion, yet adhering to the GMV($\widehat{\boldsymbol{\Sigma}}$) still requires to invest more than 11 times of one's wealth and produces the second highest turnover exceeding that of the LASSO approaches by nearly 6 times.

It is not surprising that the naive portfolio features a low turnover rate as typically little trading is required to meet the $1/N$ rule at the beginning of a new investment period. It is perhaps more surprising that there exists a portfolio (CN) that exhibits lower turnover than the naive strategy. As stated earlier in Section 3.5.4, the CN strategy can be interpreted as an optimized version of the naive strategy with similar properties, since a small (but positive) fraction is invested in every asset. The weight range of

$\text{GMV}_+$ and $\text{MSR}_+$ exceeds that of other portfolios that allow only nonnegative weights and display a somewhat higher turnover. In contrast to $1/N$ and CN, the short sale constrained optimization actually sets weights to zero and thus eliminates assets from the portfolio. Portfolios without short-selling share the common feature that their exposure is limited to one.

Table 3.7.: Analysis of weights for the CRSP dataset

| | short-selling | | | | put option | | | |
|---|---|---|---|---|---|---|---|---|
| Strategy | turn | Min | Max | $\xi$ | turn | Min | Max | $\xi$ |
| $\text{MSR}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ | 2444 | -768 | 750 | 882 | 0.65 | −0.86 | 0.82 | 1.00 |
| $\text{GMV}(\widehat{\boldsymbol{\Sigma}})$ | 6.76 | −8.97 | 12.10 | 11.43 | 0.59 | −0.79 | 1.05 | 1.00 |
| 1/N | 0.10 | 0.13 | 0.13 | 1.00 | 0.10 | 0.13 | 0.13 | 1.00 |
| covariance regularization (GMV) | | | | | | | | |
| CN | 0.09 | 0.01 | 0.58 | 1.00 | 0.09 | 0.01 | 0.58 | 1.00 |
| 1F | 0.30 | −0.75 | 2.37 | 2.22 | 0.13 | −0.33 | 1.06 | 1.00 |
| LW1F | 1.16 | −2.69 | 6.26 | 4.28 | 0.27 | −0.62 | 1.46 | 1.00 |
| LWNL | 1.25 | −2.16 | 2.98 | 4.05 | 0.31 | −0.53 | 0.74 | 1.00 |
| DCCNL | 3.16 | −1.95 | 7.38 | 3.49 | 0.90 | −0.55 | 2.15 | 1.00 |
| weight constraints | | | | | | | | |
| $\text{GMV}_+$ | 0.27 | 0.00 | 13.29 | 1.00 | 0.27 | 0.00 | 13.29 | 1.00 |
| $\text{MSR}_+$ | 0.57 | 0.00 | 13.02 | 1.00 | 0.57 | 0.00 | 13.02 | 1.00 |
| KML | 1.14 | −4.24 | 9.93 | 2.44 | 0.47 | −1.71 | 4.14 | 1.00 |
| BJL | 1.16 | −4.28 | 9.85 | 2.48 | 0.47 | −1.70 | 4.05 | 1.00 |
| blocking strategies | | | | | | | | |
| block(eq) | 2.50 | −3.10 | 5.22 | 4.32 | 0.59 | −0.70 | 1.25 | 1.00 |
| block(ord) | 2.63 | −3.63 | 5.88 | 4.31 | 0.62 | −0.86 | 1.44 | 1.00 |
| block(SR) | 2.71 | −4.29 | 6.96 | 4.28 | 0.65 | −1.04 | 1.72 | 1.00 |

The table reports the turnover, average minimum and maximum weights and the average exposure across all investment periods for 15 portfolio strategies applied to the CRSP dataset comprising 760 stocks. The estimation period spans 1000 trading days with rebalancing every 50 days. The out-of-sample period ranges from December 26th. 2003 to November 14th. 2019.

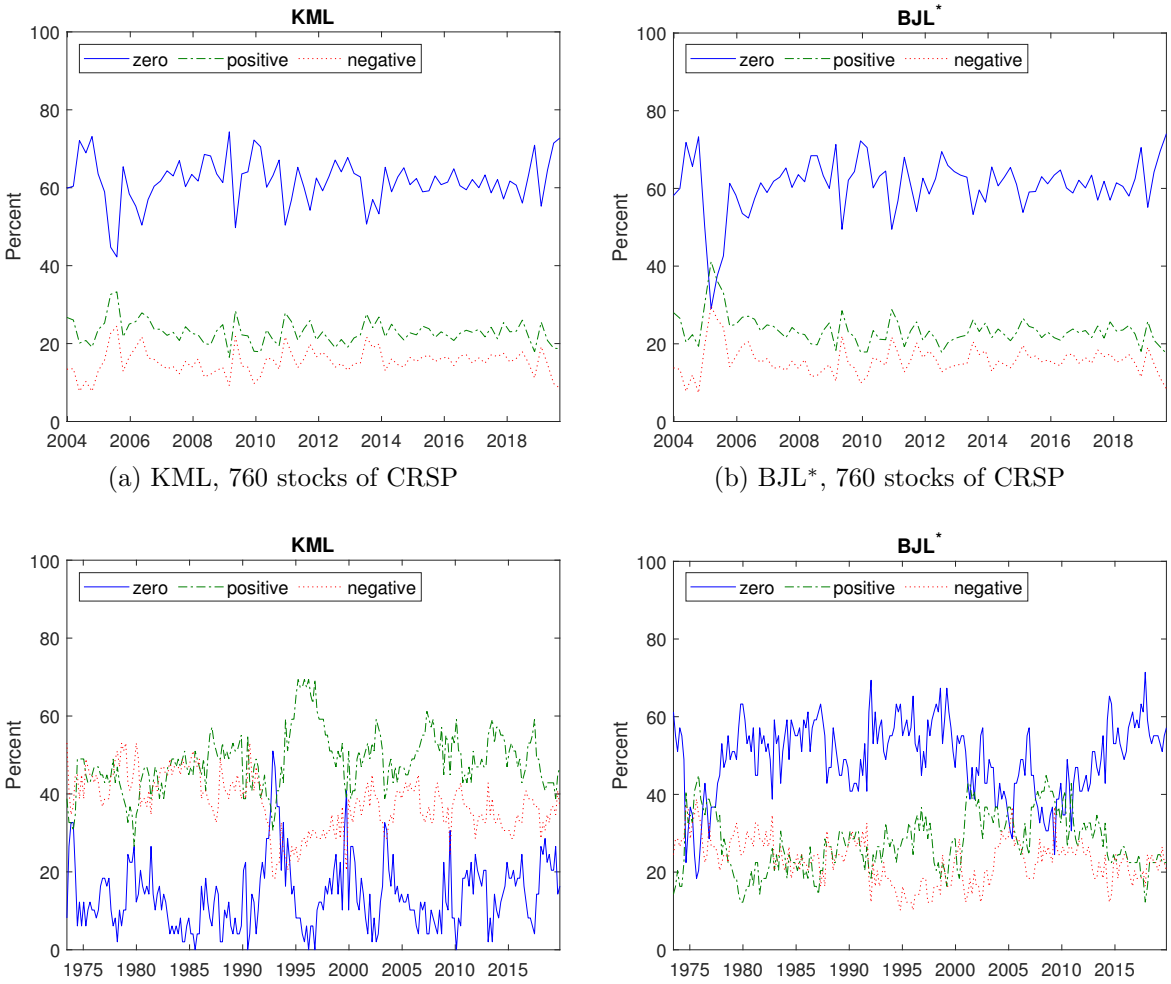
Among the portfolios that allow short-selling, the 1F portfolio attains the lowest turnover and exposure rate. The turnover for the LASSO and static Ledoit-Wolf approaches is of similar magnitude and improves substantially as compared to the $\text{GMV}(\widehat{\boldsymbol{\Sigma}})$ portfolio. There are, however, important differences between the LASSO and static LW portfolios: The L1-regularization eliminates weights from the portfolio and thus limits exposure subject to the size of the tuning parameter $\lambda$. As mentioned earlier, we use

cross-validation in order to select $\lambda$ for each investment period. Figures 3.6a and 3.6b show the percentage shares of zero, positive and negative weights throughout the out-of-sample period for KML and BJL*, respectively. In both cases, the average proportion of eliminated weights exceeds 60% while the number of weights with a negative sign rarely exceeds 20%. These smaller shares of negative weights for the short-selling strategy lead to a considerable reduction in the exposure as compared to the LW strategies. While the exposure of the LASSO portfolios is below 2.5, the LW1F and LWNL portfolios require to invest more than four times of one's wealth on average. The exposure of the DCCNL is slightly below that of the static LW strategies, but the turnover is the highest among all regularization strategies and about half the turnover of the GMV($\widehat{\mathbf{\Sigma}}$) strategy. The blocking strategies require somewhat less trading than DCCNL, but clearly rank behind the static LW and LASSO approaches.

The four columns on the right-hand side of Table 3.7 report the weight measures for the put option strategy. By definition, the exposure rate of the put option strategy corresponds to one, since this strategy requires a normalization such that $\sum |w_i| = 1$. As mentioned in Section 3.3, the intuitive explanation is that a put option shifts the minus sign of the weight to the asset's return and thus implies a positive investment in an asset with an inverse return without leverage-effect. Hence, betting on a falling stock price does not require to incur debts by borrowing a share as for the short-selling strategy. This reduces the weight range and substantially lowers the turnover for portfolios that apply the put option strategy. Therefore, the gap between the highest and lowest turnover rates is much smaller. The naive and CN strategies still attain the lowest turnover rates directly before the 1F strategy. They precede the $\text{GMV}_+$ and static LW approaches, which now yield comparable turnover rates, followed by the LASSO strategies. The decline in turnover is most pronounced for the plug-in estimators, which have a similar turnover as the blocking portfolios. Consistent with the results for short-selling strategy, the DCCNL yields a comparatively high turnover.

#### 49 industry portfolios

Table 3.8 reports the weight measures for the 49 industry portfolios. While the turnover rates of all portfolios are considerably lower in comparison to the CRSP dataset, the ranking of the strategies is similar to that in Table 3.7 with one important exception: The GMV($\widehat{\mathbf{\Sigma}}$) portfolio exhibits a turnover that is now in the range of the KML and static LW strategies. This matches the insights from the performance analysis in Section 3.5.4 indicating that the regularization strategies bring only small improvements in this low-concentration ratio setting.




(a) KML, 760 stocks of CRSP

(b) BJL*, 760 stocks of CRSP

(c) KML, 49 industry portfolios

(d) BJL*, 49 industry portfolios

Figure 3.6.: LASSO weights

The figure displays the percentage of zero, positive and negative weights for the LASSO approaches across the out-of-sample time periods for both the CRSP dataset and the 49 industry portfolios.

Table 3.8.: Analysis of weights for the 49 industry portfolios

| | short-selling | | | | put option | | | |
|---|---|---|---|---|---|---|---|---|
| Strategy | turn | Min | Max | $\xi$ | turn | Min | Max | $\xi$ |
| MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) | 15.85 | -195 | 199 | 21.35 | 0.26 | −7.09 | 13.15 | 1.00 |
| GMV($\widehat{\boldsymbol{\Sigma}}$) | 0.51 | −18.50 | 67.43 | 3.30 | 0.16 | −5.51 | 21.43 | 1.00 |
| 1/N | 0.05 | 2.04 | 2.04 | 1.00 | 0.05 | 2.04 | 2.04 | 1.00 |
| | | | | covariance regularization (GMV) | | | | |
| CN | 0.05 | 0.30 | 6.71 | 1.00 | 0.05 | 0.30 | 6.71 | 1.00 |
| 1F | 0.34 | −16.34 | 43.18 | 2.83 | 0.11 | −5.65 | 15.75 | 1.00 |
| LW1F | 0.47 | −17.25 | 63.49 | 3.18 | 0.15 | −5.34 | 20.89 | 1.00 |
| LWNL | 0.48 | −17.14 | 62.85 | 3.18 | 0.16 | −5.32 | 20.71 | 1.00 |
| DCCNL | 2.72 | −15.80 | 55.69 | 3.14 | 0.86 | −4.91 | 18.85 | 1.00 |
| | | | | weight constraints | | | | |
| $\text{GMV}_+$ | 0.11 | 0.00 | 59.41 | 1.00 | 0.11 | 0.00 | 59.41 | 1.00 |
| $\text{MSR}_+$ | 0.34 | 0.00 | 47.36 | 1.00 | 0.34 | 0.00 | 47.36 | 1.00 |
| KML | 0.49 | −17.22 | 68.84 | 2.92 | 0.18 | −5.83 | 25.15 | 1.00 |
| BJL* | 0.61 | −18.97 | 85.10 | 2.33 | 0.25 | −7.97 | 38.51 | 1.00 |
| | | | | blocking strategies | | | | |
| block(eq) | 0.85 | −11.95 | 32.98 | 2.40 | 0.35 | −4.73 | 14.47 | 1.00 |
| block(ord) | 0.91 | −15.34 | 37.72 | 2.42 | 0.37 | −6.03 | 15.96 | 1.00 |
| block(SR) | 0.91 | −15.93 | 41.55 | 2.41 | 0.38 | −6.46 | 18.27 | 1.00 |

The table reports the turnover, average minimum and maximum weights and the average exposure across all investment periods for 15 portfolio strategies applied to the 49 industry portfolios. The estimation period spans 1000 trading days with rebalancing every 50 days. The out-of-sample period ranges from June 18th. 1973 - January 15th. 2020.

It is not surprising that the average minimum and maximum weights are larger in absolute values, since the investible wealth is distributed among fewer assets. Interestingly, the BJL* strategy features the widest weight range behind the MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) portfolio, but yields the lowest exposure rate among the strategies that involve short sales. This results from the fact that the BJL* estimator effectively limits short sales and sets nearly 50% of the 49 weights to zero (Figure 3.6d). In contrast, the KML estimator eliminates on average only 14% of the assets (Figure 3.6c) and in fact has a larger exposure than for

the CRSP dataset. As depicted in Figure 3.6, the proportion of zero weights is smaller for the 49 industry portfolios as compared to the CRSP dataset, which is particularly pronounced for the KML strategy. Since the exposure rate decreases for nearly all portfolios in comparison to the CRSP dataset, the KML now ranks in the midfield behind the blocking strategies but ahead of the LW portfolios.

In line with the results for the CRSP dataset, the turnover rates reduce substantially when the put option strategy is applied. This does, however, barely affect the ranking with regard to the turnover, except for the fact that the MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) portfolio moves upwards before the blocking and DCCNL portfolios. Since the exposure rate is constrained to equal one for all strategies that invest in put options, the minimum and maximum weights are smaller in absolute values. Therefore, the put-option strategy effectively reduces turnover by limiting the exposure even in this environment with a low-concentration ratio.

## 3.6. Conclusion

In this paper, we consider a variety of aspects related to empirical challenges for portfolio selection. The popular MSR portfolio requires reliable estimates of expected returns and the covariance matrix, as estimation errors in both moments can lead to devastating effects on out-of-sample performance. The plug-in estimator based on the sample moments yields extreme portfolio weights that perform poorly out-of-sample and involve excessive leverage and turnover. Possible solutions are either to regularize the estimated weights of the MSR portfolios (e.g. short sale constraints, L1-regularization) or to ignore the information on the first moment and estimate the GMV portfolio. The latter solution solely requires to estimate the covariance matrix of returns. For the dataset comprising 49 industry portfolios, where $T >> N$, the plug in estimator of the GMV portfolio attains competitive out-of-sample results. However, obtaining reasonable estimates is more challenging in case $N$ is considerable relative to $T$, as for the 760 CRSP stocks, and requires, for instance, effective regularization approaches for the covariance matrix or the weights.

In our empirical application, we compare several such strategies. We observe that the considered LASSO and static LW-shrinkage approaches attain a quite similar out-of-sample performance when estimating the GMV portfolio for the CRSP dataset and clearly improve upon the simple regularization estimators. Perhaps surprisingly, we find that it is difficult to outperform the LW1F strategy with other LW-type estimators. This relatively simple estimator shrinks the sample covariance matrix towards the single-

factor model, which is a particularly suitable shrinkage target in finance applications. Estimating the LASSO variant of the KM-regression yields a similar performance as the LW1F strategy, but features an important advantage: The L1-regularization eliminates weights from the portfolio and significantly reduces the exposure factor. In addition, applying L1-regularization to the normalized version of the BJ-regression yields estimated weights for the MSR portfolio that yield a comparable performance for the CRSP dataset and improve upon the performance of KML and LW1F for the 49 industry portfolios in terms of portfolio return and Sharpe ratio.

In addition to regularization, a further approach that helps to improve the portfolio performance is to artificially reduce the concentration ratio for the considered dataset by forming blocks of assets and compute the plug-in GMV estimator for each individual block. Despite sacrificing some data information through the use of blocking, the performance appears to benefit from the reduced concentration ratio as these strategies belong to the top-performing portfolios for both of our applications. The assets may be assigned systematically, e.g. according to their sample mean. While the direct usage of sample means for weight estimation is problematic due to large estimation errors, the indirect usage of the sample means for sorting the assets into blocks can be useful, since mistakes (e.g. assigning an asset to the wrong block) are far less severe. We observe that attaching larger weights to the blocks comprising the best-performing assets increases the portfolio return and Sharpe ratio in our applications.

A further empirical challenge is to cope with the many negative estimates of the weights that result when applying the classical short-selling strategy that lead to excessive turnover, leverage and exposure. One possibility is to fully exclude negative weights from the optimization and estimate short sale constrained portfolios for either the MSR or GMV portfolio. The portfolio risk can, however, be further reduced by using some effective regularization strategies while allowing (some) short sales. Another option is to apply a suitable put option strategy, which entails a new normalization scheme, where the sum of the absolute weights equals one. Such a strategy leads to a substantial reduction in portfolio risk, turnover and extreme weights, but also to a noticeable decline in portfolio return. In contrast to short sale bans, this strategy allows to bet on falling stock prices while keeping the exposure fixed at one.

# Chapter 4.

# Quantifying Downside Risk: A comparative Study of Value at Risk and Expected Shortfall

## 4.1. Introduction

Value at Risk (VaR) and Expected Shortfall (ES) are two widely employed downside risk measures that play an important role in determining regulatory capital requirements for trading books of banks. In simple terms, VaR denotes a threshold for losses, such as for a portfolio of assets, which is not exceeded with a high probability (e.g., 99%). On the other hand, ES represents the expected loss in case the actual loss exceeds the VaR threshold (see Section 4.2.1 for detailed definitions). This implies that ES exceeds VaR for the same level of confidence.

The Basel III Accords stipulate that ES at the 97.5% confidence level replaces VaR at the 99% confidence level as the risk measure for capital requirement calculations. This transition is motivated by the fact that the ES captures tail risks, in contrast to VaR (BCBS, 2016, 2019). For instance, the VaR at the 99% confidence level provides no information about the magnitude of extreme losses with a probability of less than 1%. Furthermore, it is often argued that ES is theoretically superior, since it fulfills the mathematical axiom of subadditivity and is hence a coherent risk measure, unlike VaR (Artzner et al., 1999). On the other hand, ES lacks the property of elicitability, which led to a debate whether ES is backtestable (Gneiting, 2011). In fact, VaR is elicitable and continues to be used as a risk measure for conducting backtests under Basel III.

Despite the theoretical differences between VaR and ES, both risk measures have to be estimated in practice. ES is estimated with more uncertainty than VaR at the same confidence level. In addition, previous studies find that the estimation of ES is less robust with regard to model misspecifications and noise in the data (see, e.g., Cont et al., 2010;

Kou et al., 2013; Kellner and Rösch, 2016). This leaves the question whether the more complex estimation of ES provides any additional insights on the riskiness of a portfolio as compared to the estimation of VaR in practical applications.

Under certain distributional assumptions, there exists a relationship between ES and VaR. For the normal distribution this relationship depends only on the confidence level and is otherwise constant. Therefore, if the data follows a normal distribution, ES does not provide any additional information compared to VaR, since VaR can be easily transformed into ES by multiplying it with the corresponding constant. However, financial market data often exhibit fat tails and assuming a normal distribution is unrealistic in most cases. The Student-t distribution, on the other hand, is popular for modeling financial market data. For this distribution, the ratio between ES and VaR depends on the degrees of freedom and thus on the heaviness of the tails in addition to the confidence level. The relationship between ES and VaR can be used to construct a simple estimator for ES. For this purpose, VaR is estimated using some established estimation method and is then multiplied by a ratio ES/VaR. By scaling the VaR upwards, this method relies more on the accuracy of the VaR estimation rather than on individual data points in the tail, as commonly employed methods for ES estimation.

This paper presents three main contributions. First, the performance of established estimation models for VaR and ES is examined in simulations and in a real-data application. While VaR is estimated more accurately than ES at the same level of confidence, lower confidence levels enhance the estimation accuracy of risk measures. Therefore, switching from 99% VaR to 97.5% ES could potentially offset this advantage of VaR. Thus, comparing the two risk measures at these regulatory relevant confidence levels is particularly interesting. Second, the objective is to explore whether simple ratio models can achieve comparable performance for ES estimation of stock data in comparison to established models. In simple ratio models, the VaR of different stocks is multiplied by the same constant to obtain ES estimates. This applies, for example, to the ES/VaR ratio for normally distributed data. Alternatively, the simple ratio can be set higher or estimated from the data to obtain a more realistic (and thus higher) ratio for stock data. Third, the performance of ratio models is examined, where the ratio incorporates additional information about the stocks, such as the industry affiliation of a company or, as for the Student-t ratio, the heaviness of the tail.

Comparing alternative estimation methods for real data is not straightforward, since the true values of risk measures are unknown. Therefore, the first application employs a bootstrap resampling procedure to generate new samples from a filtered dataset consisting of 760 stocks. This enables a comparison of estimation methods in terms of bias,

variance and MSE. In a second application, the performance is analyzed using the joint scoring function of Fissler and Ziegel (2016). The benchmark models that are considered in these applications are (filtered) historical simulation and two models based on extreme value theory.

The remainder of the paper is structured as follows. Section 4.2 provides a literature review discussing the key theoretical differences between VaR and ES and their practical implications for stock data. Section 4.3 derives the relationship between ES and VaR for the normal and Student-t distribution and investigates the performance of the t-ratio in a simulation. Section 4.4 presents the results of the bootstrap resampling procedure and Section 4.5 shows the results of the performance evaluation using the joint scoring function. Section 4.6 concludes.

## 4.2. Definitions and Properies of VaR and ES

### 4.2.1. VaR and ES

The formulations of VaR and ES in the academic literature differ and depend on whether the return variable $R$ or the loss variable $L$ is considered for some financial asset and whether *high* or *low-level* terminology is used for the confidence or probability levels of these risk measures. In this paper, we use the more common notation and consider the random loss variable $L$ of a financial position defined on a probability space $(\Omega, \mathcal{F}, P)$ and high-level terminology for the confidence level $\alpha \in (0, 1)$ with values close to 1. Note that $L = -R$ so that losses are positive numbers in this setting. A risk measure $p$ (e.g. VaR or ES) is a functional defined on a set of random variables $\mathcal{L}$ that maps a random variable $L \in \mathcal{L}$ into the real numbers $\mathbb{R}$. Thus, $p(L)$ reports a level of risk that can be used to determine the capital amount to back a position with loss $L$.

Both risk measures are *law-invariant* meaning that two random variables $L_1$ and $L_2$ with the same distribution $F_{L_1} = F_{L_2}$ yield the same risk measure value $p(L_1) = p(L_2)$. A law-invariant risk measure depends on the random variable $L \in \mathcal{L}$ only through its distribution $F_L$ and can thus be interpreted as a statistical functional defined on a space of distribution functions $\mathcal{P}$ that maps $F_L$ to the real line. We will use $p$ for both, $p(L)$ and $p(F_L)$, to simplify notation.

For an asset with loss distribution $F_L$, the VaR denotes a loss threshold that is not exceeded with probability $\alpha$ over some period of time. More formally, the VaR at confidence level $\alpha$ is given by

$$\mathrm{VaR}_\alpha(L) = \inf\{l \in \mathbb{R} : P(L \leq l) \geq \alpha\} \tag{4.1}$$

and is thus the $\alpha$-quantile of the loss distribution, $q_\alpha(L)$. The ES denotes the expected loss in case the VaR is exceeded. For a continuous loss distribution, the ES is given by the conditional expected value above the VaR,[1]

$$\mathrm{ES}_\alpha(L) = \mathbb{E}[L|L \geq VaR_\alpha(L) = q_\alpha(L)]. \tag{4.2}$$

There is a vast literature on the desirable properties of risk measures and, in this context, on the theoretical pros and cons of VaR and ES in terms of these properties. The debate revolves primarily around the properties coherence and elicitability with the ES satisfying coherence but not being elicitable and, vice versa, the VaR fulfilling elicitability but not coherence due to its lack of subadditivity. While these theoretical differences have been widely discussed, their practical implications are less well understood, in particular with regard to the transition of 99% VaR to 97.5% ES in Basel III. This section aims to provide important definitions and highlight theoretical differences between VaR and ES.

### 4.2.2. Coherence, subadditivity and fat tails

In their seminal paper, Artzner et al. (1999) recommend the use of *coherent risk measures* for effective risk management. They are defined as follows.

**Definition 4.2.1 (Coherent risk measures)** *A risk measure $p(\cdot)$ is coherent if it satisfies*

1. *Monotonicity: For all $L_1, L_2 \in \mathcal{L}$ it holds that*
   $L_1 \leq L_2 \Rightarrow p(L_1) \leq p(L_2)$.
2. *Positive homogeneity: For all $L \in \mathcal{L}$ and any $\lambda \geq 0$ it holds that*
   $p(\lambda L) = \lambda p(L)$.
3. *Translation invariance: For all $L \in \mathcal{L}$ and every $c \in \mathbb{R}$ it holds that*
   $p(L - c) = p(L) - c$.
4. *Subadditivity: For all $L_1, L_2 \in \mathcal{L}$ it holds that*
   $p(L_1 + L_2) \leq p(L_1) + p(L_2)$.

Following the work of Artzner et al. (1999), other papers published around the turn of the millennium particularly emphasize the importance of coherence (Acerbi and Tasche, 2002; Tasche, 2002). For example, Acerbi and Tasche (2002, p.380) claim that "To avoid

[1] See, for instance, Emmer et al. (2015) for a more general definition.

confusion, if a measure is not coherent, we just choose not to call it a risk measure at all." As both risk measures, VaR and ES, meet the first three conditions of coherence, the discussion regarding the advantages and disadvantages of VaR and ES mainly focuses on the subadditivity property. Artzner et al. (1999) demonstrate that ES is subadditive, while VaR does not generally have this property.

Subadditivity ensures that the total risk of a portfolio composed of multiple assets cannot exceed the sum of the risks of its individual assets. Put simply, the subadditivity property reflects the idea that diversification reduces risk. There exist several examples in the literature that show that VaR is not generally subadditive (see e.g. Artzner et al. 1999 or Daníelsson et al. 2013). These examples, however, represent relatively extreme situations that do not necessarily match the characteristics of nonderivative financial assets. Hence, the question remains whether the general lack of subadditivity, as a theoretical deficiency, excludes VaR as a risk measure for portfolios and stocks in practice.

There are several studies that examine the conditions under which VaR is subadditive. Although Artzner et al. (1999) show that VaR is subadditive for normally distributed data in the relevant tail region (i.e. when $\alpha > 0.5$ ), this is not a practical scenario for financial market data, which typically displays fat tails (see, e.g., Mandelbrot, 1963; Fama, 1965). However, further studies reveal that the VaR is also subadditive for many practical applications with financial market data, as long as the tails of the corresponding loss distribution are not *super fat* (cf. Garcia et al., 2007; Ibragimov and Walden, 2007; Ibragimov, 2009; Daníelsson et al., 2013). *Super fat tails*, as stated by Daníelsson et al. (2013), refers to a scenario where the first moment of the distribution does not exist, such as for the Cauchy distribution.

To better understand when a distribution has *super fat tails*, it is helpful to look at the formal definition of fat-tailed distributions based on the notion of regular variation, as presented, for instance, in Daníelsson et al. (2013).[2]

[2]Daníelsson et al. (2013) point out that the commonly used definition of fat-tailed distributions, which is "higher than normal kurtosis", is not accurate. This is because there exist examples of distributions with high kurtosis and thin tails. In addition, according to definition 4.2.2, kurtosis is not defined for $\gamma \leq 4$.

**Definition 4.2.2 (Fat-tailed distribution)** *A distribution function $F(x)$ has fat tails if it varies regularly at infinity with tail index $\gamma > 0$, that is*

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\gamma}, \quad \forall x > 0. \tag{4.3}$$

The tail of a regularly varying distribution can be represented by the product of a power function and a slowly varying function, i.e. $1 - F(x) = x^{-\gamma}L(x)$. We call $L(x)$ slowly varying if $L(tx)/L(t) \to 1$ as $t \to \infty$ for all $x > 0$ (e.g. the log function). Hence, the tails of a regularly varying distribution essentially decay according to a power function as $x \to \infty$, where the rate of decay is determined by the tail index $\gamma$. Put simply, the lower the tail index, the thicker the tails of the distribution. A *super fat-tailed* distribution exhibits a tail index below one. Another important aspect is that the tail index determines the number of moments that are finite: As tails get thicker, the computation of moments $\mathbb{E}(x^m) = \int x^m f(x)dx$ is affected by increasingly large observations (due to the explosion of $x^m$) causing moments of order $m > \gamma$ to be infinite.

Daníelsson et al. (2013) extend the work of Garcia et al. (2007), Ibragimov and Walden (2007) and Ibragimov (2009) and show that the VaR is subadditive in the relevant tail region for financial assets with (jointly) regularly varying non-degenerate tails in case $\gamma > 1$. Therefore, the VaR is subadditive for a wide range of distributions that are also relevant for modeling financial market data, such as the t-distribution with degrees of freedom larger than one. The t-distribution varies regularly at infinity with the tail index equaling the degrees of freedom. For the losses of nonderivative financial assets, it is quite realistic to assume that the first moment (the mean) exists, so that the VaR is subadditive and thus coherent. For example, Jansen and De Vries (1991) find in their empirical application that the tail index of stocks and stock indices is between 3 and 5. When assets occasionally experience severe losses (e.g. defaultable bonds[3]), the tails can be so heavy that the tail index is less than 1. However, ES is not defined in this case, as it depends on the existence of the first moment. In addition, subadditivity is a controversial assumption when $\gamma < 1$, since diversification can increase risk when portfolio components exhibit super fat tails, see Ibragimov and Walden (2007) and Ibragimov (2009). Hence, it can be concluded that the general lack of subadditivity alone does not exclude the use of VaR as a risk measure for practical applications with (portfolios of) stocks and stock indices.

[3]Further examples would be options, portfolios including short positions or insurance contracts from the insurer's perspective.

### 4.2.3. Elicitability and conditional elicitability

Another important theoretical property that gained a lot of attention in the academic debate concerning the pros and cons of risk measures is *elicitability*, since Gneiting (2011) proved that ES is not elicitable in contrast to VaR. The concept of elicitability was introduced by Osband (1985) and extended in Lambert et al. (2008) and Gneiting (2011) and relates to the evaluation of point forecasts by means of *scoring functions*. A scoring function $S$ is a loss function, or predictive error function in forecasting terminology, that maps point forecasts $x$ and realizations of a random variable $L$ to $\mathbb{R}_+$.[4]

**Definition 4.2.3 (Elicitability)** *The statistical functional $p$ is elicitable with respect to $\mathcal{P}$ if there exists a scoring function $S$ that is strictly consistent for $p$ relative to $\mathcal{P}$, i.e.*

$$\mathbb{E}(S(p(F_L), L)) < \mathbb{E}(S(x, L))$$

*for all $x \neq p(F_L)$.*

Elicitability is an important property, since it implies that the optimal forecast $p^*(F_L) = x^*$ for a risk measure $p$ can be found by minimizing the expected value of a scoring function,

$$p^*(F_L) = x^* = \arg\min_x \mathbb{E}[S(x, L)]. \tag{4.4}$$

In addition, competing forecasts of elicitable functionals may be compared using their expected scores and thus elicitability is useful for forecast ranking and comparative backtesting (cf. Gneiting, 2011; Emmer et al., 2015; Nolde and Ziegel, 2017). In practice, the true distribution $F_L$ is unknown and the expected score is approximated by the mean score for $T$ forecast cases,

$$\bar{S} = \frac{1}{T}\sum_{t=1}^{T} S(x_t, l_t), \tag{4.5}$$

where $x_1, ..., x_T$ denote point forecasts and $l_1, ..., l_T$ are realizations of the random variable $L$. The performance of $M$ alternative estimation methods can be compared by computing $\bar{S}_m$ for $m = 1, ..., M$, since the mean scores tend to be lowest for the most accurate forecasts.

The VaR is an elicitable risk measure for which strictly consistent scoring functions exist. More generally, all quantiles are elicitable and the respective scoring functions are characterized in, for example, Gneiting and Raftery (2007) and Gneiting (2011) and are given by,

$$S(v, l) = (1 - \alpha)G(v) + \mathbb{1}\{l > v\}(G(l) - G(v)) + h(l), \tag{4.6}$$

[4]See, for instance, Bellini and Bignozzi (2015) for a more rigorous definition of scoring functions.

where $v = \text{VaR}_\alpha$, $G$ is a strictly increasing function[5] and $h$ is integrable. In contrast, the ES is not elicitable and thus there exists no natural empirical score in order to compare alternative ES forecasts. However, some statistical functionals that are not elicitable individually can be elicitable jointly with other functionals. In terms of risk measures, this means that the $k$-dimensional vector of the true risk measures $\mathbf{p} = (p_1, ..., p_k)$ minimizes the expected loss of a scoring function $S(\mathbf{x}, L)$ with $\mathbf{x} = (x_1, ..., x_k)$. Fissler and Ziegel (2016) show that this is the case for ES, which is jointly elicitable with VaR[6] and that strictly consistent scoring functions for evaluating the pair $(\text{VaR}_\alpha, \text{ES}_\alpha)$ take the following form,

$$\begin{aligned} S(v, e, l) = \mathbb{1}\{l > v\}(-G_1(v) + G_1(l) - G_2(e)(v - l)) \\ + (1 - \alpha)(G_1(v) - G_2(e)(e - v) + \mathcal{G}_2(e)), \end{aligned} \tag{4.7}$$

where $v = \text{VaR}_\alpha$ and $e = \text{ES}_\alpha$ for notational convenience. The functions $G_1, G_2$ and $\mathcal{G}_2$ fulfill certain properties, among which is the condition that $G_1$ is an increasing function, $\mathcal{G}_2$ is the antiderivative of $G_2$ and $\mathcal{G}_2$ is strictly increasing and strictly concave.[7] Hence, minimizing any member of the scoring functions in (4.7) yields the true pair $(\text{VaR}_\alpha, \text{ES}_\alpha)$. Four concrete versions that have been used in the literature are presented in Table 1 of Taylor (2020). The joint elicitability is a weaker concept than elicitability itself, since, as exemplified by ES, the joint elicitability of $(p_1, ..., p_k)$ does not imply that $p_i$ is elicitable for each $i = 1, ...k$. On the other hand, for every $p_i$ that is elicitable with the corresponding scoring function $S_i$, the vector $(p_1, ..., p_k)$ is jointly elicitable with $S(\mathbf{x}, L) = \sum_{i=1}^{k} S_i(x_i, L)$ (Kou and Peng, 2016).

The lack of elicitability has sparked a controversial debate on whether it affects the ability to perform backtests for the ES. Acerbi and Szekely (2014) claim that elicitability is not a concern for backtesting per se, but only for the relative ranking of alternative models. Fissler et al. (2016) and Nolde and Ziegel (2017) address this point and differentiate between two different types of backtests, namely traditional and comparative backtests. The purpose of traditional backtests is model validation and involves testing a null hypothesis of the form "$H_0$ : The available estimates for the risk measure are correct". In contrast, comparative backtests aim to compare and rank the performance

[5] As pointed out by Gneiting (2011), the condition that $G$ is strictly increasing is the requirement for strict consistency of the scoring function in (4.6). If $G$ is only increasing, then $S$ is consistent for $\text{VaR}_\alpha$ (or quantiles in general).

[6] Another prominent example is the variance, which is not elicitable on its own, but only jointly with the mean.

[7] This is the condition for strict consistency of the scoring functions in (4.7). When $\mathcal{G}_2$ is increasing and concave, the scoring functions in (4.7) are consistent for the pair $(\text{VaR}_\alpha, \text{ES}_\alpha)$.

of alternative estimation methods using strictly consistent scoring functions. In light of this distinction, it is therefore correct that comparative backtests exploit the elicitability property. Nevertheless, conducting traditional backtests is more difficult for ES than for VaR, since ES lacks the property of identifiability, which is closely linked to elicitabiliy. According to Steinwart et al. (2014), both concepts are even equivalent when considering point-valued functionals (i.e. the $k = 1$ case) under some additional assumptions, see also Nolde and Ziegel (2017) and Fissler and Hoga (2023).

Based on Nolde and Ziegel (2017, Definition 2), we define identifiability for the $k = 1$ case as follows:

**Definition 4.2.4 (Identifiability)** *The statistical functional $p$ is identifiable with respect to $\mathcal{P}$ if there is a function $V$ such that*

$$\mathbb{E}(V(x, L)) = 0 \quad \Leftrightarrow \quad x = p(F_L)$$

*for all $L$ with distribution $F_L$ in $\mathcal{P}$.*

For identifiable risk measures, the identification function $V$ can be used to perform traditional backtests by testing whether the sample analog of $V$ is (close to) zero. The VaR is identifiable via the *hit sequence* and traditional backtests are based on the corresponding identification function $V(v, l) = 1 - \alpha - \mathbb{1}\{l > v\}$. In practice, it is tested if the hit sequence is close to $1 - \alpha$. In contrast, the ES is not separately identifiable, but only jointly with the VaR (see Nolde and Ziegel (2017) for the definition of joint identifiability).

Despite the lack of elicitability and identifiability, there are several proposals in the literature for backtesting ES. For instance, Acerbi and Szekely (2014) introduce three nonparametric backtests for ES. Kou and Peng (2016), however, argue that these backtests, as well as all other approaches known to them for backtesting the ES, are indirect backtests. For risk measures, a direct backtest tests whether an estimated risk measure equals the unknown true value of the risk measure, while an indirect approach tests a related quantity, such as the entire tail or loss distribution (as in Acerbi and Szekely, 2014), or a linear approximation of the risk measure (as in Emmer et al., 2015). Indirect backtests can also be based on the joint elicitability of risk measures such as the ones proposed in Fissler et al. (2016) for the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$.

There are two main problems with using indirect backtests for ES that require additional input variables. Firstly, if the null hypothesis of a correct model is rejected, it remains unclear if the ES estimate is incorrect or if the rejection is the result of a poor estimation of another variable. Secondly, from a regulatory perspective, the additional

input variables may not necessarily be available, as financial institutions are not required to disclose them (Kou and Peng, 2016; Bayer and Dimitriadis, 2022). A possible solution to these problems was recently proposed by Bayer and Dimitriadis (2022), since they suggested ES backtests based on the Mincer-Zarnowitz regression that only require ES forecasts and realized losses as input variables.

Fissler et al. (2016) and Nolde and Ziegel (2017) propose to complement the traditional backtests used in regulatory practice with comparative backtests. Specifically, Nolde and Ziegel (2017) suggest a two-stage procedure: In the first stage, a traditional backtest is conducted. In the second stage, provided that the first stage is passed, a comparative backtest is performed to compare a financial institution's internal model against a regulator's standard model. When solely traditional backtests are used for model validation, this could provide an incentive to minimize the risk measure estimate under the condition that the backtest is passed, rather than aiming for the best possible forecasts. To incorporate comparative backtests into regulation, a standardized method for risk measure estimation is needed for the comparison with internally generated forecasts from financial institutions. Nolde and Ziegel (2017) propose the filtered historical simulation (FHS) method as a potential candidate for this. For ES, it remains challenging to find a reliable comparative backtesting method due to the lack of elicitability. One possible solution is to use the scoring functions in (4.7) based on the joint elicitability of VaR and ES, however it is unclear whether a poor estimation of the ES, VaR or both is responsible for the failure of an estimation method in a comparative backtest, see also examples 5 and 6 in Kou and Peng (2016). Dimitriadis and Schnaitmann (2021) recently proposed an encompassing test for situations where only ES forecasts are available. However, the authors also recommend to publish both ES and VaR forecasts as a standard practice given their joint elicitability.

Overall, it appears that backtesting ES predictions is more challenging than backtesting VaR predictions, since the ES lacks elicitability (and identifiability). However, the discovery by Fissler and Ziegel (2016) that ES is elicitable (and identifiable) jointly with VaR opens up new possibilities for evaluating ES predictions, leading Bayer and Dimitriadis (2022) to conclude that "[...] the ES is an appropriate candidate for being the standard risk measure in practice". Kou and Peng (2016) are more skeptical and favor quantile-based risk measures such as the median shortfall.[8] They claim that the joint elicitability of "[...] $(\mathrm{ES}_\alpha, \mathrm{VaR}_\alpha)$ does not lead to a reliable method for evaluating forecasts for ES."

[8] Kou and Peng (2016) refer to the median shortfall as the median of the tail loss distribution, e.g. the VaR at level $(1+\alpha)/2$.

### 4.2.4. Robustness

The modeling of risk measures such as VaR and ES entails two major challenges in practice. First, there is a high degree of model uncertainty in the estimation of risk measures, since the "true" underlying model is unknown. This creates perils for modeling, such as model misspecification or parameter estimation errors. Second, the estimation of risk measures is based on either historical or simulated data, which may not accurately predict future risks. Problems can arise in the estimation, for example, due to limited data or contaminated datasets. With ES set to replace VaR as the regulatory risk measure for determining capital requirements under Basel III, it is important to consider which risk measure is more robust in the face of these uncertainties. However, robustness has received less attention in the academic literature than coherence and elicitability and lacks a uniform definition. The following presents the definition of robustness according to Kou et al. (2013) and then discusses some important contributions to the topic.

Kou et al. (2013) address the uncertainties mentioned above and refer to a risk measure as robust if (i) it can accommodate model misspecifications and (ii) exhibits statistical robustness with respect to changes in the data, see also He et al. (2022). The authors emphasize the importance of having a robust regulatory risk measure that provides reliable results and that can be consistently implemented in all institutions. If a risk measure lacks robustness, two institutions with equal risk profiles that use different methods for estimating the risk measure could face completely different capital requirements in case both estimation methods pass the regulatory backtests. This incentivizes the preference for estimation methods that output lower values for the risk measure. Kou et al. (2013) refer to the *robustness of law*, which involves designing a law in such a way that different judges come to the same decision when applying it.

Kellner and Rösch (2016) address the first aspect of robustness and conduct an empirical analysis to compare the model risk of ES at level 97.5% and VaR at level 99%. The authors quantify the so-called legal robustness, which measures the mean absolute deviation among different estimation methods that previously passed a traditional backtest. In addition, the sensitivity of the risk measures to errors in the estimation of the model parameters is quantified by evaluating the ratio of partial derivatives of the risk measures with respect to the corresponding parameter. The study finds that the variability of estimates between reasonable models for $\mathrm{ES}_{97.5\%}$ is higher than for $\mathrm{VaR}_{99\%}$ in most cases. In addition, estimating the ES carries a higher risk of parameter misspecification, as the estimates have a higher variability when the estimated parameter deviates from the true value. It is particularly problematic that the estimates of $\mathrm{ES}_{97.5\%}$ are less reliable than $\mathrm{VaR}_{99\%}$ estimates during adverse market conditions and that heavier

tails in the loss distribution lead to an increased difference in model risk between the two risk measures, since reliable estimates in terms of capital requirements are needed precisely under these circumstances. Kellner and Rösch (2016) conclude that the ES is more sensitive to regulatory arbitrage and that there is a trade-off between the ability to incorporate extreme events and model risk due to misspecification of parameters and higher variability in estimates.

Cont et al. (2010) focus on the second aspect of robustness mentioned above and investigate the impact of a small change in the dataset (i.e. adding a new data point) on various estimation methods for evaluating the robustness of VaR and ES using sensitivity functions, which is a tool from robust statistics. They conclude that ES lacks robustness in this regard, while the VaR exhibits bounded sensitivity. They demonstrate that there is a fundamental contradiction between subadditivity (and hence coherence) and the robustness of risk measures with regard to the dataset used for their computation. It turns out that the estimation method plays an important role for this form of robustness, as a parametric model, for instance, can react differently to the addition of a data point than a model based on the empirical loss distribution. This sensitivity varies considerably across different ES models. Cont et al. (2010) also show that the popular historical simulation (HS) estimator is significantly more robust for VaR than for ES when adding a new observation. Furthermore, the ES is more sensitive to the size of the data point.

In the study by He et al. (2022), the most important results from the literature on robust statistics are summarized. It is shown that VaR is more robust than ES with regard to four methods of robust statistics, namely influence functions, asymptotic breakdown points, finite sample breakdown points and Hampel robustness. For example, VaR has a bounded influence function while that of ES is unbounded. Put simply, the asymptotic breakdown point is a measure of the proportion of outliers that an estimator can tolerate before its behavior becomes arbitrary as the sample size approaches infinity. A higher breakdown point indicates greater robustness to outliers. While the asymptotic breakdown point is $1 - \alpha$ for $\text{VaR}_\alpha$, it is 0 for $\text{ES}_\alpha$. In practice, the computation of ES may be substantially affected by adding one data point to a finite sample. For more details, refer to He et al. (2022) and the references therein.

In summary, it can be concluded that the VaR is more robust than the ES with regard to noise in the dataset and model misspecifications. However, this observation must be balanced against the ability of ES to respond to tails events. When robustness primarily refers to the sensitivity to outliers as in Cont et al. (2010), it is not surprising that the estimation of ES as a conditional expectation is more sensitive than the estimation of a quantile. In fact, ES was implemented as a regulatory risk measure in Basel III to

capture possible extreme losses in the tail of the distribution, which VaR is not sensitive to. Extreme values can occur in financial market data and are not necessarily outliers, see the discussion in Emmer et al. (2015). Nevertheless, it is desirable for a regulatory risk measure that alternative estimation methods that are accepted by traditional backtests respond similarly to changes in the dataset. The sensitivity of ES with respect to data and model assumptions makes it challenging to implement this risk measure consistently across all financial institutions in the current regulatory setting, as each institution uses its own data and internal models to calculate risk measures. This promotes regulatory arbitrage in both the model selection and the data used for computation.

### 4.2.5. Summary of properties

The previous sections discuss some important concepts regarding risk measures and their differences with respect to the two popular risk measures VaR and ES. ES is often considered the theoretically superior risk measure, as it is subadditive and coherent, but mainly because it accounts for tail risks beyond the VaR. In contrast, ES is not elicitable and is less robust with regard to model misspecifications and noise in the data. Therefore, there is a trade-off between subadditivity and sensitivity to extreme events on the one hand, and elicitability and robustness on the other.

In light of this tradeoff and the current regulatory requirements for backtesting, it is unclear whether adopting $\mathrm{ES}_{97.5\%}$ as regulatory risk measure is justified. As mentioned above, the lack of subadditivity does not seem to be a relevant issue in most practical applications and in cases where VaR is not subadditive, this concept is controversial. The lack of elicitability complicates backtesting and model selection for the ES, although recent publications in this field have made progress. While ES theoretically covers extreme events beyond VaR and captures both their size and likelihood, it is unclear to what extent an estimated ES can actually quantify these tail risks in practice, given the lack of robustness.

There is no consensus in the literature about the most suitable risk measure for regulation. While, for example, Emmer et al. (2015) and Bayer and Dimitriadis (2022) consider ES to be the appropriate risk measure in practice, other authors such as Cont et al. (2010), Kou and Peng (2016) and He et al. (2022) are more skeptical and prefer quantile-based risk measures.

## 4.3. ES/VaR ratios for the normal and t-distribution

Despite the theoretical differences discussed in Section 4.2, there exists a relationship between VaR and ES under certain distributional assumptions. This section presents VaR and ES for location-scale families, followed by the closed-form solutions of both risk measures for the standard normal distribution and the standard t-distribution. Based on this, the ratios between ES and VaR for both distributions are derived and the performance of the t-ratio estimator is examined in a simulation study.

### 4.3.1. VaR and ES for location scale families

Consider a random variable $X$ from the location-scale family of distributions with a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\sigma \in \mathbb{R}_+$ such that there exists a standardized random variable $Z = (X - \mu)/\sigma$. Since both VaR and ES satisfy the properties of positive homogeneity and translation invariance, it holds that

$$\mathrm{VaR}_\alpha(X) = \mu + \sigma \mathrm{VaR}_\alpha(Z), \text{ and} \tag{4.8}$$

$$\mathrm{ES}_\alpha(X) = \mu + \sigma \mathrm{ES}_\alpha(Z) \tag{4.9}$$

for all location-scale families. In the following, we consider the derivations of VaR and ES for two popular location-scale families, the normal and the Student-t distribution.

### 4.3.2. Normal distribution

Suppose that the random variable $Z$ has a standard normal distribution, thus $Z \sim N(0,1)$. Then it holds that

$$P(Z \leq \Phi^{-1}(\alpha)) = \Phi(\Phi^{-1}(\alpha)) = \alpha,$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution and $\Phi^{-1}(\cdot)$ is the corresponding quantile function. Since $\mathrm{VaR}_\alpha(Z) = \Phi^{-1}(\alpha)$ according to the definition in (4.1), the VaR for normally distributed losses $L$ with location parameter $\mu$ and scale parameter $\sigma$ is

$$\mathrm{VaR}_\alpha(L) = \mu + \sigma\Phi^{-1}(\alpha). \tag{4.10}$$

Let $\phi(z)$ denote the density function of $Z$. Then, according to the definition in (4.2), a closed-form solution for the ES is obtained by computing the following integral,

$$\begin{aligned} ES_\alpha(Z) &= \frac{1}{1-\alpha}\int_{\Phi^{-1}(\alpha)}^{\infty} z\phi(z)dz \\ &= \frac{1}{(1-\alpha)\sqrt{2\pi}}\int_{\Phi^{-1}(\alpha)}^{\infty} z\,\exp\left(-\frac{1}{2}z^2\right)dz. \end{aligned}$$

With the substitution $u(z) = -\frac{1}{2}z^2$ and taking the limit yields

$$\begin{aligned} ES_\alpha(Z) &= -\frac{1}{(1-\alpha)\sqrt{2\pi}}\lim_{b\to\infty}\int_{u(\Phi^{-1}(\alpha))}^{u(b)}\exp(u)\,du \\ &= -\frac{1}{(1-\alpha)\sqrt{2\pi}}\lim_{b\to\infty}\left[\exp(u)\right]_{u(\Phi^{-1}(\alpha))}^{u(b)} \\ &= -\frac{1}{(1-\alpha)\sqrt{2\pi}}\lim_{b\to\infty}\left(\exp\left(-\frac{1}{2}b^2\right)-\exp\left(-\frac{1}{2}(\Phi^{-1}(\alpha))^2\right)\right) \\ &= \frac{1}{(1-\alpha)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\Phi^{-1}(\alpha)\right)^2\right) \\ &= \frac{1}{(1-\alpha)}\phi(\Phi^{-1}(\alpha)) \end{aligned} \tag{4.11}$$

(see, e.g., McNeil et al., 2015, p.70). For $L \sim N(\mu, \sigma^2)$, ES follows as

$$ES_\alpha(L) = \mu + \sigma ES_\alpha(Z). \tag{4.12}$$

### 4.3.3. Student-t distribution

The VaR and ES for returns that have a generalized Student-t distribution with $\nu$ degrees of freedom can be derived in a similar manner.[9] The VaR results as

$$\mathrm{VaR}_\alpha(L) = \mu + \sigma t_\nu^{-1}(\alpha), \tag{4.13}$$

where $t_\nu^{-1}(\alpha)$ denotes the $\alpha$-quantile of the standard Student-t distribution. For the derivation of ES, suppose that the random variable $Z$ has a standard Student-t distribution with $\nu$ degrees of freedom where $f_\nu(z)$ denotes the density function and $t_\nu^{-1}(\alpha)$ is the $\alpha$-quantile of the distribution. The closed-form solution for ES is obtained by

[9] Note that the variance of the generalized Student-t distribution is given by $\nu\sigma^2/(\nu-2)$ and does not equal $\sigma^2$.

computing

$$\begin{aligned} ES_\alpha(Z) &= \frac{1}{(1-\alpha)} \int_{t_\nu^{-1}(\alpha)}^{\infty} z \; f_\nu(z) dz \\ &= \frac{1}{(1-\alpha)} \underbrace{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}}_{:=c} \int_{t_\nu^{-1}(\alpha)}^{\infty} z \left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}} dz, \end{aligned} \tag{4.14}$$

where $\Gamma(x) = (x-1)!$ denotes the Gamma function. Substituting $u(z) = 1 + \frac{z^2}{\nu}$ yields

$$\begin{aligned} ES_\alpha(Z) &= \frac{c\nu}{2(1-\alpha)} \lim_{b\to\infty} \int_{u(t_\nu^{-1}(\alpha))}^{u(b)} u^{-\frac{\nu+1}{2}} du \\ &= \frac{c\nu}{(1-\alpha)(1-\nu)} \lim_{b\to\infty} \left[u^{\frac{1-\nu}{2}}\right]_{u(t_\nu^{-1}(\alpha))}^{u(b)} \\ &= \frac{c\nu}{(1-\alpha)(1-\nu)} \lim_{b\to\infty} \left(\left(1 + \frac{b^2}{\nu}\right)^{\frac{1-\nu}{2}} - \left(1 + \frac{(t_\nu^{-1}(\alpha))^2}{\nu}\right)^{\frac{1-\nu}{2}}\right) \\ &= \frac{c\nu}{(1-\alpha)(\nu-1)} \left(1 + \frac{(t_\nu^{-1}(\alpha))^2}{\nu}\right)^{\frac{1-\nu}{2}}, \quad \text{for } \nu > 1, \\ &= \frac{\nu}{(1-\alpha)(\nu-1)} \left(1 + \frac{(t_\nu^{-1}(\alpha))^2}{\nu}\right) \underbrace{c\left(1 + \frac{(t_\nu^{-1}(\alpha))^2}{\nu}\right)^{-\frac{\nu+1}{2}}}_{f_\nu(t_\nu^{-1}(\alpha))} \\ &= \frac{\nu + (t_\nu^{-1}(\alpha))^2}{(1-\alpha)(\nu-1)} f_\nu(t_\nu^{-1}(\alpha)) \end{aligned} \tag{4.15}$$

(see, e.g., McNeil et al., 2015, p.70). For a generalized Student-t distributed loss variable $L$, the ES results as

$$ES_\alpha(L) = \mu + \sigma ES_\alpha(Z), \tag{4.16}$$

accordingly.

### 4.3.4. Ratios

The closed-form solutions for the ES in (4.11) and (4.15) can be used to explicitly calculate the ratio ES/VaR for the normal and Student-t distributions. For losses that have a standard normal distribution, the ratio is

$$\frac{ES_\alpha}{VaR_\alpha} = \frac{1}{(1-\alpha)\Phi^{-1}(\alpha)} \phi\left(\Phi^{-1}(\alpha)\right). \tag{4.17}$$

Hence, the normal-ratio is a function of $\alpha$. For the standard Student-t distribution, the ratio depends on the degrees of freedom $\nu$ and results as

$$\frac{ES_\alpha}{VaR_\alpha} = \frac{\nu + (t_\nu^{-1}(\alpha))^2}{(1-\alpha)(\nu-1)t_\nu^{-1}(\alpha)} f_\nu(t_\nu^{-1}(\alpha)). \tag{4.18}$$

Table 4.1 displays the VaR, ES and ratio for the normal and Student-t distribution with various degrees of freedom at some prominent confidence levels. The table shows that the t-ratio increases with tail heaviness, while the ES and VaR are related by a small constant in the normal case.

Table 4.1.: VaR, ES and ratios for the Student-t and normal distribution

| | VaR | | | ES | | | ratio | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 99% | 97.5% | 95% | 99% | 97.5% | 95% | 99% | 97.5% | 95% |
| $t_3$ | 4.54 | 3.18 | 2.35 | 7.00 | 5.04 | 3.87 | 1.54 | 1.58 | 1.65 |
| $t_4$ | 3.75 | 2.78 | 2.13 | 5.22 | 3.99 | 3.20 | 1.39 | 1.44 | 1.50 |
| $t_5$ | 3.36 | 2.57 | 2.02 | 4.45 | 3.52 | 2.89 | 1.32 | 1.37 | 1.43 |
| $t_6$ | 3.14 | 2.45 | 1.94 | 4.03 | 3.26 | 2.71 | 1.28 | 1.33 | 1.40 |
| $t_7$ | 3.00 | 2.36 | 1.89 | 3.77 | 3.09 | 2.59 | 1.26 | 1.31 | 1.37 |
| $t_8$ | 2.90 | 2.31 | 1.86 | 3.59 | 2.97 | 2.51 | 1.24 | 1.29 | 1.35 |
| $t_9$ | 2.82 | 2.26 | 1.83 | 3.46 | 2.88 | 2.45 | 1.23 | 1.28 | 1.34 |
| $t_{10}$ | 2.76 | 2.23 | 1.81 | 3.36 | 2.82 | 2.41 | 1.22 | 1.27 | 1.33 |
| $t_{15}$ | 2.60 | 2.13 | 1.75 | 3.10 | 2.64 | 2.28 | 1.19 | 1.24 | 1.30 |
| $t_{20}$ | 2.53 | 2.09 | 1.72 | 2.98 | 2.56 | 2.22 | 1.18 | 1.23 | 1.29 |
| $N(0,1)$ | 2.33 | 1.96 | 1.64 | 2.67 | 2.34 | 2.06 | 1.15 | 1.19 | 1.25 |

This table displays values for the VaR, ES and the ratio ES/VaR for the standard Student-t distribution for various degrees of freedom and for the standard normal distribution.

### 4.3.5. Simulation study

In this section, the results of a simple simulation study for the risk measures VaR and ES are presented for the confidence levels 97.5% and 99% that are relevant for regulatory practice. The study uses a standard t-distribution with degrees of freedom ranging between three and six as the data-generating process to create a simple laboratory setting in which the true values of VaR and ES are known. This allows to compare the performance with regard to bias, variance and MSE. The choice of degrees of freedom aims to simulate heavy tails as they are often observed in stock data, while maintaining the existence of the first two moments.

In the simulation, the maximum likelihood estimator (MLE) is compared to the historical simulation (HS) estimator for both risk measures at the two confidence levels 97.5% and 99%. In addition, for ES, the ratio estimator is computed and compared to

the other two estimation methods. In this laboratory setting, the MLE is the asymptotically most efficient estimator and is obtained by plugging in the maximum likelihood estimate for the degrees of freedom into (4.15). On the other hand, HS is one of the most popular estimators in practice (European Banking Authority, 2021). Given a sample of past observations $x_{t-1}, ..., x_{t-w}$ where $w$ denotes the length of the estimation window, estimates for the VaR and the ES are obtained as the empirical $\alpha$-quantile and the mean of all observations exceeding VaR, respectively, i.e.

$$\widehat{VaR}_\alpha^{HS} = \widehat{F}^{-1}(\alpha), \tag{4.19}$$

$$\widehat{ES}_\alpha^{HS} = \frac{\sum_{i=1}^{w} x_{t-i} \cdot \mathbb{1}\{x_{t-i} > \widehat{VaR}_\alpha^{HS}\}}{\sum_{i=1}^{w} \mathbb{1}\{x_{t-i} > \widehat{VaR}_\alpha^{HS}\}}, \tag{4.20}$$

where $\widehat{F}$ denotes the empirical distribution. The HS estimator is a simple nonparametric method that makes only one assumption about the distribution of the data, which is that the data is i.i.d. For this simulation, the multiplier for the ratio estimator is selected using a data-driven approach. The t-ratio estimator is simply obtained by multiplying $\widehat{VaR}_\alpha^{HS}$ with the ratio in (4.18), where the degrees of freedom are replaced by the maximum likelihood estimate for $\nu$.[10] The simulations are based on 2000 replications. For each simulation run, data for 400 artificial stocks are simulated, with 100 for each degree of freedom.

Table 4.2 reports the results of the simulations for estimating the two risk measures at the relevant confidence levels for an estimation window of 1000 periods. This corresponds to approximately four years of daily data in a financial market application. The estimators are evaluated in terms of bias, variance and MSE. It is not surprising that the MLE outperforms the other estimators, as it is the asymptotically most efficient estimator under these simulation conditions. For real data, on the other hand, it becomes necessary to make an assumption about the unknown loss distribution, and accordingly, a pseudo-MLE is employed in practice.

The results in Table 4.2 show that estimating ES by means of MLE and HS leads to higher variance and MSE compared to estimating the corresponding VaR counterparts at the same confidence levels for all considered degrees of freedom. Additionally, the absolute bias is higher in most cases. The differences are more pronounced for lower degrees of freedom, i.e. when the tails are heavier. For example, the MSE for both

[10] In the optimization carried out in Matlab 2020a, the only condition for the degrees of freedom is that $\nu > 0$. The maximum likelihood estimation involves numerical optimization techniques and aims to find the parameter values that maximize the likelihood of observing the given data. This optimization process can also yield estimates for $\nu$ that are non-integer.

Table 4.2.: Simulation results for t-distributed data

| | VaR 99% | | VaR 97.5% | | ES 99% | | | ES 97.5% | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dof | MLE | HS | MLE | HS | MLE | HS | ratio | MLE | HS | ratio |
| | | | | | Bias | | | | | |
| 3 | 0.39 | 3.03 | 0.04 | 0.89 | 2.71 | −8.05 | 6.82 | 1.18 | −2.37 | 2.53 |
| 4 | 0.02 | 1.32 | −0.09 | 0.35 | 0.87 | −5.47 | 2.69 | 0.31 | −1.84 | 0.95 |
| 5 | −0.09 | 0.88 | −0.13 | 0.26 | 0.36 | −4.36 | 1.65 | 0.08 | −1.51 | 0.60 |
| 6 | −0.16 | 0.53 | −0.15 | 0.18 | 0.11 | −3.83 | 0.99 | −0.05 | −1.39 | 0.39 |
| Avg. | 0.04 | 1.44 | −0.08 | 0.42 | 1.01 | −5.43 | 3.04 | 0.38 | −1.78 | 1.12 |
| | | | | | Variance | | | | | |
| 3 | 8.78 | 28.98 | 2.15 | 6.67 | 51.59 | 226.44 | 101.36 | 16.76 | 50.93 | 28.39 |
| 4 | 4.15 | 13.25 | 1.16 | 3.73 | 18.24 | 62.10 | 36.29 | 6.64 | 16.56 | 12.05 |
| 5 | 2.65 | 8.30 | 0.80 | 2.65 | 9.96 | 29.40 | 20.03 | 3.88 | 8.82 | 7.42 |
| 6 | 1.95 | 6.09 | 0.62 | 2.11 | 6.66 | 18.17 | 13.56 | 2.71 | 5.92 | 5.38 |
| Avg. | 4.38 | 14.16 | 1.18 | 3.79 | 21.61 | 84.03 | 42.81 | 7.50 | 20.56 | 13.31 |
| | | | | | MSE | | | | | |
| 3 | 8.78 | 29.08 | 2.15 | 6.68 | 51.66 | 227.08 | 101.83 | 16.78 | 50.99 | 28.45 |
| 4 | 4.15 | 13.26 | 1.16 | 3.73 | 18.25 | 62.40 | 36.37 | 6.64 | 16.59 | 12.05 |
| 5 | 2.65 | 8.31 | 0.80 | 2.66 | 9.96 | 29.59 | 20.06 | 3.88 | 8.84 | 7.42 |
| 6 | 1.95 | 6.09 | 0.62 | 2.11 | 6.66 | 18.32 | 13.57 | 2.71 | 5.94 | 5.38 |
| Avg. | 4.38 | 14.19 | 1.18 | 3.79 | 21.63 | 84.35 | 42.96 | 7.50 | 20.59 | 13.33 |

This table reports simulation results for estimating VaR and ES by MLE and HS at the 99%, 97.5% confidence levels. The data is generated by a standard t-distribution with 3-6 degrees of freedom. The results are based on 400 assets (100 for each d.o.f.), 1000 time periods and 2000 replications and all performance numbers are multiplied by 100.

MLE and HS is almost eight times higher when the risk measures are compared at the 97.5% confidence level and for three degrees of freedom. For six degrees of freedom, the MSE of the MLE estimator is still more than four times as large and the MSE of the HS estimator is nearly tripled. The simulation results show that the estimation of $ES_{99\%}$ is a difficult task, in particular for the HS estimator. This is because, even with a relatively long estimation window of 1000 periods, the HS estimator is based on only ten observations. As a result, each of these individual observations has a substantial impact on the outcome of the ES estimation. When the degrees of freedom are low and the tails are correspondingly heavier, the likelihood of extreme observations increases, thereby affecting the computation of ES and leading to increased variability.

In light of the transition from $VaR_{99\%}$ to $ES_{97.5\%}$ as the regulatory risk measure, the comparison between these two risk measures is particularly relevant for regulatory purposes. The simulation results for $VaR_{99\%}$ and $ES_{97.5\%}$ are listed in columns two and three, as well as the last three columns of Table 4.2, respectively. The MLE consistently shows lower values for variance and MSE for $VaR_{99\%}$ as compared to $ES_{97.5\%}$. This difference is again particularly pronounced for low degrees of freedom. For three degrees of freedom, the MLE exhibits an almost twice as high MSE for $ES_{97.5\%}$ in comparison to $VaR_{99\%}$. For six degrees of freedom, the MSE for the $ES_{97.5\%}$ is still about 40% higher. The bias is relatively low for both risk measures, so the MSE is mainly driven by the variance. In comparison to the MLE, the HS estimator has higher levels of bias and variance overall. However, similar to the MLE, estimating $VaR_{99\%}$ with HS provides advantages compared to estimating $ES_{97.5\%}$ in terms of variance and MSE for low degrees of freedom. For three degrees of freedom, the MSE of the HS estimator for $ES_{97.5\%}$ is about 75% higher than for $VaR_{99\%}$. When the degrees of freedom are 5-6, the differences in MSE between the two risk measures are minor.

One major challenge for estimating the ES at the 97.5% confidence level is the high variability in estimation when tails are heavy, in particular for the widely-used HS estimator. In contrast, the HS estimator shows lower bias, variance and MSE values when used to estimate the VaR at the 97.5% confidence level. To mitigate the issue, the ratio estimator approach proposes to compute ES by multiplying the VaR at the 97.5% level with a ratio, which can lead to a more stable estimate of ES. In this simulation study, the ratio is estimated from the data and depends on the estimated degrees of freedom, reflecting the heaviness of the tails. When the VaR is multiplied by a constant that is larger than one, the variance increases due to its property $V(aX) = a^2V(X)$ for a random variable $X$ and any constant $a$. Accordingly, the variance of the ratio estimator for $ES_{97.5\%}$ is higher than that of the HS estimator for $VaR_{97.5\%}$. However, Table 4.2

shows that for $ES_{97.5\%}$, the ratio estimator exhibits a lower variance and MSE compared to the HS estimator for all considered degrees of freedom, as well as a lower absolute average bias. The improvement in variance and MSE is most noticeable for low degrees of freedom. Moreover, the ratio estimator for $ES_{97.5\%}$ achieves values for bias, variance and MSE that are slightly lower than those of the HS estimator for $VaR_{99\%}$. Recall that $VaR_{99\%}$ was the former regulatory risk measure for determining capital requirements in Basel II. These findings suggest that the implementation of a ratio estimator could offer practical utility by stabilizing ES estimates at the confidence level 97.5% and thus reduce the variability in the calculation of capital requirements. For $ES_{99\%}$, the ratio estimator also yields significantly lower values for variance and MSE in comparison to the HS estimator, although the overall level remains relatively high.

The simulation results for an estimation window of 500 periods are presented in Table C.10 in Appendix C. Despite a noticeably higher overall level of bias, variance and MSE, the relative performance of the estimators for the two risk measures appears similar to that for an estimation window of 1000. For the same confidence level, the MLE and HS estimators exhibit substantially lower bias, variance and MSE values for the estimation of VaR compared to ES. For the estimation of ES, the ratio estimator shows markedly lower variance and MSE values than the HS estimator, while the performance of the ratio estimator for $ES_{97.5\%}$ is again slightly better than that of the HS estimator for $VaR_{99\%}$. These results demonstrate the challenging nature of obtaining accurate ES estimates with smaller estimation windows.

## 4.4. Bootstrap resampling application

In this section, we examine the performance of some prominent estimation methods on real data and compare them to ratio models, where the estimation of the ES is obtained by multiplying the VaR with a constant. The results of the empirical application are based on daily stock data from the Center for Research in Security Prices (CRSP) for the time period 26 December 2003 until 14 November 2019, yielding 4000 daily returns WRDS (2020). The dataset comprises 760 stocks for which we have an almost complete return history for the considered time period.[11] To ensure a certain quality standard of the stocks included in our investment universe, we restrict attention to stocks that were constituents of the S&P 500 index on the last trading day of the years between 2010 and 2019 plus constituents of the NASDAQ composite index. All included stocks are

[11] We allow for <1% of missing returns during the considered time period. The few missing values are replaced by zeros.

common shares that are traded on the NYSE or NASDAQ stock exchanges with no more than ten recorded trading days without trading volume. We chose these requirements to avoid the inclusion of illiquid stocks with high spreads.

### 4.4.1. Application setup

For real data, the performance comparison of alternative estimation methods for VaR and ES is considerably more challenging than for simulated data, since the true values of these risk measures are unknown for real data. Under the laboratory conditions of the simulation study in Section 4.3.5, the true values for VaR and ES could be determined according to the formulas in Section 4.3.3, allowing for the comparison of various estimation methods in terms of bias, variance and MSE.

To mimic such laboratory conditions for the real dataset mentioned above, the complete set of 4000 (filtered) time series of the 760 stocks is considered as the population from which the true VaR and ES for each stock are determined based on the empirical distribution. Bootstrapping is then employed to generate new samples from this dataset by resampling the data points. Specifically, for each new sample, T time points are randomly drawn with replacement from the original 4000 time points of the population. This results in each new sample having the dimension T $\times$ 760. Thus, for each new sample, there is a new composition of time points, while maintaining the grouping of stocks to preserve the correlations among them. In total, $B$ new samples are generated using this procedure and the VaR and ES are calculated for each sample using various estimation methods. This approach allows to compare the performance of alternative estimation approaches with regard to average bias, variance and MSE based on the deviation of the sample-based estimates from the true values obtained from the population.

In order to apply this bootstrap procedure, the time series should be i.i.d. However, financial market data often exhibit autocorrelation and heteroskedasticity. Therefore, it is essential to apply suitable filtering techniques to the time series of the 760 stocks. In this context, we assume that the series of losses is of the form

$$L_t = \mu_t + \varepsilon_t \quad \text{with} \quad \varepsilon_t = \sigma_t Z_t \tag{4.21}$$

where $\mu_t$ denotes the conditional mean (location), $\sigma_t$ represents the conditional standard deviation (scale) and both parameters are $\mathcal{F}_{t-1}$ measurable, the information available up to time point $t-1$. The series $Z_t$ are i.i.d. innovations with zero mean and unit variance with distribution function $F_Z$. A widely employed approach for capturing the dynamics in financial market data involves combining an ARMA model to estimate the

conditional mean $\mu_t$ with a GARCH model to estimate the conditional variance $\sigma_t^2$. In this study, we specifically employ an AR(1)-GARCH(1,1) model to filter the data, which is a common approach used by many other researchers in their empirical studies focused on forecasting risk measures, e.g. Kellner and Rösch (2016), Nolde and Ziegel (2017) or Li and Wang (2023). The AR(1)-GARCH(1,1) model is specified as

$$L_t = \phi_0 + \phi_1 L_{t-1} + \varepsilon_t \quad \text{with} \quad \varepsilon_t = \sigma_t Z_t, \tag{4.22}$$

$$\sigma_t^2 = \omega + \delta_1 \varepsilon_{t-1}^2 + \delta_2 \sigma_{t-1}^2 \tag{4.23}$$

where $\mu_t = \mathbb{E}[L_t|\mathcal{F}_{t-1}] = \phi_0 + \phi_1 L_{t-1}$ and $Z_t \overset{i.i.d.}{\sim} f_Z$. The sequence of i.i.d. innovations can be obtained using the following two step procedure:

1. The parameters $\mu_t$ and $\sigma_t$ are estimated using maximum likelihood, assuming a specific distribution for $Z_t$.
2. Compute the standardized residuals using the estimated parameters $\hat{\mu}_t$ and $\hat{\sigma}_t$,

$$\hat{Z}_t = \frac{L_t - \hat{\mu}_t}{\hat{\sigma}_t}. \tag{4.24}$$

The risk measures can be computed based on these standardized residuals. For the loss function $L_t$, the risk measure estimates can simply be obtained by $\hat{p}_t(L_t) = \hat{\mu}_t + \hat{\sigma}_t \hat{p}_t(\hat{Z}_t)$, where $\hat{p}_t$ denotes an estimator for a risk measure at time $t$.

A simple assumption for the distribution of $Z_t$ is $N(0,1)$. While this assumption often leads to a misspecified model for stock returns, an AR(1)-GARCH(1,1) filter with normally distributed innovations removes a large portion of autocorrelation and heteroskedasticity and generates approximately i.i.d. distributed residuals, see Kuester et al. (2006). In this application, a Student-t distribution is assumed for filtration to better account for the heavy tails typically observed in stock returns. This assumption changes the functional form of the likelihood function and the corresponding additional parameter, the degrees of freedom $\nu_t$, is estimated in the first step described above, together with the parameters $\mu_t$ and $\sigma_t$. The AR(1)-GARCH(1,1) model with t-distributed innovations yields approximately i.i.d. residuals $\hat{Z}_t$ based on Ljung-Box test results. This model represents an improvement over a filter that uses normal innovations. On the other hand, more complex models such as GJR-GARCH or a model employing a skewed-t distribution do not seem to provide any improvement in terms of filtration for the dataset used.

For the bootstrap experiment, samples with $T = 1000$ and $T = 500$ observations are examined. The estimation methods that are described in Section 4.4.2 are used

to compute VaR and ES for the confidence levels 97.5% and 99% that are particularly relevant with regard to the Basel regulations. All results are derived from $B = 2000$ generated samples.

### 4.4.2. Estimators

For the bootstrap experiment, three prominent estimation methods serve as a benchmark, and their performance is compared to various ratio models. These three estimation methods include the well-known estimation method (filtered) historical simulation, which is considered in most related studies and is also a popular estimation method used by banks in practice. In addition, two estimation methods based on extreme value theory are included. Estimation methods based on extreme value theory are considered, for example, in Kuester et al. (2006), Kellner and Rösch (2016), Nolde and Ziegel (2017) and Taylor (2019). According to the empirical results in Kellner and Rösch (2016), all three models that serve as a benchmark in this section are among the best models for estimating $\mathrm{ES}_{97.5\%}$ and pass their conducted backtests. All estimations in this section are computed based on the samples consisting of standardized residuals.

#### FHS

The nonparametric estimation method known as filtered historical simulation (FHS) is similar to the HS method presented in equations (4.19) and (4.20), with the difference that the estimators are computed based on standardized residuals. Hence, the estimator for $\mathrm{VaR}_\alpha^{\mathrm{FHS}}$ is the empirical $\alpha$-quantile of the bootstrap sample $Z_b$ and $\widehat{\mathrm{ES}}_\alpha^{\mathrm{FHS}}$ corresponds to the average of the observations above $\widehat{\mathrm{VaR}}_\alpha^{\mathrm{FHS}}$.

For the simulations in Section 4.3.5, no filtration is necessary, since the simulated data is i.i.d. However, it is important to note that data filtration is important for financial market returns. FHS is one of the most investigated methods for the estimation of VaR and ES. Regarding ES, Patton et al. (2019) refer to FHS as "perhaps the best existing model for ES". Nolde and Ziegel (2017) propose FHS as a potential "standard model" against which alternative models should be compared to in comparative backtesting, as it is flexible and performs well in many situations.

The unconditional historical simulation method (HS), on the other hand, would be applied directly to the raw data without prior filtration. Generally, HS exhibits a relatively poor performance since it violates the i.i.d. assumption that underlies this method. Examples of its applications can be found in Righi and Ceretta (2015) and Patton et al. (2019). Nonetheless, unconditional HS is likely the most commonly used method by

banks due to its simplicity and the relatively smooth risk measures it provides, which do not lead to rapid changes in capital requirements, see Pérignon and Smith (2010). According to the European Banking Authority (2021), 72% of the participating banks reported to use some form of historical simulation for VaR computation, without differentiating between filtered and unfiltered HS.

### EVT

We consider two estimation methods, the GPD and Hill approach, which are based on extreme value theory (EVT). EVT is a branch of statistics that specifically focuses on modeling the tail regions of distributions rather than the entire distribution. EVT provides important results regarding limiting distributions for extreme observations in large samples. Due to space constraints, we provide only a concise overview of EVT and the two estimation methods. For a detailed exploration of EVT, refer to McNeil et al. (2015) or Daníelsson (2011) for an intuitive introduction.

Consider the sequence of i.i.d. random variables $X_1, ..., X_T$ and its maximum $M_T = \max(X_1, ..., X_T)$. The fundamental result of classical EVT is that the limiting distribution for normalized maxima from samples of i.i.d. random variables is in the family of the generalized extreme value (GEV) distribution with distribution function $H_\xi(x)$.[12] In this case, it is commonly stated that the random variable $X$ with distribution function $F$ is in the maximum domain of attraction of an extreme value distribution, denoted as $F \in \text{MDA}(H_\xi)$. The value of the parameter $\xi$ plays a key role in EVT analysis, since for $F \in \text{MDA}(H_\xi)$, the tails of the distribution fall into the categories Fréchet when $\xi > 0$, Gumbel when $\xi = 0$ or Weibull when $\xi < 0$, irrespective of the shape of the distribution function $F$. In the context of stock returns, the first category is particularly relevant, as $\xi > 0$ implies heavy tails where the tails decline by a power law at rate $1/\xi$. The parameter $\xi$ is referred to as shape parameter and is the reciprocal of the tail index $\gamma$, see also Definition 4.2.2.

The block maxima method, which involves dividing a sample into blocks and using the respective corresponding block maxima to estimate the GEV distribution is considered inefficient in terms of data utilization. Therefore, modern EVT analysis focuses on large observations that exceed a particular threshold, known as peaks over threshold (POT). Similar to how the GEV distribution represents the limiting distribution of normalized maxima, the generalized Pareto (GP) distribution is the limiting distribution of normalized data exceeding a threshold. When the POT method is applied in practice,

[12] See, e.g., McNeil et al. (2015, p. 136).

an appropriate threshold $u$ must be chosen, which is usually set to an upper order statistic, since EVT methods are typically applied to the right tail of a distribution. The choice of the threshold involves a bias-variance tradeoff. A higher value for $u$ increases the variability in the estimation of the parameters of the GP distribution, while a lower value of $u$ leads to bias, since the GP distribution assumption is only valid in the tails.

We consider two estimators based on the POT approach, the GPD and Hill estimator. Since EVT methods rely on the i.i.d. assumption, the estimation methods should be applied to appropriately filtered data, such as the standardized residuals $Z$ (McNeil and Frey, 2000). For a sufficiently large threshold $u$, the threshold exceedences $Z - u$ follow a GP distribution, i.e.

$$Z - u | Z > u \sim \mathrm{GP}(\beta, \xi) \tag{4.25}$$

with shape parameter $\xi$ and $\beta > 0$ denotes a scale parameter, see e.g. Nolde and Ziegel (2017). For the GPD estimator, the two parameters are estimated by fitting a GPD model to the excess losses in each bootstrap sample using MLE. McNeil et al. (2015) derive the formulas for VaR and ES, which are expressed as

$$\widehat{VaR}_\alpha^{GPD} = u + \frac{\hat{\beta}}{\hat{\xi}} \left( \left( \frac{1-\alpha}{k/T} \right)^{-\hat{\xi}} - 1 \right) \quad \text{and} \tag{4.26}$$

$$\widehat{ES}_\alpha^{GPD} = \frac{\widehat{VaR}_\alpha^{GPD}}{1-\hat{\xi}} + \frac{\hat{\beta} - \hat{\xi} u}{1-\hat{\xi}}, \tag{4.27}$$

where the threshold $u$ is set to the $(k+1)$-upper order statistic in each bootstrap sample and $k$ denotes the number of threshold exceedences. As mentioned previously, the selection of the threshold is an important task in practice. We adopt the approach used in Nolde and Ziegel (2017) and use the 12% most extreme observations for the GPD estimator. For instance, if the sample contains $T = 1000$ observations, the number of threshold exceedances used for fitting the GPD model would be $k = 120$.

The approach of Hill (1975) offers an alternative way to estimate the shape parameter or tail index. When using the Hill method, it is assumed that the underlying distribution is in the MDA of the Fréchet distribution and is therefore heavy-tailed. This means that the tail of the distribution is regularly varying and can be represented as the product of a power function and a slowly varying function, see Section 4.2.2 for details. Based on the assumption that the slowly varying function is constant above a threshold $u^{(H)}$, the

Hill estimator for the shape parameter corresponds to

$$\hat{\xi}^{(H)} = \frac{1}{\hat{\gamma}^{(H)}} = \frac{1}{k^{(H)}} \sum_{i=1}^{k^{(H)}} \ln Z_{(1)} - \ln Z_{(k+1)}, \tag{4.28}$$

where $Z_{(1)} \geq Z_{(2)} \geq ... \geq Z_{(k^{(H)}+1)}$ represent the descending order statistics of the bootstrap sample $Z_b$ and $k^{(H)}$ denotes the number of observations considered in the tail for the Hill approach. Embrechts et al. (1997) present various derivation methods for the estimator of $\xi^{(H)}$ (or $\gamma^{(H)}$, respectively).

The derivation of the risk measures VaR and ES is based on the Hill tail estimator, see Embrechts et al. (1997, p. 331 ff.) and McNeil et al. (2015, p. 160). Since the slowly varying function is considered as a constant $c$, the tail distribution is of the form $\bar{F}(z) = cz^{-1/\xi}$ for $z > u$. By using $\bar{F}(u) = cu^{-1/\xi}$, the constant can be represented as $c = \bar{F}(u)u^{1/\xi}$. The expression for the right tail of the distribution follows as

$$\bar{F}(z) = \bar{F}(u) \left(\frac{z}{u}\right)^{-1/\xi}.$$

The Hill tail estimator is obtained by using $\hat{\xi}^{(H)}$ as the shape parameter, $u^{(H)} = Z_{(k^{(H)}+1))}$ as the threshold and replacing $\bar{F}(u)$ with the empirical estimator $k/T$. The VaR at level $\alpha$ is obtained by inverting the Hill tail estimator,

$$\widehat{\mathrm{VaR}}_\alpha^{\mathrm{Hill}}(Z) = Z_{k^{(H)}+1} \left(\frac{1-\alpha}{k/T}\right)^{-\hat{\xi}^{(H)}}. \tag{4.29}$$

The ES is derived using (4.29) and the definition of ES as an integral, which results in

$$\widehat{\mathrm{ES}}_\alpha^{\mathrm{Hill}}(Z) = \frac{Z_{k^{(H)}+1}}{1-\hat{\xi}^{(H)}} \left(\frac{1-\alpha}{k/T}\right)^{-\hat{\xi}^{(H)}} = \frac{\widehat{\mathrm{VaR}}_\alpha^{\mathrm{Hill}}(Z)}{1-\hat{\xi}^{(H)}}. \tag{4.30}$$

From equation (4.30) follows that the Hill-based ES estimator can be viewed as a type of ratio estimator. Similar to the t-ratio estimator, the ES is obtained by multiplying the VaR by a parameter that depends on the thickness of the tail. However, in contrast to the t-ratio estimator, the Hill-based ES estimator implicitly also depends on the number of exceedances $k^{(H)}$. Additionally, the t-ratio estimator can accommodate scenarios where the distribution is not very heavy-tailed with correspondingly higher degrees of freedom.

In our application, we choose the 5% most extreme observations for the Hill estimator, hence $k^{(H)} = 50$ in a sample of $T = 1000$. Simulation evidence from McNeil et al. (2015)

suggests that the Hill estimator performs well for relatively small values of $k^{(H)}$, that is $20 - 75$ observations in sample of 1000, and can outperform the GPD (and FHS) estimator for distributions with well-behaved regularly varying tails. However, the GPD estimator seems to be less sensitive to the threshold choice and is more flexible, since it is also applicable to non-heavy-tailed data.

#### ES-normal-ratio

For the ES normal-ratio (n-ratio) estimator, the three VaR estimators $\widehat{\text{VaR}}_\alpha^{\text{FHS}}(Z)$, $\widehat{\text{VaR}}_\alpha^{\text{GPD}}(Z)$ and $\widehat{\text{VaR}}_\alpha^{\text{Hill}}(Z)$ are multiplied with the constant in (4.17). For the normal ratio, the underlying assumption is that the ratio is constant for a given $\alpha$. The corresponding multipliers can be found in the last row of Table 4.1. Since the ratios are based on the normal distribution and do not account for heavy tails, it is expected that the ES will be underestimated in most cases for stock returns.

#### ES-t-ratio

For the t-ratio estimator, the three VaR estimators are multiplied by the ratio in (4.18), where the degrees of freedom are replaced by the estimator $\hat{\nu}$, which is estimated for each bootstrap sample and each asset individually. This ratio estimator is a relatively flexible model as it can capture various degrees of tail-heaviness through the degrees of freedom. Additionally, since an individual ratio is estimated for each asset, this approach can also be applied to portfolios consisting of a single asset.

#### ES-p-ratio

To obtain a constant ratio for all assets within each bootstrap sample, which incorporates data information, the constant p-ratio based on the squared differences between ES and VaR estimates is considered. For a dataset with $i = 1, ..., N$ stocks, the estimator for the ratio is obtained by

$$\min_p \left( \sum_{i=1}^{N} \widehat{\text{ES}}_{\alpha,i}^{FHS} - p\widehat{\text{VaR}}_{\alpha,i}^{FHS} \right)^2 . \tag{4.31}$$

Similar to the normal ratio, the VaR values of all assets are multiplied by the same constant $p$ to obtain estimates for ES. However, in contrast to the normal ratio, $p$ is estimated based on data. Since stock returns often exhibit heavier tails than the normal distribution, the p-ratio will generally be higher than the normal ratio. Additionally,

the p-ratio is recalculated for each bootstrap sample. In this application, nonparametric estimates of ES and VaR are used to compute $p$ due to its widespread use in practice.

The p-ratio approach is conceptually similar to the probability equivalent level of VaR and ES (PELVE) approach introduced by Li and Wang (2023). The authors establish a relationship between ES and VaR based on the probability level $1 - \epsilon = \alpha$ and solve for some (single) asset the equation

$$\widehat{\mathrm{ES}}_{1-c\epsilon}^{FHS} = \widehat{\mathrm{VaR}}_{1-\epsilon}^{FHS}, \tag{4.32}$$

for the constant $c \in [1, 1/\epsilon]$. It is important to note that Li and Wang (2023) consider PELVE as a distributional index that increases as the distribution exhibits heavier tails, rather than using it for estimating ES. An alternative approach to estimate ES would be to consider the modified relationship

$$\widehat{\mathrm{ES}}_{1-\epsilon}^{FHS} = \widehat{\mathrm{VaR}}_{1-\epsilon/c}^{FHS}. \tag{4.33}$$

The p-ratio approach in (4.31) is based on the relationship between ES and VaR at the same level of confidence $\alpha$. Since $1 - \epsilon/c \geq \alpha$, the estimation of ES in (4.33) is based on a less accurate VaR estimate compared to $\mathrm{VaR}_\alpha$, as lower confidence levels enhance estimation accuracy. The results of an (unreported) simulation study with t-distributed data indicate that ES estimates based on (4.31) exhibit a substantially lower MSE than ES estimates based on (4.33). This serves as the motivation for using (4.31) to estimate the p-ratio in this application.

#### ES-i-ratio

The assumption of an identical ratio across all assets, as it is presumed in the case of the normal and p-ratio, is highly restrictive. For the ES-i-ratio, the 760 companies in the dataset are categorized into 11 different industries according to their Standard Industrial Classification (SIC) code. In order to estimate branch-specific ratios, the formula in (4.31) is used for each industry.

### 4.4.3. Results

In this section, the results of the bootstrap experiment are presented. For VaR, the performance of the three benchmark estimation methods is evaluated. For ES, the n-ratio, t-ratio, p-ratio and i-ratio are considered for all three benchmark methods. In

total, three models are compared for VaR and 15 models are compared for ES.[13] The performance measures for VaR and ES include the average values of bias, variance and MSE for the 760 stocks in the dataset. For each estimation method, the values of bias, variance and MSE are calculated with respect to the *true* risk measures for each of the 760 assets and then average values are computed. The *true* values are based on the empirical distribution of the complete filtered time series, as described in Section 4.4.1 Tables 4.3 - 4.6 present the average values for bias, variance and MSE.[14] Additionally, the average rank for the absolute bias, variance and MSE is provided. Each method is assigned a rank for each asset and an average value is computed for all assets. Since 15 estimation methods are considered for ES, the average rank ranges from 1 (best) to 15 (worst). For VaR, the average rank ranges from 1 to 3.

Tables 4.3 and 4.4 show the performance of the estimation methods for $\text{VaR}_{97.5\%}$ and $\text{ES}_{97.5\%}$ for an estimation period of $T = 1000$. For VaR, the FHS method exhibits the lowest bias, while the EVT-based methods have a slightly lower MSE due to a lower variance. Consequently, the FHS method achieves the lowest rank for bias, while the EVT-based methods obtain lower ranks for variance and MSE. However, the overall performance measured by MSE is relatively similar among the three benchmark methods. The ranking of benchmark estimators is reversed for ES. FHS exhibits the lowest MSE, closely followed by the GPD estimator. Hill performs worst among the benchmark estimators in terms of all performance measures and average ranks. The Hill estimator is appropriate only when the distribution is actually fat-tailed, since it assumes that the underlying distribution is in the MDA of the Fréchet distribution. Therefore, $\xi^{(H)}$ is confined to positive values. However, the fat-tailed assumption is not adequate in every case. In comparison, the GPD estimates a negative value for $\xi$ in approximately 20% of all estimations.[15] For ES-Hill, this appears to be a larger problem than for VaR-Hill, since the ES estimator divides the VaR value by $1 - \xi^{(H)}$, which can lead to the positive bias. In general, it can be observed that the variance and MSE for the ES estimators are higher than for VaR estimation at the same level of $\alpha$, which is consistent with the simulation results.

[13] The results in this section do not include the fully-parametric (MLE) approach to avoid making the list of models, especially for ES, too extensive. First, fully-parametric models are not used frequently in practice (according to European Banking Authority (2021), 6% of the banks use a fully-parametric approach for VaR calculation). Second, these models are not among those that pass the backtests conducted in Kellner and Rösch (2016). The computations for this study also do not indicate improved performance of the fully-parametric models. In Section 4.5, however, the FP-t approach is considered as a benchmark model.

[14] Note that $\overline{\text{variance}} + \overline{\text{bias}}^2 \neq \overline{\text{MSE}}$, as we consider average values of bias, variance and MSE.

[15] In total, there are $1,520,000$ estimations for 760 assets times 2000 bootstrap samples.

Table 4.3.: Performance results for VaR 97.5%

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | −0.0010 | 1.30 | 0.0150 | 2.64 | 0.0152 | 2.26 |
| GPD | 0.0301 | 2.30 | 0.0125 | 2.12 | 0.0145 | 1.96 |
| Hill | −0.0311 | 2.40 | 0.0115 | 1.25 | 0.0137 | 1.78 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 97.5% VaR with 1000 periods and 2000 replications.

Table 4.4.: Performance results for ES 97.5%

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | −0.0152 | 2.40 | 0.0933 | 13.31 | 0.0936 | 8.45 |
| GPD | −0.0041 | 2.06 | 0.0945 | 13.32 | 0.0951 | 8.64 |
| Hill | 0.0755 | 6.56 | 0.1021 | 14.51 | 0.1124 | 10.59 |
| n-FHS | −0.4804 | 13.54 | 0.0214 | 3.50 | 0.2958 | 13.48 |
| n-GPD | −0.4432 | 12.56 | 0.0177 | 2.13 | 0.2504 | 12.30 |
| n-Hill | −0.5163 | 14.59 | 0.0163 | 1.26 | 0.3163 | 14.51 |
| t-FHS | −0.0514 | 6.52 | 0.0513 | 11.74 | 0.0736 | 7.16 |
| t-GPD | −0.0060 | 5.59 | 0.0485 | 11.23 | 0.0630 | 5.78 |
| t-Hill | −0.0925 | 7.19 | 0.0456 | 9.99 | 0.0676 | 6.19 |
| p-FHS | −0.0051 | 9.17 | 0.0313 | 7.71 | 0.0827 | 7.33 |
| p-GPD | 0.0398 | 8.73 | 0.0260 | 5.89 | 0.0693 | 5.88 |
| p-Hill | −0.0485 | 8.26 | 0.0240 | 4.29 | 0.0641 | 4.93 |
| i-FHS | −0.0219 | 7.96 | 0.0322 | 8.51 | 0.0714 | 6.09 |
| i-GPD | 0.0233 | 7.61 | 0.0275 | 7.13 | 0.0592 | 4.78 |
| i-Hill | −0.0643 | 7.26 | 0.0256 | 5.50 | 0.0579 | 3.89 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 97.5% ES with 1000 periods and 2000 replications.

Turning to the ratio estimators for $\text{ES}_{97.5\%}$ estimation, the n-ratio estimators have the lowest variance among all considered models, but also by far the highest absolute bias and MSE. The low variance results from multiplying the low-variance $\text{VaR}_{97.5\%}$ benchmark estimators by a relatively small constant that is displayed in Table 4.1. However, this small constant is simultaneously responsible for the substantial negative bias observed for the n-ratio estimators, since it is too small for capturing the characteristics of fat-tailed distributions, leading to an underestimation of $\text{ES}_{97.5\%}$.

The other ratio estimators exhibit a lower variance and MSE in comparison to the benchmark models, but with higher absolute bias. The p-ratio estimator employs the same multiplier for each asset, which is determined from the data, in contrast to the n-ratio estimator. This simple approach reduces the MSE compared to the benchmark

methods. The decrease in MSE is more pronounced for the EVT-based estimators reflecting that the VaR estimates are better for these models. It is an interesting observation that this restrictive ratio model leads to an improved MSE and average ranks in such a diverse investment universe consisting of 760 different stocks. However, the relationship between ES and VaR does not seem to correspond to a single natural constant, since subdividing the assets into industry sectors based on their SIC codes further improves the MSE, as evidenced by the results for the i-ratio models. The MSE is lowest for the i-Hill estimator, which results from the low MSE in the VaR-Hill estimation.

Figure 4.1 shows the distribution of the estimated p-ratios and i-ratios for each of the 11 considered industries for all 2000 bootstrap samples. While the p-ratios are quite narrowly distributed in the range of 1.38 and 1.48, the industry ratios show distributions around considerably lower ratios (e.g., Mining) as well as substantially higher ratios (e.g., Services). The models based on i-ratios exhibit the lowest MSE and seem to gain an advantage from avoiding the computation of individual ratios for each asset - a process that can lead to estimation errors in the t-ratio method. On the other hand, the i-ratio models do not have the restrictive assumption of the p-ratio models that the ratios for all assets are identical.[16]

In terms of MSE, the performance of the t-ratio estimators for FHS and GPD falls in between their p-ratio and i-ratio counterparts. The average ranks for absolute bias improve in comparison to the other ratio estimators but are higher than those of the benchmark models. Nevertheless, the MSE is considerably reduced as compared to the benchmark models: for t-FHS by approximately 21% compared to FHS, for t-GPD by nearly 34% compared to GPD and for t-Hill by almost 40% compared to Hill. The variance of the t-ratio estimators is higher than that of the other ratio estimators, since the ratio is estimated individually for each asset and depends on the estimated degrees of freedom. However, the variance is considerably lower as compared to the benchmark estimators with reductions ranging from 45% to 55%. The t-ratio approach is particularly interesting, as it can also be used for ES estimation for individual portfolios.

Tables 4.5 and 4.6 show the performance of the estimators for VaR and ES at the 99% confidence level. The comparative performance is similar to the 97.5% confidence level, even though the general levels of bias, variance and MSE are higher. It is worth noting

[16] As an alternative to grouping assets by industries, the assets could also be divided into groups using a data-driven method. For instance, the following two-step procedure could be implemented for this purpose. In a first step, the t-ratios for all assets are determined. In a second step, the vector of t-ratios can be split into $k$ clusters using $k$-means clustering. The resulting cluster centroids can then serve as ratio estimates for the $k$ clusters. However, some experiments indicate that this approach does not yield any further improvement with regard to MSE.

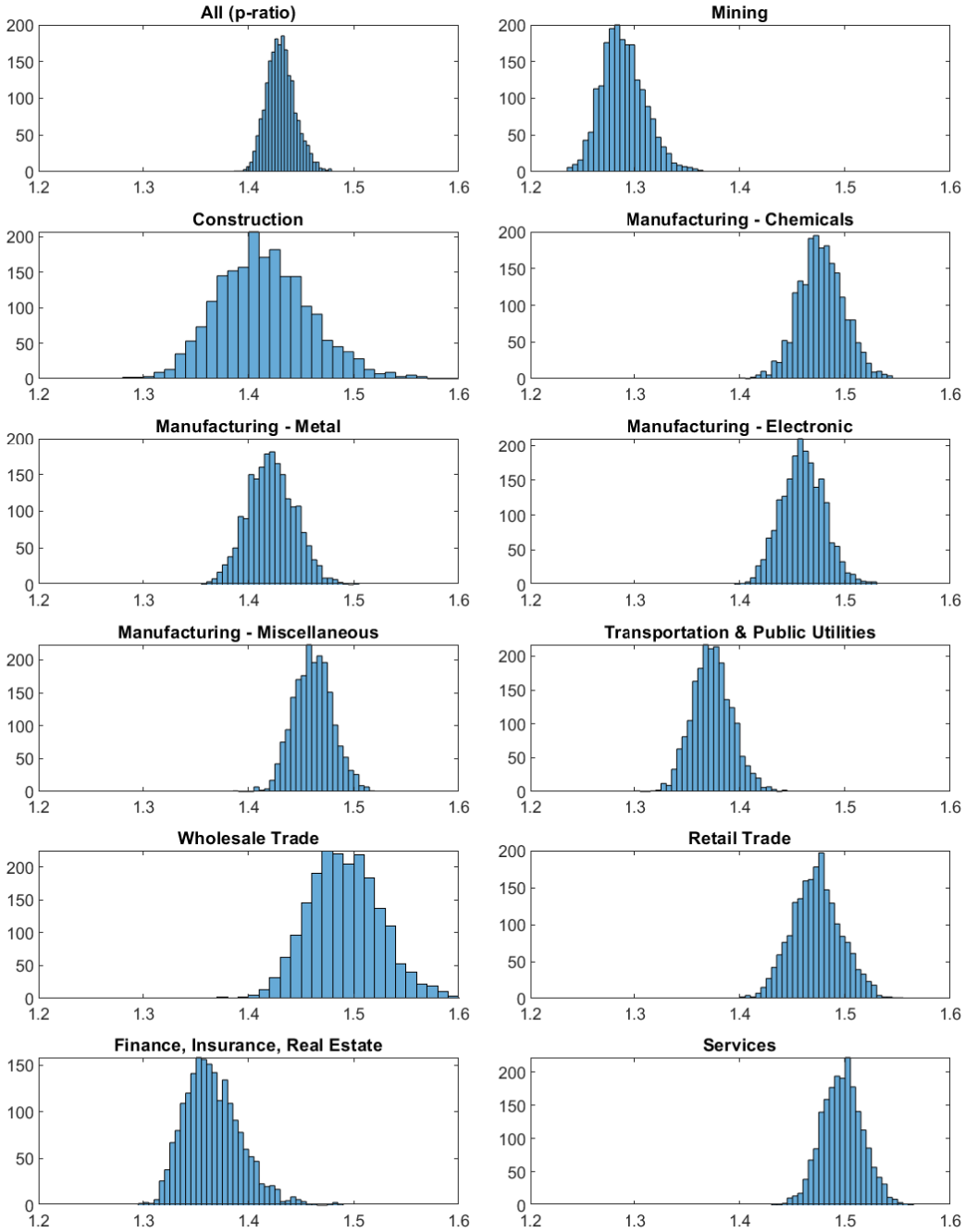



Figure 4.1.: p-ratios and i-ratios for the 2000 bootstrap samples

that for the benchmark models, the variance and MSE for $\text{VaR}_{99\%}$ are lower than for the $\text{ES}_{97.5\%}$. This observation is particularly interesting in light of the transition from $\text{VaR}_{99\%}$ to $\text{ES}_{97.5\%}$ as the regulatory risk measure under Basel III.

Table 4.5.: Performance results for VaR 99%

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | 0.0065 | 1.37 | 0.0686 | 2.75 | 0.0700 | 2.54 |
| GPD | 0.0635 | 2.54 | 0.0453 | 1.51 | 0.0570 | 1.84 |
| Hill | 0.0131 | 2.09 | 0.0465 | 1.75 | 0.0517 | 1.63 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 99% VaR with 1000 periods and 2000 replications.

Table 4.6.: Performance results for ES 99%

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | −0.0479 | 2.60 | 0.3806 | 13.45 | 0.3832 | 9.02 |
| GPD | −0.0910 | 3.68 | 0.3496 | 13.21 | 0.3656 | 8.80 |
| Hill | 0.2071 | 8.35 | 0.3683 | 14.37 | 0.4579 | 11.94 |
| n-FHS | −0.7440 | 13.33 | 0.0900 | 4.49 | 0.8034 | 13.27 |
| n-GPD | −0.6786 | 12.18 | 0.0594 | 1.54 | 0.6486 | 11.31 |
| n-Hill | −0.7364 | 13.51 | 0.0610 | 1.76 | 0.7443 | 12.92 |
| t-FHS | −0.1584 | 7.08 | 0.2027 | 12.03 | 0.3083 | 8.11 |
| t-GPD | −0.0784 | 5.31 | 0.1563 | 9.84 | 0.2145 | 4.95 |
| t-Hill | −0.1485 | 6.41 | 0.1561 | 10.07 | 0.2409 | 5.70 |
| p-FHS | −0.0520 | 9.05 | 0.1359 | 8.53 | 0.3025 | 7.81 |
| p-GPD | 0.0289 | 8.25 | 0.0895 | 4.55 | 0.2144 | 5.29 |
| p-Hill | −0.0426 | 8.23 | 0.0917 | 5.02 | 0.2329 | 5.27 |
| i-FHS | −0.0506 | 7.70 | 0.1429 | 9.09 | 0.2701 | 6.84 |
| i-GPD | 0.0321 | 7.19 | 0.0982 | 5.88 | 0.1941 | 4.45 |
| i-Hill | −0.0401 | 7.14 | 0.0995 | 6.18 | 0.2073 | 4.34 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 99% ES with 1000 periods and 2000 replications.

In comparison to $\text{ES}_{97.5\%}$, the t-ratio, p-ratio and i-ratio models display a more sizable reduction in MSE for $\text{ES}_{99\%}$ as compared to their benchmark counterparts in all cases except one. For instance, the MSE for p-GPD is roughly 27% lower than that of GPD for ES at the 97.5% level and it decreases by approximately 41% for $\text{ES}_{99\%}$. The reduction in MSE for i-GPD and t-GPD is around 9 and 8 percentage points, respectively. The Hill ratio estimators, p-FHS and i-FHS also exhibit higher percentage reductions in MSE for $\text{ES}_{99\%}$. The only exception is t-FHS, where the reduction remains relatively constant.

For the estimation at the 99% confidence level, the EVT-based methods appear to provide more stable estimates in comparison to FHS. With an estimation period of 1000 data points, the FHS estimation relies on only a few observations in the tail, which makes the estimation uncertain and susceptible to the influence of individual data points. In contrast, EVT-based methods fit a smooth function to the tail, which helps to stabilize the risk measure estimates. For $\text{VaR}_{99\%}$, despite having a higher bias, the EVT-based methods exhibit a lower MSE than FHS due to a lower variance. The data-driven ratio estimators for GPD and Hill benefit from this lower variance, showing improved performance compared to their FHS counterparts. Among these estimators, the i-GPD estimator has the lowest MSE. For the data-driven EVT ratio estimators, the average absolute bias is lower compared to their benchmark counterparts, while the average ranks are higher. Similar to $\text{ES}_{97.5\%}$, the Hill estimator has a relatively large bias.

Tables C.1-C.4 in Appendix C show the performance of the risk measures for an estimation period of $T = 500$. In general, the absolute bias, variance and MSE are considerably higher compared to an estimation period of $T = 1000$. However, the key findings remain the same: VaR estimation is more accurate than ES estimation and the data-driven ratio models offer a considerable stabilization of ES estimates with a substantially lower MSE. Notably, the GPD model exhibits a high variance for ES when $T = 500$. As with $T = 1000$, 12% of the data are used as exceedances for the GPD model. However, in some instances, this appears to be insufficient for the ML optimization to converge. The performance of GPD improves when, for instance, 20% of the data are selected as exceedances for $T = 500$, resulting in a reduction of the variance from 0.7045 to 0.1613 for $\text{ES}_{97.5\%}$. Nevertheless, this highlights a practical drawback of EVT-based models, as they necessitate the selection of a threshold. The EVT-based ratio models are less affected by this issue, since they rely on the performance of VaR estimation. Estimating ES at high confidence levels with with short estimation periods is particularly challenging. The results in Table C.4 show that the benchmark models are even inferior to the n-ratio models in two cases, while the data-driven ratio models display a considerable reduction in MSE as compared to the benchmark models.

## 4.5. Performance based on scoring functions

The objective of this application is to investigate the out-of-sample performance of several VaR and ES models for portfolios based on the scoring function in (4.7). In practice, it is relevant to be able to compare the performance of risk measure estimates for individual assets, such as portfolios. The results are based on the same dataset as in Section 4.4,

from which the returns of the well-known $1/N$ (naive) portfolio and the mean-variance portfolio with short-sale constraints are obtained. Additionally, we examine the scores for the three prominent single-asset portfolios McDonald's, Nvidia and Bank of America.

### 4.5.1. Application setup and portfolios

To obtain forecasts for the risk measures, a rolling-window procedure is applied. For an estimation period of 1000 daily data points, an AR(1)-GARCH(1,1) model with $t$-innovations is estimated and the forecasts of the risk measures for time period $t+1$ are obtained as

$$\widehat{\mathrm{VaR}}_\alpha(L_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\widehat{\mathrm{VaR}}_\alpha(Z_{t+1}) \quad \text{and}$$
$$\widehat{\mathrm{ES}}_\alpha(L_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1}\widehat{\mathrm{ES}}_\alpha(Z_{t+1}).$$

For example, the first 1000 periods are used to produce a forecast for $t = 1001$. Then, the estimation window shifts to $2:1001$ to produce a forecast for $t = 1002$ and so on. The out-of-sample period covers the time span from December 26, 2003 to November 14, 2019. However, for the mean-variance portfolio with short-sale constraints, the out-of-sample period starts later on December 17, 2007, since the first time periods are used for estimating the portfolio weights. For comparison purposes, a shorter estimation period of 500 is also considered, while maintaining the same out-of-sample time span.

For the simultaneous forecast evaluation of the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, we use the scoring function in (4.7). This requires a choice for $G_1$ and $G_2$. We follow Nolde and Ziegel (2017, eq. 2.24) and choose $G_1(x) = 0$ and $G_2(x) = 1/x$ so that $\mathcal{G}(x) = \ln(x)$, see also Patton et al. (2019, eq. 6). This yields the scoring function

$$S(v, e, l) = \mathbb{1}(l > v)\frac{l - v}{e} + (1 - \alpha)\left(\frac{v}{e} - 1 + \ln(e)\right). \tag{4.34}$$

As pointed out by Nolde and Ziegel (2017) and Patton et al. (2019), this choice of the scoring function is appealing since it generates loss differences between competing scores that are homogeneous of degree zero in the relevant tail region. Patton and Sheppard (2009) show that this property improves the power of Diebold and Mariano (1995) tests in the context of volatility forecasts.

In this application, five portfolios with different risk profiles are considered. Firstly, we examine two diversified portfolios: the $1/N$ portfolio and the mean-variance portfolio with short sale constrains ($\mathrm{MV}_+$). The $1/N$ portfolio does not require the estimation of portfolio weights, since an equal weight of $1/N$ is invested in each asset in a potential in-

vestment universe consisting of $N$ assets. The weights of the $\text{MV}_+$ portfolio are obtained by solving the classical mean-variance problem subject to the constraint that the portfolio weights are nonnegative. Empirical findings from numerous studies demonstrate that the $\text{MV}_+$ portfolio exhibits significantly improved out-of-sample performance and more stable weights as compared to the mean-variance portfolio without constraints, see e.g. Jagannathan and Ma (2003). In addition to the diversified portfolios, the stocks of McDonald's (MCD), Nvidia (NVDA) and Bank of America (BAC) are considered as single-asset portfolios. These three companies belong to different industries and have a high level of recognition.

Table 4.7 shows that the returns of the portfolios under analysis exhibit different risk profiles. The two diversified portfolios have the lowest standard deviation and the smallest maximum loss over the observed time span. MCD is included in the analysis because it has a relatively defensive risk profile that is more similar to that of the diversified portfolios. In contrast, NVDA and BAC exhibit significantly higher volatility over the observed time period. For example, the standard deviation of NVDA is more than three times higher than that of the $MV_+$ portfolio. Moreover, NVDA and BAC also experience notable instances of both substantial losses and gains, as reflected in their maximum daily losses and returns. The kurtosis for all portfolios is higher than normal. It may be surprising that the single-asset portfolios exhibits positive skewness. However, Albuquerque (2012) finds a difference in the sign of the skewness between returns at the aggregate stock market level and the individual firm level. While negative skewness is commonly observed for aggregate stock market returns, individual stock returns often exhibit positive skewness, see also Jondeau et al. (2019).

Table 4.7.: Descriptive statistics for the five portfolios

| Method | naive | $\text{MV}_+$ | MCD | NVDA | BAC |
|---|---|---|---|---|---|
| mean | 0.0602 | 0.0484 | 0.0418 | 0.1527 | 0.0465 |
| stddev. | 1.3167 | 1.1766 | 1.4382 | 3.8325 | 2.8779 |
| skewness | −0.1276 | −0.2471 | 0.0336 | 0.6187 | 0.9071 |
| kurtosis | 9.1079 | 6.8539 | 9.4737 | 16.1463 | 31.1804 |
| min | −10.3830 | −7.6461 | −12.8170 | −35.2335 | −28.9694 |
| max | 10.8257 | 9.4200 | 9.3895 | 42.4145 | 35.2691 |

### 4.5.2. Estimators

To assess the performance of risk measures for individual assets, a total of 11 estimation methods are examined. These methods include the (unfiltered) HS and fully-parametric-t (FP-t) as additional benchmarks, which are known to perform worse than, for example,

FHS in most cases. However, according to Patton et al. (2019), it is difficult to distinguish good models from each other using scoring functions, but it is possible to distinguish the worst models from the good ones. The analysis also includes the baseline models FHS, GPD and Hill.

Two types of ratio models are considered. Firstly, a completely naive model is assessed, where the VaR estimates from FHS, GPD and Hill are simply multiplied by a factor of 1.4 to obtain 97.5% ES estimates. Although this factor is not based on a rigorous estimation, it roughly corresponds to the average of all ratios considered in Section 4.4. As shown in Table 4.1, this already represents a ratio suitable for distributions with relatively heavy tails when $\alpha = 97.5\%$. Hence, this naive approach will overestimate the ratio for well-diversified portfolios such as the $1/N$ strategy. Nonetheless, it is interesting to compare the scores of such a completely naive (and partially incorrect) approach with sophisticated methods for estimating ES. For estimating $\text{ES}_{99\%}$, the simple ratio is reduced to 1.35. To include a data-driven ratio estimator, the t-ratio estimator is also considered, which provides an individual ratio for each time point and asset. The t-ratio estimator is obtained by multiplying the VaR values from FHS, GPD and Hill with the corresponding t-ratios.

In this application, the p-ratio and industry-specific ratio are not considered, since estimating these ratios requires a dataset consisting of multiple assets. While determining the p-ratio from the five assets would be feasible, the focus here is to specifically examine estimation methods applicable in the single-asset case. To save space, the poorly performing n-ratio model is not further explored.

### 4.5.3. Results

Table 4.8 shows the average scores for $(\text{VaR}_{97.5\%},\text{ES}_{97.5\%})$, which are determined as the average out-of-sample scores using the joint scoring function in (4.34). As described in Section 4.2.3, the true values of the risk measures minimize the expected value of the scoring function, hence lower average scores in Table 4.8 indicate better performance. The ranks of the individual methods are presented in parentheses, while the last column of Table 4.8 displays the average rank. Evaluating ES forecasts using the joint scoring function is not straightforward for two reasons. Firstly, scores for good models tend to be close to each other (Patton et al., 2019). Secondly, VaR and ES forecasts are evaluated simultaneously. However, the considered ratio models employ the same VaR forecasts as their benchmark counterparts, thus differences in scores arise from the ES forecasts.

The scores of the unconditional HS estimator stand out as particularly poor and rank last for all portfolios. This confirms that inferior models can be effectively distinguished

Table 4.8.: Average scores and ranks for $(\mathrm{VaR}_{97.5\%}, \mathrm{ES}_{97.5\%})$

| Method | naive | | MV+ | | MCD | | NVDA | | BAC | | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | 1.3588 | (11) | 1.3009 | (11) | 1.2075 | (11) | 2.1954 | (11) | 2.1848 | (11) | 11 |
| t-FP | 1.0326 | (9) | 1.0960 | (3) | 1.1662 | (10) | 2.1277 | (10) | 1.6274 | (10) | 8.4 |
| FHS | 1.0281 | (2) | 1.0950 | (2) | 1.0971 | (9) | 2.0182 | (7) | 1.5887 | (9) | 5.8 |
| GPD | 1.0295 | (4) | 1.0962 | (5) | 1.0946 | (8) | 2.0189 | (8) | 1.5869 | (7) | 6.4 |
| Hill | 1.0300 | (5) | 1.0972 | (6) | 1.0942 | (7) | 2.0175 | (6) | 1.5884 | (8) | 6.4 |
| s-FHS | 1.0300 | (6) | 1.0977 | (7) | 1.0912 | (3) | 2.0121 | (3) | 1.5809 | (1) | 4.0 |
| s-GPD | 1.0317 | (7) | 1.1004 | (10) | 1.0899 | (2) | 2.0131 | (4) | 1.5817 | (4) | 5.4 |
| s-Hill | 1.0325 | (8) | 1.0986 | (9) | 1.0898 | (1) | 2.0197 | (9) | 1.5829 | (5) | 6.4 |
| t-FHS | 1.0256 | (1) | 1.0949 | (1) | 1.0942 | (6) | 2.0073 | (1) | 1.5814 | (2) | 2.2 |
| t-GPD | 1.0282 | (3) | 1.0961 | (4) | 1.0916 | (4) | 2.0082 | (2) | 1.5816 | (3) | 3.2 |
| t-Hill | 1.0326 | (10) | 1.0983 | (8) | 1.0925 | (5) | 2.0137 | (5) | 1.5841 | (6) | 6.8 |

This table shows the scaled average scores for the estimation of $(\mathrm{VaR}_{97.5\%}, \mathrm{ES}_{97.5\%})$ with a rolling window of 1000 time periods. The corresponding ranks are presented in brackets. The last column contains the average rank for each estimation method.

from better models for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ forecasts using the joint scoring function. The t-FP methods also exhibits higher scores in most cases and has the second-worst average rank. The best average ranks are achieved by t-FHS and t-GPD, which display lower scores in comparison to their benchmark counterparts for all portfolios. The t-Hill estimator exhibits lower scores only for single stock portfolios compared to Hill. The naive simple-ratio (s-ratio) models perform impressively well for the single-stock portfolios and surprisingly outperform the benchmark models in most cases. For the diversified portfolios, the s-ratio models show slightly worse scores compared to the benchmark models. It is, however, remarkable that the scores of the s-ratio models, based on the results of Diebold-Mariano tests, do not differ significantly from the benchmark models, even though the simple ratio of 1.4 is likely set too high for well-diversified portfolios (cf. Tables 4.1 and 4.9).

Diebold-Mariano (DM) tests can be used to evaluate whether the average scores differ significantly from each other. For illustration, Tables 4.9 and 4.10 show DM t-statistics for the naive and BAC portfolios. The t-statistics for $\mathrm{MV}_+$, MCD and NVDA are provided in Tables C.7-C.9 in Appendix C. The results are presented in a similar manner as in Patton et al. (2019): Positive t-statistics suggest that the column model is superior to the row model. A t-value with an absolute value greater than 1.96 indicates a significant difference in average scores at a 95% confidence level. As noted by Patton et al. (2019), detecting statistically significant differences between sophisticated models can be challenging. This is supported by the results for the naive portfolio in Table 4.9.

Table 4.9.: Diebold-Mariano t-statistics for the naive portfolio, ES 97.5%

| Method | HS | t-FP | FHS | GPD | Hill | s-FHS | s-GPD | s-Hill | t-FHS | t-GPD | t-Hill |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | · | 4.65 | 4.75 | 4.71 | 4.73 | 4.77 | 4.73 | 4.72 | 4.81 | 4.75 | 4.67 |
| t-FP | −4.65 | · | 0.93 | 0.79 | 0.57 | 0.29 | 0.11 | 0.02 | 1.08 | 0.88 | −0.01 |
| FHS | −4.75 | −0.93 | · | −0.66 | −0.65 | −0.32 | −0.69 | −1.25 | 1.10 | 0.00 | −1.18 |
| GPD | −4.71 | −0.79 | 0.66 | · | −0.18 | −0.08 | −0.44 | −0.86 | 1.13 | 0.88 | −0.84 |
| Hill | −4.73 | −0.57 | 0.65 | 0.18 | · | −0.01 | −0.29 | −0.78 | 1.07 | 0.59 | −1.10 |
| s-FHS | −4.77 | −0.29 | 0.32 | 0.08 | 0.01 | · | −0.80 | −0.56 | 1.10 | 0.35 | −0.31 |
| s-GPD | −4.73 | −0.11 | 0.69 | 0.44 | 0.29 | 0.80 | · | −0.20 | 1.59 | 0.87 | −0.12 |
| s-Hill | −4.72 | −0.02 | 1.25 | 0.86 | 0.78 | 0.56 | 0.20 | · | 2.61 | 1.66 | −0.04 |
| t-FHS | −4.81 | −1.08 | −1.10 | −1.13 | −1.07 | −1.10 | −1.59 | −2.61 | · | −1.04 | −1.40 |
| t-GPD | −4.75 | −0.88 | 0.00 | −0.88 | −0.59 | −0.35 | −0.87 | −1.66 | 1.04 | · | −1.12 |
| t-Hill | −4.67 | 0.01 | 1.18 | 0.84 | 1.10 | 0.31 | 0.12 | 0.04 | 1.40 | 1.12 | · |

This table reports the Diebold-Mariano t-statistics comparing the average scores for the naive portfolio for $(\text{VaR}_{97.5\%}, \text{ES}_{97.5\%})$. Positive t-statistics show that the column model exhibits a lower average score than the row model.

Only the unconditional HS model displays significantly inferior average scores compared to all other models, and t-FHS performs significantly better than p-Hill. Although all other entries in the t-FHS column are positive, they are notably below 1.96. For the single-stock portfolios, the absolute t-values are considerably higher. For BAC, the s-ratio models as well as t-FHS and t-GPD frequently show statistically significant improvements compared to the benchmark models. The columns with ratio models feature positive t-values in the majority of the cases. HS is significantly worse in all cases and t-FP is significantly worse in almost all cases as compared to the other models.

Turning to risk measures at a higher $\alpha$-level, Table 4.11 presents the performance for $(\text{VaR}_{99\%},\text{ES}_{99\%})$. Similar to Section 4.4.3, it is observed that the performance of FHS and t-FHS deteriorates at this high confidence level, although t-FHS remains superior to FHS. The estimation of ES using FHS is quite uncertain for $\alpha = 0.99$. As observed in Section 4.4.3, the EVT methods demonstrate better performance, with t-GPD achieving the lowest average rank of 2.8. The s-ratio models exhibit lower average scores for the single-stock portfolios. Tables C.5 and C.6 in Appendix C contain the average scores for an estimation period of 500 days, showing qualitatively similar results. The ratio models outperform the benchmark models in most cases, in particular for the single-stock portfolios. For $\alpha = 0.99$ and an estimation period of 500, the s-ratio models display the lowest average ranks, as all types of estimation seem to entail significant uncertainty for ES. It can be concluded that t-FHS performs well for $\alpha = 97.5\%$ and a reasonable estimation period, while t-GPD consistently belongs to the better models with comparatively low average scores, irrespective of the confidence level or estimation period.

Table 4.10.: Diebold-Mariano t-statistics for the BAC portfolio, ES 97.5%

| Method | HS | t-FP | FHS | GPD | Hill | s-FHS | s-GPD | s-Hill | t-FHS | t-GPD | t-Hill |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | · | 5.91 | 6.50 | 6.55 | 6.52 | 6.56 | 6.59 | 6.57 | 6.54 | 6.58 | 6.55 |
| t-FP | −5.91 | · | 1.90 | 2.02 | 1.90 | 2.53 | 2.49 | 2.06 | 2.32 | 2.32 | 1.87 |
| FHS | −6.50 | −1.90 | · | 0.71 | 0.13 | 2.33 | 1.93 | 1.97 | 2.66 | 2.65 | 1.23 |
| GPD | −6.55 | −2.02 | −0.71 | · | −0.54 | 1.35 | 2.01 | 1.46 | 1.25 | 2.86 | 0.71 |
| Hill | −6.52 | −1.90 | −0.13 | 0.54 | · | 2.04 | 1.95 | 2.59 | 2.11 | 2.58 | 1.31 |
| s-FHS | −6.56 | −2.53 | −2.33 | −1.35 | −2.04 | · | −0.23 | −0.43 | −0.25 | −0.20 | −0.55 |
| s-GPD | −6.59 | −2.49 | −1.93 | −2.01 | −1.95 | 0.23 | · | −0.32 | 0.07 | 0.05 | −0.45 |
| s-Hill | −6.57 | −2.06 | −1.97 | −1.46 | −2.59 | 0.43 | 0.32 | · | 0.37 | 0.48 | −0.64 |
| t-FHS | −6.54 | −2.32 | −2.66 | −1.25 | −2.11 | 0.25 | −0.07 | −0.37 | · | −0.06 | −0.59 |
| t-GPD | −6.58 | −2.32 | −2.65 | −2.86 | −2.58 | 0.20 | −0.05 | −0.48 | 0.06 | · | −0.65 |
| t-Hill | −6.55 | −1.87 | −1.23 | −0.71 | −1.31 | 0.55 | 0.45 | 0.64 | 0.59 | 0.65 | · |

This table reports the Diebold-Mariano t-statistics comparing the average scores for the BAC portfolio for $(\text{VaR}_{97.5\%}, \text{ES}_{97.5\%})$. Positive t-statistics show that the column model exhibits a lower average score than the row model.

Table 4.11.: Average scores and ranks for $(\text{VaR}_{99\%}, \text{ES}_{99\%})$

| Method | naive | | MV+ | | MCD | | NVDA | | BAC | | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | 1.6254 | (11) | 1.5408 | (11) | 1.5054 | (11) | 2.4937 | (11) | 2.5418 | (11) | 11 |
| t-FP | 1.2206 | (7) | 1.2686 | (4) | 1.4199 | (10) | 2.4090 | (10) | 1.8566 | (10) | 8.2 |
| FHS | 1.2256 | (10) | 1.2742 | (8) | 1.3726 | (9) | 2.3309 | (9) | 1.8431 | (9) | 9.0 |
| GPD | 1.2217 | (8) | 1.2682 | (3) | 1.3646 | (6) | 2.3303 | (8) | 1.8266 | (6) | 6.2 |
| Hill | 1.2113 | (1) | 1.2716 | (6) | 1.3636 | (5) | 2.3223 | (5) | 1.8261 | (5) | 4.4 |
| s-FHS | 1.2183 | (6) | 1.2809 | (10) | 1.3670 | (7) | 2.3210 | (4) | 1.8336 | (7) | 6.8 |
| s-GPD | 1.2179 | (5) | 1.2728 | (7) | 1.3551 | (1) | 2.3200 | (3) | 1.8177 | (3) | 3.8 |
| s-Hill | 1.2117 | (2) | 1.2697 | (5) | 1.3580 | (2) | 2.3282 | (7) | 1.8160 | (1) | 3.4 |
| t-FHS | 1.2223 | (9) | 1.2753 | (9) | 1.3722 | (8) | 2.3170 | (2) | 1.8363 | (8) | 7.2 |
| t-GPD | 1.2168 | (4) | 1.2675 | (2) | 1.3596 | (3) | 2.3153 | (1) | 1.8180 | (4) | 2.8 |
| t-Hill | 1.2129 | (3) | 1.2649 | (1) | 1.3634 | (4) | 2.3238 | (6) | 1.8172 | (2) | 3.2 |

This table shows the scaled average scores for the estimation of $(\text{VaR}_{99\%}, \text{ES}_{99\%})$ with a rolling window of 1000 time periods. The corresponding ranks are presented in brackets. The last column contains the average rank for each estimation method.

## 4.6. Conclusion

The transition from 99% VaR to 97.5% ES as the regulatory risk measure according to the Basel Accords leaves the question whether the estimation of 97.5% ES provides practical value in quantifying risk. This paper examines the key theoretical differences between these two risk measures and compares their estimation performance through simulation studies and real-data applications. Additionally, it is explored whether ES can be well approximated by so-called ratio models, that involve to multiply the VaR with a factor above one.

ES is often considered the theoretically superior risk measure, since it is subadditive and incorporates tail risks. However, ES is not elicitable and less robust to model misspecifications and outliers in the dataset. Hence, there is a trade-off between subadditivity and sensitivity to extreme events on the one hand, and elicitability and robustness on the other. As a result, there is no consensus in the literature regarding the most suitable risk measure for regulatory purposes.

The results of the simulation study and bootstrap resampling application suggest that $\text{ES}_{97.5\%}$ is estimated with considerably higher variability than VaR, even when considering the higher confidence level of 99% for VaR. The results in this paper also show that the estimation of ES can be improved in terms of MSE by multiplying VaR by a suitable constant. Such ratio models benefit from the relatively accurate estimation of $\text{VaR}_{97.5\%}$, which is upscaled to obtain $\text{ES}_{97.5\%}$ estimates.

The applications in Sections 4.4 and 4.5 indicate that even rudimentary estimated ratios or rather naively chosen ratios can enhance the performance of the considered benchmark models FHS, GPD and Hill. Such ratios either improve the performance or achieve a comparable quality. However, the ES/VaR ratio does not seem to be a fixed constant. Ratio models that incorporate additional information about companies (such as the industry sector) or the distribution of data (such as tail heaviness) outperform the benchmark and simple ratio models.

Among the considered ratio models, the t-ratio model is particularly interesting as it can be applied to single-asset portfolios. For this model, the ratio is based on the degrees of freedom of the Student-t distribution, which reflects the heaviness of the tails. Notably, the t-ratio model based on the VaR from the GPD model (t-GPD) consistently ranks among the best-performing models, irrespective of the considered confidence levels and estimation periods.

In conclusion, the findings of this paper suggest that quantile-based risk measures improve estimation accuracy. In case accurate estimation of risk measures is an impor-

tant goal in regulation, this should be taken into account. Although ES theoretically accounts for tail risks, it remains uncertain how effectively an estimated ES quantifies these risks in practice due to higher estimation uncertainty. An alternative approach could involve upscaling the more accurately estimated $\text{VaR}_{97.5\%}$ with the scaling factor incorporating information on individual or market risk. Future research could further explore and develop ratio models that, for example, consider economic conditions.

# Appendix A.

# Appendix of Chapter 2

Details of the data generating processes used in Section 2.7.4

### DGP6, (Bai, 2009)

We consider the following model with two regressors, $k = 2$, and $r = 2$ unobserved factors:

$$y_{it} = \beta_1 x_{1,it} + \beta_2 x_{2,it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{it}, \tag{A.1}$$

with $\beta_1 = 1$, $\beta_2 = 3$, $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \lambda_{2,i})'$ and $\boldsymbol{f}_t = (f_{1,t}, f_{2,t})'$. The two regressors are generated as

$$x_{1,it} = \mu_1 + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \boldsymbol{\iota}' \boldsymbol{\lambda}_i + \boldsymbol{\iota}' \boldsymbol{f}_t + \eta_{1,it} \tag{A.2}$$

$$x_{2,it} = \mu_2 + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \boldsymbol{\iota}' \boldsymbol{\lambda}_i + \boldsymbol{\iota}' \boldsymbol{f}_t + \eta_{2,it} \tag{A.3}$$

with $\boldsymbol{\iota}' = (1, 1)$. Hence, both regressors are correlated with the loadings, the factors and the product of both. The unobserved factors and loadings follow standard normal distributions,

$$f_{j,t} \overset{iid}{\sim} \mathcal{N}(0, 1) \text{ for } j = 1, 2,$$
$$\lambda_{j,i} \overset{iid}{\sim} \mathcal{N}(0, 1) \text{ for } j = 1, 2,$$

where $j = 1, 2$ denotes the factor subscript. The regression error is generated as

$$u_{it} \overset{iid}{\sim} \mathcal{N}(0, 4)$$

and the idiosyncratic components of the regressors are generated as

$$\eta_{l,it} \overset{iid}{\sim} \mathcal{N}(0,1) \text{ for } l = 1, 2,$$

where $l$ indicates the regressor and $\mu_l = 1$ for $l = 1, 2$.

**DGP7, (Chudik et al., 2011)**

This simulation setup is based on a model with two regressors and three unobserved factors,

$$y_{it} = \beta_1 x_{1,it} + \beta_2 x_{2,it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{it} \tag{A.4}$$

where $\beta_1 = \beta_2 = 1$, $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \lambda_{2,i}, \lambda_{3,i})'$ and $\boldsymbol{f}_t = (f_{1,t}, f_{2,t}, f_{3,t})'$. The regressors are generated according to

$$x_{1,it} = \boldsymbol{\gamma}_{1,i}' \boldsymbol{f}_t + \eta_{1,it}, \tag{A.5}$$

$$x_{2,it} = \boldsymbol{\gamma}_{2,i}' \boldsymbol{f}_t + \eta_{2,it}, \tag{A.6}$$

where $\boldsymbol{\gamma}_{1,i}$ and $\boldsymbol{\gamma}_{2,i}$ denote $r$-dimensional vectors of loadings for the regressors that are independent of the loadings in the DGP of the dependent variable, $\boldsymbol{\lambda}_i$. The unobserved factors are generated as independent AR(1) processes,

$$f_{j,t} = 0.5 f_{j,t-1} + \upsilon_{f_{j,t}},\ j = 1, 2, 3;\ t = -49, ..., 0, 1, ..., T$$
$$\upsilon_{f_{j,t}} \overset{iid}{\sim} \mathcal{N}(0, 1 - 0.5^2),\ f_{j,-50} = 0.$$

In order to reduce the effect of the initial value, the first 50 observations of $f_{j,t}$ are discarded. The factor loadings in the DGP of $y_{it}$ are generated as

$$\lambda_{j,i} \overset{iid}{\sim} \mathcal{N}(0,1) \text{ for } j = 1, 2, 3$$

and are independently distributed from the factor loadings in the DGPs of the regressors,

$$\gamma_{l,j,i} \overset{iid}{\sim} \mathcal{N}(0,1) \text{ for } l = 1, 2;\ j = 1, 2, 3$$

where $l$ denotes the index for the regressor $x_{l,it}$. The regression errors exhibit mild heteroskedasticity and are generated as

$$u_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_i^2), \text{ where } \sigma_i^2 \overset{iid}{\sim} \mathcal{U}(0.5, 1.5).$$

The idiosyncratic components of the regressors are generated according to

$$\eta_{l,it} = \rho_{\nu_{l,i}} \eta_{l,it-1} + \nu_{j,it} \text{ for } l = 1, 2; \ t = -49, ..., 0, 1, ...T$$
$$\nu_{l,it} \overset{iid}{\sim} \mathcal{N}(0, 1 - \rho^2_{\nu_{j,i}}), \ \eta_{l,i,-50} = 0, \ \rho_{\nu_{l,i}} \overset{iid}{\sim} \mathcal{U}(0.05, 0.95) \text{ for } l = 1, 2.$$

The first 50 observations of $\eta_{l,t}$ are discarded as "burn-in" period.

**DGP8, (Ahn et al, 2013):**

For this DGP, we consider a model with $k = 2$ and $r = 2$,

$$y_{it} = \beta_1 x_{1,it} + \beta_2 x_{2,it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{it} \tag{A.7}$$

where $\beta_1 = \beta_2 = 1$, $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \lambda_{2,i})'$ and $\boldsymbol{f}_t = (f_{1,t}, f_{2,t})$. The regressors are generated by

$$x_{1,it} = \lambda_{1,i} f_{1,t} + \lambda_{1,i} + f_{1,t} + \eta_{1,it} + \mu_{1,i} \tag{A.8}$$
$$x_{2,it} = \lambda_{2,i} f_{2,t} + \lambda_{2,i} + f_{2,t} + \eta_{2,it} + \mu_{2,i} \tag{A.9}$$

DGP8 differs from DGP6 in that the regressor $x_{l,it}$ for $l = 1, 2$ is only correlated with one factor $f_{j,t}$, the loadings $\lambda_{j,i}$ and the product $\lambda_{j,i} f_{j,t}$ for $j = 1, 2$, but is independent of the other factor and loadings. The unobserved factors follow a uniform distribution,

$$f_{j,t} \overset{iid}{\sim} \mathcal{U}(0, 2) \text{ for } j = 1, 2,$$

and the loadings follow a normal distribution,

$$\lambda_{j,i} \overset{iid}{\sim} \mathcal{N}(0, 4) \text{ for } j = 1, 2.$$

The regression errors are generated by an AR(1) process,

$$u_{it} = \rho u_{i,t-1} + \nu_{it} \text{ for } t = -49, ..., 0, 1, ...T,$$
$$\text{where } \rho = 0.5, \ \nu_{it} \sim \mathcal{N}(0, 1) \text{ and } u_{i,-50} = 0.$$

The first 50 time observations of $u_{it}$ are discarded. The idiosyncractic components of the regressors are

$$\eta_{l,it} \overset{iid}{\sim} \mathcal{N}(0,1) \text{ and } \mu_{l,i} \overset{iid}{\sim} \mathcal{N}(0,1) \text{ for } l = 1, 2.$$

# Appendix B.

# Appendix of Chapter 3

## B.1. 49 industry portfolios with adjusted time period

Table B.1 reports performance results for the 49 industry portfolios with an out-of-sample period ranging from December 26th. 2003 to November 14th. 2019 matching the investment period for the CRSP dataset. The first thing we notice is that the absolute performance deteriorates in comparison to the results that comprise the full out-of-sample period (Table 3.6) and are much more similar to the performance results for the CRSP dataset (Table 3.5). For the short-selling strategy, the relative ranking of the portfolios is largely consistent with the results reported in Table 3.6. The main difference is that the performance of the DCCNL falls behind the static LW approaches, which is accordance with the results for the CRSP dataset. In addition, the simple 1F portfolio attains the second highest Sharpe ratio after the BJL$^*$ portfolio, which stems, however, from a higher portfolio return. For the put option strategy, the blocking strategies rank behind the static-LW and LASSO portfolios in terms of Sharpe ratio.

## B.2. L1-regularization: choice of $\lambda$

The estimation of LASSO regressions requires to choose the tuning parameter $\lambda$. There are various possibilities to choose $\lambda$ but the most common approach is k-fold cross-validation: The available data for estimation is split into $k$ folds and the model is fitted for each value of $\lambda$ to $k-1$ of these folds (the training set) and evaluated on the fold that was excluded in the estimation (test set). This process is repeated for each of the $k$ folds and for the classical cross-validation, one selects the $\lambda$ that minimizes the overall mean squared error (MSE). In Table B.2 we show results for the cross-validation procedure with the following alternative error functions.

Table B.1.: Performance measures for the 49 industry portfolios with adapted out-of-sample period

| | short-selling | | | put option | | |
|---|---|---|---|---|---|---|
| Strategy | SR | $\mu$ | $\sigma$ | SR | $\mu$ | $\sigma$ |
| MSR($\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$) | 0.22 | 245.89 | 1136.10 | 1.00 | 2.24 | 2.25 |
| GMV($\widehat{\boldsymbol{\Sigma}}$) | 1.32 | 13.07 | 9.89 | 1.38 | 3.58 | 2.60 |
| 1/N | 0.69 | 13.27 | 19.16 | 0.69 | 13.27 | 19.16 |
| | | | covariance regularization | | | |
| CN | 0.78 | 13.78 | 17.73 | 0.78 | 13.78 | 17.73 |
| 1F | 1.41 | 16.57 | 11.72 | 1.45 | 5.52 | 3.81 |
| LW1F | 1.36 | 13.39 | 9.86 | 1.43 | 3.82 | 2.67 |
| LWNL | 1.35 | 13.32 | 9.87 | 1.41 | 3.81 | 2.71 |
| DCCNL | 1.17 | 11.71 | 9.98 | 1.10 | 3.14 | 2.85 |
| | | | weight constraints | | | |
| $GMV_+$ | 1.01 | 14.02 | 13.83 | 1.01 | 14.02 | 13.83 |
| $MSR_+$ | 0.94 | 14.90 | 15.84 | 0.94 | 14.90 | 15.84 |
| KML | 1.33 | 13.20 | 9.91 | 1.41 | 4.24 | 3.02 |
| BJL* | 1.53 | 17.05 | 11.17 | 1.52 | 6.55 | 4.30 |
| | | | blocking strategies | | | |
| block(eq) | 1.29 | 14.44 | 11.23 | 1.26 | 5.35 | 4.25 |
| block(ord) | 1.40 | 15.50 | 11.10 | 1.35 | 5.64 | 4.18 |
| block(SR) | 1.37 | 15.10 | 11.00 | 1.32 | 5.51 | 4.17 |

The table reports the annualized out-of-sample Sharpe ratios, returns and standard deviations of 15 portfolio strategies applied to the 49 industry portfolios. The estimation period spans 1000 trading days with rebalancing every 50 days. The out-of-sample period ranges from December 26th. 2003 - November 14th. 2019.

- CV-MSE: We conduct 5-fold cross-validation and choose the $\lambda$ that minimizes the MSE.
- CV-STD: We conduct 5-fold cross-validation and choose the $\lambda$ that minimizes the portfolio standard deviation.
- CV-SR: We conduct 5-fold cross-validation and choose the $\lambda$ that maximizes the portfolio Sharpe Ratio.
- CV-1SE: We conduct 5-fold cross-validation and choose the largest value of $\lambda$ such that the MSE is within one standard error of the minimum. This promotes a more regularized model than CV-MSE.

In addition, we consider the following information criteria to guide the choice of $\lambda$.

- AIC: We choose the $\lambda$ that minimizes $2k - 2\ln(\hat{L})$, where $\hat{L}$ is the maximized value of the likelihood function, $k$ denotes the degrees of freedom.
- BIC: We choose the $\lambda$ that minimizes the $k\ln(N) - 2\ln(\hat{L})$.

Table B.2.: Alternative methods for choosing $\lambda$

| | short-selling | | | put option | | |
|---|---|---|---|---|---|---|
| Strategy | SR | $\mu$ | $\sigma$ | SR | $\mu$ | $\sigma$ |
| | | | KML | | | |
| MSE | 0.80 | 7.65 | 9.52 | 0.86 | 3.51 | 4.09 |
| STD | 0.76 | 7.25 | 9.49 | 0.86 | 3.47 | 4.05 |
| SR | 0.60 | 8.31 | 13.96 | 0.57 | 7.34 | 12.77 |
| 1SE | 0.81 | 7.88 | 9.73 | 0.87 | 4.92 | 5.66 |
| AIC | 0.78 | 8.44 | 10.76 | 0.74 | 6.46 | 8.68 |
| BIC | 0.80 | 9.62 | 12.04 | 0.76 | 8.51 | 11.19 |
| | | | BJL* | | | |
| MSE | 0.79 | 7.51 | 9.54 | 0.85 | 3.44 | 4.04 |
| STD | 0.76 | 7.24 | 9.51 | 0.86 | 3.48 | 4.06 |
| SR | 0.64 | 8.93 | 14.03 | 0.61 | 7.86 | 12.92 |
| 1SE | 0.80 | 7.76 | 9.72 | 0.86 | 4.84 | 5.60 |
| AIC | 0.79 | 8.45 | 10.76 | 0.75 | 6.51 | 8.67 |
| BIC | 0.80 | 9.75 | 12.11 | 0.76 | 8.60 | 11.26 |

The table reports the performance of alternative methods to choose $\lambda$ for the LASSO approaches applied to the CRSP dataset.

The results reported in Table B.2 show that the classical cross-validation minimizing the MSE belongs to the best-performing approaches to select $\lambda$. The alternative approaches CV-STD and CV-1SE yield similar performances. Guiding the choice of $\lambda$ by

minimizing AIC or BIC yields Sharpe ratios that are in a similar range as CV-MSE, which are, however, accompanied by larger portfolio standard deviations. The CV-SR approach performs worst in terms of Sharpe ratio and exhibits the largest standard deviations.

# Appendix C.

# Appendix of Chapter 4

Table C.1.: Performance results for VaR 97.5% (est. window 500)

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | 0.0008 | 1.37 | 0.0309 | 2.78 | 0.0312 | 2.58 |
| GPD | 0.0303 | 2.24 | 0.0251 | 2.09 | 0.0272 | 1.88 |
| Hill | −0.0359 | 2.39 | 0.0230 | 1.13 | 0.0257 | 1.54 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 97.5% VaR with 500 periods and 2000 replications.

Table C.2.: Performance results for ES 97.5% (est. window 500)

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | −0.0590 | 4.70 | 0.1745 | 12.62 | 0.1782 | 9.04 |
| GPD | −0.0065 | 1.89 | 0.7034 | 13.85 | 0.7044 | 10.44 |
| Hill | 0.0955 | 7.01 | 0.2280 | 14.57 | 0.2408 | 12.07 |
| n-FHS | −0.4782 | 13.70 | 0.0440 | 3.82 | 0.3153 | 13.30 |
| n-GPD | −0.4431 | 12.73 | 0.0357 | 2.09 | 0.2681 | 11.57 |
| n-Hill | −0.5220 | 14.75 | 0.0328 | 1.13 | 0.3383 | 14.51 |
| t-FHS | −0.0466 | 5.86 | 0.1070 | 11.98 | 0.1280 | 8.15 |
| t-GPD | −0.0034 | 5.16 | 0.0995 | 11.31 | 0.1137 | 6.69 |
| t-Hill | −0.0968 | 6.94 | 0.0933 | 10.00 | 0.1157 | 6.58 |
| p-FHS | −0.0492 | 8.75 | 0.0620 | 7.96 | 0.1137 | 6.61 |
| p-GPD | −0.0074 | 8.11 | 0.0506 | 5.83 | 0.0915 | 4.49 |
| p-Hill | −0.1012 | 8.33 | 0.0465 | 4.01 | 0.0935 | 3.89 |
| i-FHS | −0.0655 | 7.46 | 0.0637 | 8.68 | 0.1051 | 5.63 |
| i-GPD | −0.0235 | 6.89 | 0.0533 | 6.99 | 0.0843 | 3.56 |
| i-Hill | −0.1165 | 7.72 | 0.0492 | 5.17 | 0.0903 | 3.45 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 97.5% ES with 500 periods and 2000 replications.

Table C.3.: Performance results for VaR 99% (est. window 500)

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | 0.0256 | 1.56 | 0.1477 | 2.88 | 0.1510 | 2.81 |
| GPD | 0.0506 | 2.48 | 0.0924 | 1.43 | 0.1026 | 1.58 |
| Hill | 0.0146 | 1.97 | 0.0950 | 1.69 | 0.1004 | 1.61 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 99% VaR with 500 periods and 2000 replications.

Table C.4.: Performance results for ES 99% (est. window 500)

| Method | Bias | $\overline{\text{Rank}}$ | Variance | $\overline{\text{Rank}}$ | MSE | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|
| FHS | −0.0836 | 3.74 | 0.7418 | 12.97 | 0.7496 | 10.57 |
| GPD | −0.0896 | 3.96 | 3.8871 | 13.79 | 3.9027 | 12.23 |
| Hill | 0.2612 | 9.14 | 0.8552 | 14.41 | 0.9621 | 13.77 |
| n-FHS | −0.7221 | 13.21 | 0.1939 | 4.93 | 0.8619 | 12.13 |
| n-GPD | −0.6935 | 12.59 | 0.1213 | 1.44 | 0.7290 | 9.34 |
| n-Hill | −0.7347 | 13.77 | 0.1247 | 1.70 | 0.8028 | 11.16 |
| t-FHS | −0.1240 | 6.34 | 0.4513 | 12.23 | 0.5351 | 9.27 |
| t-GPD | −0.0893 | 5.04 | 0.3244 | 9.85 | 0.3823 | 5.77 |
| t-Hill | −0.1397 | 6.16 | 0.3258 | 10.13 | 0.4047 | 6.53 |
| p-FHS | −0.0839 | 8.80 | 0.2831 | 8.74 | 0.4373 | 7.49 |
| p-GPD | −0.0489 | 7.89 | 0.1765 | 4.30 | 0.3009 | 3.84 |
| p-Hill | −0.0991 | 8.33 | 0.1812 | 4.83 | 0.3270 | 4.38 |
| i-FHS | −0.0835 | 7.32 | 0.2972 | 9.23 | 0.4155 | 6.73 |
| i-GPD | −0.0470 | 6.68 | 0.1924 | 5.55 | 0.2874 | 3.19 |
| i-Hill | −0.0979 | 7.03 | 0.1954 | 5.91 | 0.3084 | 3.61 |

This table reports the performance results of the bootstrap resampling experiment for the estimation of 99% ES with 500 periods and 2000 replications.

Table C.5.: Average scores and ranks for $(\mathrm{VaR}_{97.5\%}, \mathrm{ES}_{97.5\%})$, (est. window 500)

| Method | naive | | MV+ | | MCD | | NVDA | | BAC | | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | 1.2610 | (11) | 1.2435 | (11) | 1.1937 | (11) | 2.0818 | (10) | 1.9347 | (11) | 10.8 |
| t-FP | 1.0300 | (3) | 1.1082 | (7) | 1.1725 | (10) | 2.1481 | (11) | 1.6175 | (10) | 8.2 |
| FHS | 1.0337 | (6) | 1.1060 | (3) | 1.1095 | (9) | 2.0068 | (9) | 1.5718 | (6) | 6.6 |
| GPD | 1.0285 | (2) | 1.1081 | (6) | 1.1042 | (7) | 2.0023 | (8) | 1.5740 | (8) | 6.2 |
| Hill | 1.0334 | (5) | 1.1083 | (8) | 1.1068 | (8) | 1.9997 | (7) | 1.5757 | (9) | 7.4 |
| s-FHS | 1.0392 | (10) | 1.1047 | (2) | 1.0988 | (4) | 1.9923 | (3) | 1.5633 | (1) | 4.0 |
| s-GPD | 1.0321 | (4) | 1.1092 | (9) | 1.0943 | (1) | 1.9884 | (1) | 1.5691 | (3) | 3.6 |
| s-Hill | 1.0338 | (7) | 1.1067 | (4) | 1.0976 | (2) | 1.9957 | (5) | 1.5717 | (5) | 4.6 |
| t-FHS | 1.0351 | (9) | 1.1040 | (1) | 1.1042 | (6) | 1.9955 | (4) | 1.5645 | (2) | 4.4 |
| t-GPD | 1.0277 | (1) | 1.1078 | (5) | 1.0983 | (3) | 1.9910 | (2) | 1.5695 | (4) | 3.0 |
| t-Hill | 1.0341 | (8) | 1.1096 | (10) | 1.1026 | (5) | 1.9973 | (6) | 1.5739 | (7) | 7.2 |

This table shows the scaled average scores for the estimation of $(\mathrm{VaR}_{97.5\%}, \mathrm{ES}_{97.5\%})$ with a rolling window of 500 time periods. The corresponding ranks are presented in brackets. The last column contains the average rank for each estimation method.

Table C.6.: Average scores and ranks for $(\mathrm{VaR}_{99\%}, \mathrm{ES}_{99\%})$, (est. window 500)

| Method | naive | | MV+ | | MCD | | NVDA | | BAC | | $\overline{\text{Rank}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | 1.4994 | (11) | 1.5110 | (11) | 1.4715 | (11) | 2.3990 | (10) | 2.2204 | (11) | 10.8 |
| t-FP | 1.2150 | (2) | 1.2978 | (3) | 1.4259 | (10) | 2.4338 | (11) | 1.8570 | (10) | 7.2 |
| FHS | 1.2174 | (6) | 1.3106 | (8) | 1.3973 | (9) | 2.3183 | (8) | 1.8165 | (9) | 8.0 |
| GPD | 1.2208 | (8) | 1.2990 | (4) | 1.3725 | (6) | 2.3310 | (9) | 1.8067 | (7) | 6.8 |
| Hill | 1.2226 | (10) | 1.3061 | (7) | 1.3715 | (5) | 2.3100 | (5) | 1.8088 | (8) | 7.0 |
| s-FHS | 1.2145 | (1) | 1.3130 | (10) | 1.3868 | (7) | 2.2833 | (1) | 1.8006 | (5) | 4.8 |
| s-GPD | 1.2165 | (5) | 1.2968 | (1) | 1.3548 | (1) | 2.3063 | (4) | 1.7968 | (2) | 2.6 |
| s-Hill | 1.2198 | (7) | 1.3013 | (5) | 1.3576 | (2) | 2.3061 | (3) | 1.7938 | (1) | 3.6 |
| t-FHS | 1.2154 | (3) | 1.3125 | (9) | 1.3959 | (8) | 2.2884 | (2) | 1.8047 | (6) | 5.6 |
| t-GPD | 1.2162 | (4) | 1.2972 | (2) | 1.3627 | (3) | 2.3101 | (6) | 1.7987 | (4) | 3.8 |
| t-Hill | 1.2211 | (9) | 1.3031 | (6) | 1.3666 | (4) | 2.3103 | (7) | 1.7969 | (3) | 5.8 |

This table shows the scaled average scores for the estimation of $(\mathrm{VaR}_{99\%}, \mathrm{ES}_{99\%})$ with a rolling window of 500 time periods. The corresponding ranks are presented in brackets. The last column contains the average rank for each estimation method.

Table C.7.: Diebold-Mariano t-statistics for the $MV_+$ portfolio, ES 97.5%

| Method | HS | t-FP | FHS | GPD | Hill | s-FHS | s-GPD | s-Hill | t-FHS | t-GPD | t-Hill |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | · | 3.67 | 3.67 | 3.67 | 3.65 | 3.60 | 3.59 | 3.63 | 3.65 | 3.67 | 3.62 |
| t-FP | −3.67 | · | 0.22 | −0.07 | −0.25 | −0.45 | −1.42 | −0.55 | 0.21 | −0.04 | −0.30 |
| FHS | −3.67 | −0.22 | · | −0.55 | −1.08 | −0.75 | −1.05 | −1.45 | 0.11 | −0.48 | −0.82 |
| GPD | −3.67 | 0.07 | 0.55 | · | −0.41 | −0.40 | −1.03 | −1.02 | 0.46 | 0.15 | −0.44 |
| Hill | −3.65 | 0.25 | 1.08 | 0.41 | · | −0.12 | −0.60 | −0.90 | 0.87 | 0.43 | −0.32 |
| s-FHS | −3.60 | 0.45 | 0.75 | 0.40 | 0.12 | · | −0.99 | −0.24 | 0.76 | 0.47 | −0.07 |
| s-GPD | −3.59 | 1.42 | 1.05 | 1.03 | 0.60 | 0.99 | · | 0.39 | 0.99 | 1.18 | 0.25 |
| s-Hill | −3.63 | 0.55 | 1.45 | 1.02 | 0.90 | 0.24 | −0.39 | · | 1.23 | 0.98 | 0.09 |
| t-FHS | −3.65 | −0.21 | −0.11 | −0.46 | −0.87 | −0.76 | −0.99 | −1.23 | · | −0.41 | −0.81 |
| t-GPD | −3.67 | 0.04 | 0.48 | −0.15 | −0.43 | −0.47 | −1.18 | −0.98 | 0.41 | · | −0.43 |
| t-Hill | −3.62 | 0.30 | 0.82 | 0.44 | 0.32 | 0.07 | −0.25 | −0.09 | 0.81 | 0.43 | · |

This table reports the Diebold-Mariano t-statistics comparing the average scores for the $MV_+$ portfolio for ES 97.5% with a rolling window of 1000 periods. Positive t-statistics show that the column model exhibits a lower average score than the row model.

Table C.8.: Diebold-Mariano t-statistics for the MCD portfolio, ES 97.5%

| Method | HS | t-FP | FHS | GPD | Hill | s-FHS | s-GPD | s-Hill | t-FHS | t-GPD | t-Hill |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | · | 1.17 | 3.38 | 3.48 | 3.46 | 3.60 | 3.68 | 3.59 | 3.43 | 3.55 | 3.42 |
| t-FP | −1.17 | · | 3.14 | 3.40 | 3.38 | 3.62 | 3.90 | 3.41 | 3.26 | 3.60 | 3.11 |
| FHS | −3.38 | −3.14 | · | 0.97 | 1.39 | 2.54 | 1.58 | 2.08 | 1.36 | 1.56 | 1.29 |
| GPD | −3.48 | −3.40 | −0.97 | · | 0.21 | 1.13 | 1.65 | 1.74 | 0.12 | 1.36 | 0.57 |
| Hill | −3.46 | −3.38 | −1.39 | −0.21 | · | 1.19 | 1.15 | 1.52 | 0.00 | 0.84 | 0.47 |
| s-FHS | −3.60 | −3.62 | −2.54 | −1.13 | −1.19 | · | 0.37 | 0.40 | −1.08 | −0.14 | −0.30 |
| s-GPD | −3.68 | −3.90 | −1.58 | −1.65 | −1.15 | −0.37 | · | 0.04 | −0.80 | −0.66 | −0.50 |
| s-Hill | −3.59 | −3.41 | −2.08 | −1.74 | −1.52 | −0.40 | −0.04 | · | −0.98 | −0.60 | −0.99 |
| t-FHS | −3.43 | −3.26 | −1.36 | −0.12 | 0.00 | 1.08 | 0.80 | 0.98 | · | 0.69 | 0.45 |
| t-GPD | −3.55 | −3.60 | −1.56 | −1.36 | −0.84 | 0.14 | 0.66 | 0.60 | −0.69 | · | −0.26 |
| t-Hill | −3.42 | −3.11 | −1.29 | −0.57 | −0.47 | 0.30 | 0.50 | 0.99 | −0.45 | 0.26 | · |

Diebold-Mariano t-statistics for the MCD portfolio. The table notes for Table C.7 apply.

Table C.9.: Diebold-Mariano t-statistics for the NVDA portfolio, ES 97.5%

| Method | HS | t-FP | FHS | GPD | Hill | s-FHS | s-GPD | s-Hill | t-FHS | t-GPD | t-Hill |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HS | · | 1.48 | 5.33 | 5.41 | 5.43 | 5.40 | 5.63 | 5.34 | 5.29 | 5.51 | 5.27 |
| t-FP | −1.48 | · | 2.99 | 3.11 | 3.06 | 3.06 | 3.34 | 2.83 | 3.30 | 3.65 | 3.12 |
| FHS | −5.33 | −2.99 | · | −0.23 | 0.19 | 1.42 | 1.31 | −0.46 | 1.81 | 1.92 | 1.10 |
| GPD | −5.41 | −3.11 | 0.23 | · | 0.33 | 1.12 | 2.60 | −0.18 | 1.51 | 2.34 | 0.96 |
| Hill | −5.43 | −3.06 | −0.19 | −0.33 | · | 1.50 | 1.27 | −0.67 | 2.09 | 2.26 | 1.35 |
| s-FHS | −5.40 | −3.06 | −1.42 | −1.12 | −1.50 | · | −0.20 | −2.20 | 1.30 | 0.65 | −0.46 |
| s-GPD | −5.63 | −3.34 | −1.31 | −2.60 | −1.27 | 0.20 | · | −1.48 | 0.85 | 1.37 | −0.12 |
| s-Hill | −5.34 | −2.83 | 0.46 | 0.18 | 0.67 | 2.20 | 1.48 | · | 2.01 | 1.81 | 1.55 |
| t-FHS | −5.29 | −3.30 | −1.81 | −1.51 | −2.09 | −1.30 | −0.85 | −2.01 | · | −0.17 | −1.93 |
| t-GPD | −5.51 | −3.65 | −1.92 | −2.34 | −2.26 | −0.65 | −1.37 | −1.81 | 0.17 | · | −1.27 |
| t-Hill | −5.27 | −3.12 | −1.10 | −0.96 | −1.35 | 0.46 | 0.12 | −1.55 | 1.93 | 1.27 | · |

Diebold-Mariano t-statistics for the NVDA portfolio. The table notes for Table C.7 apply.

Table C.10.: Simulation results for t-distributed data (est. window 500)

| | VaR 99% | | VaR 97.5% | | ES 99% | | | ES 97.5% | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dof | MLE | HS | MLE | HS | MLE | HS | ratio | MLE | HS | ratio |
| | | | | | Bias | | | | | |
| 3 | 0.61 | 6.34 | 0.00 | 1.46 | 5.06 | −16.91 | 14.02 | 2.14 | 2.11 | 4.49 |
| 4 | 0.05 | 3.19 | −0.18 | 0.80 | 1.77 | −10.69 | 6.20 | 0.64 | 1.31 | 2.06 |
| 5 | −0.22 | 1.77 | −0.28 | 0.45 | 0.63 | −8.83 | 3.29 | 0.09 | 0.60 | 1.10 |
| 6 | −0.41 | 1.13 | −0.36 | 0.19 | 0.04 | −7.58 | 2.03 | −0.22 | 0.35 | 0.52 |
| Avg. | 0.01 | 3.11 | −0.21 | 0.73 | 1.88 | −11.00 | 6.39 | 0.66 | 1.09 | 2.04 |
| | | | | | Variance | | | | | |
| 3 | 17.69 | 59.73 | 4.32 | 13.74 | 105.67 | 441.18 | 212.86 | 34.13 | 107.62 | 58.93 |
| 4 | 8.37 | 26.65 | 2.33 | 7.58 | 37.17 | 119.56 | 74.24 | 13.48 | 34.36 | 24.66 |
| 5 | 5.29 | 16.55 | 1.59 | 5.37 | 20.03 | 56.76 | 40.41 | 7.79 | 18.26 | 15.03 |
| 6 | 3.92 | 12.09 | 1.24 | 4.27 | 13.45 | 35.57 | 27.27 | 5.48 | 12.38 | 10.94 |
| Avg. | 8.82 | 28.75 | 2.37 | 7.74 | 44.08 | 163.27 | 88.70 | 15.22 | 43.16 | 27.39 |
| | | | | | MSE | | | | | |
| 3 | 17.69 | 60.13 | 4.32 | 13.76 | 105.92 | 444.03 | 214.83 | 34.18 | 107.67 | 59.13 |
| 4 | 8.37 | 26.75 | 2.33 | 7.58 | 37.20 | 120.70 | 74.62 | 13.48 | 34.38 | 24.70 |
| 5 | 5.29 | 16.58 | 1.59 | 5.37 | 20.04 | 57.54 | 40.52 | 7.79 | 18.26 | 15.04 |
| 6 | 3.92 | 12.11 | 1.24 | 4.27 | 13.45 | 36.14 | 27.32 | 5.48 | 12.38 | 10.94 |
| Avg. | 8.82 | 28.89 | 2.37 | 7.75 | 44.15 | 164.61 | 89.32 | 15.23 | 43.17 | 27.45 |

This table reports simulation results for estimating VaR and ES by MLE and HS at the 99%, 97.5% confidence levels. The data is generated by a standard t-distribution with 3-6 degrees of freedom. The results are based on 400 assets (100 per d.o.f.), 500 time periods and 2000 replications and all performance numbers are multiplied by 100.

# Bibliography


Acerbi, C. and Szekely, B. (2014). Back-testing expected shortfall. *Risk*, 27(11):76–81.

Acerbi, C. and Tasche, D. (2002). Expected shortfall: a natural coherent alternative to value at risk. *Economic notes*, 31(2):379–388.

Ahn, S. and Horenstein, A. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Ahn, S., Lee, Y., and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics*, 101:219–255.

Ahn, S., Lee, Y., and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics*, 174:1–14.

Albuquerque, R. (2012). Skewness in stock returns: Reconciling the evidence on firm versus aggregate returns. *The Review of Financial Studies*, 25(5):1630–1673.

Ao, M., Yingying, L., and Zheng, X. (2019). Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3):203–228.

Bai, J. (2009). Panel data model with interative fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Baltagi, B. (2005). *Econometric Analysis of Panel Data*. John Wiley & Sons Inc., New York, 3 edition.

Bayer, S. and Dimitriadis, T. (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, 20(3):437–471.

BCBS (2016). Minimum capital requirements for market risk. BIS online publication no. d352, Basel Committee on Banking Supervision: Bank for International Settlements.

BCBS (2019). Minimum capital requirements for market risk. BIS online publication no. d457, Basel Committee on Banking Supervision: Bank for International Settlements.

Bekker, P. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, 62(3):657–681.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance*, 15(5):725–733.

Benartzi, S. and Thaler, R. H. (2001). Naive diversification strategies in defined contribution saving plans. *American economic review*, 91(1):79–98.

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. *The review of financial studies*, 4(2):315–342.

Breitung, J. (2015). The analysis of macroeconomic panel data. In Baltagi, B., editor, *The Oxford Handbook of Panel Data*, chapter 15, pages 453–492. Oxford University Press.

Breitung, J. and Hansen, P. (2021). Alternative estimation approaches for the factor augmented panel data model with small T. *Empirical Economics*, 60:327–351.

Breitung, J. and Mann, K. (2017). Assessing the forward premium puzzle: A factor-augmented panel data approach. *International Currency Exposure*, page 265.

Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance*, 54(2):655–671.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Chopra, V. K. and Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, 19:6–11.

Chudik, A., Pesaran, M., and Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal*, 14:C45–C90.

Cont, R., Deguest, R., and Scandolo, G. (2010). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative finance*, 10(6):593–606.

Daníelsson, J. (2011). *Financial risk forecasting*. John Wiley & Sons.

Daníelsson, J., Jorgensen, B. N., Samorodnitsky, G., Sarma, M., and de Vries, C. G. (2013). Fat tails, VaR and subadditivity. *Journal of econometrics*, 172(2):283–291.

De Nard, G., Ledoit, O., and Wolf, M. (2019). Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics*.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009a). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management science*, 55(5):798–812.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009b). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The review of Financial studies*, 22(5):1915–1953.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Dimitriadis, T. and Schnaitmann, J. (2021). Forecast encompassing tests for the expected shortfall. *International Journal of Forecasting*, 37(2):604–621.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *arXiv preprint arXiv:1312.1645*.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*, 37(2):363–375.

European Banking Authority (2021). EBA report results from the 2020 market risk benchmarking exercise. EBA/REP/2021/05, European Banking Authority.

Everaert, G. and De Groote, T. (2016). Common correlated effects estimation of dynamic panels with cross-sectional dependence. *Econometric Reviews*, 35(3):428–463.

Eyster, E. and Weizsäcker, G. (2016). Correlation neglect in portfolio choice: Lab evidence. *SSRN working paper 2914526*.

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, 38(1):34–105.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Fissler, T. and Hoga, Y. (2023). Backtesting systemic risk forecasts using multi-objective elicitability. *Journal of Business & Economic Statistics*, (just-accepted):1–30.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and osband's principle. *The Annals of Statistics*, 44(4):1680–1707.

Fissler, T., Ziegel, J. F., and Gneiting, T. (2016). Expected shortfall is jointly elicitable with value at risk - implications for backtesting. *Risk*, pages 58–61.

French, K. R. (2020). Data library. mba.tuck.dartmouth.edu/pages/faculty/ken.french. accessed on 07.05.2020.

Frey, C. and Pohlmeier, W. (2016). Bayesian shrinkage of portfolio weights. *Available at SSRN 2730475*.

Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401.

Garcia, R., Renault, É., and Tsafack, G. (2007). Proper conditioning for coherent VaR in portfolio management. *Management Science*, 53(3):483–494.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Greenaway-McGrevy, R., Han, C., and Sul, D. (2012). Asymptotic distribution of factor augmented estimators for panel regression. *Journal of Econometrics*, 169(1):48–53.

Hayakawa, K. (2012). GMM estimation of short dynamic panel data models with interactive fixed effects. *Journal of the Japan Statistical Society*, 42(2):109–123.

He, X. D., Kou, S., and Peng, X. (2022). Risk measures: robustness, elicitability, and backtesting. *Annual Review of Statistics and Its Application*, 9:141–166.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174.

Holtz-Eakin, D., Newey, W., and Rosen, H. (1988). Estimating vector autoregressions with panel data. *Econometrica*, 56(6):1371–1395.

Ibragimov, R. (2009). Portfolio diversification and value at risk under thick-tailedness. *Quantitative Finance*, 9(5):565–580.

Ibragimov, R. and Walden, J. (2007). The limits of diversification when losses may be large. *Journal of banking & finance*, 31(8):2551–2569.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1683.

Jansen, D. W. and De Vries, C. G. (1991). On the frequency of large stock returns: Putting booms and busts into perspective. *The review of economics and statistics*, pages 18–24.

Jondeau, E., Zhang, Q., and Zhu, X. (2019). Average skewness matters. *Journal of Financial Economics*, 134(1):29–47.

Juodis, A. and Sarafidis, V. (2018). Fixed T dynamic panel data estimators with multifactor errors. *Econometric Reviews*, 37(8):893–929.

Juodis, A. and Sarafidis, V. (2020). A linear estimator for factor-augmented fixed-T panels with endogenous regressors. *Journal of Business & Economic Statistics*.

Karabiyik, H., Palm, F. C., and Urbain, J.-P. (2019a). Econometric analysis of panel data models with multifactor error structures. *Annual Review of Economics*, 11:495–522.

Karabiyik, H., Urbain, J.-P., and Westerlund, J. (2019b). CCE estimation of factor-augmented regression models with more factors than observables. *Journal of Applied Econometrics*, 34(2):268–284.

Kellner, R. and Rösch, D. (2016). Quantifying market risk with value-at-risk or expected shortfall? – consequences for capital requirements and model risk. *Journal of Economic Dynamics and Control*, 68:45–63.

Kempf, A. and Memmel, C. (2006). Estimating the global minimum variance portfolio. *Schmalenbach Business Review*, 58(4):332–248.

Kou, S. and Peng, X. (2016). On the measurement of economic tail risk. *Operations Research*, 64(5):1056–1072.

Kou, S., Peng, X., and Heyde, C. C. (2013). External risk measures and basel accords. *Mathematics of Operations Research*, 38(3):393–417.

Kuester, K., Mittnik, S., and Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1):53–89.

Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138.

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(5):603–621.

Ledoit, O. and Wolf, M. (2004a). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Ledoit, O. and Wolf, M. (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360–384.

Ledoit, O. and Wolf, M. (2017a). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388.

Ledoit, O. and Wolf, M. (2017b). Numerical implementation of the quest function. *Computational Statistics & Data Analysis*, 115:199–223.

Ledoit, O. and Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5):3043–3065.

Lee, N., Moon, H., and Zhou, Q. (2017). Many IVs estimation of dynamic panel regression models with measurement error. *Journal of Econometrics*, 200(2):251–259.

Li, H. and Wang, R. (2023). PELVE: Probability equivalent level of VaR and ES. *Journal of Econometrics*, 234(1):353–370.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, 7(3-4):271–300.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management*. Princeton University Press.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized'optimal? *Financial analysts journal*, 45(1):31–42.

Moon, H. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.

Moon, H. and Weidner, W. (2019). Nuclear norm regularized estimation of panel regression models. *cemmap Working Paper*, CWP14/19.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11(4):1833–1874.

Okhrin, Y. and Schmid, W. (2006). Distributional properties of portfolio weights. *Journal of econometrics*, 134(1):235–256.

Osband, K. H. (1985). *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley.

Pakel, C., Shephard, N., Sheppard, K., and Engle, R. F. (2020). Fitting vast dimensional time-varying covariance models. *Journal of Business & Economic Statistics*, pages 1–17.

Patton, A. J. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In *Handbook of financial time series*, pages 801–838. Springer.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of econometrics*, 211(2):388–413.

Pérignon, C. and Smith, D. R. (2010). The level and quality of value-at-risk disclosure by commercial banks. *Journal of Banking & Finance*, 34(2):362–377.

Pesaran, M. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.

Righi, M. B. and Ceretta, P. S. (2015). A comparison of expected shortfall estimation models. *Journal of Economics and Business*, 78:14–47.

Robertson, D. and Sarafidis, V. (2015). IV estimation of panels with factor residuals. *Journal of Econometrics*, 185(2):526–541.

Sarafidis, V. and Wansbeek, T. (2012). Cross-sectional dependence in panel data analysis. *Econometric Reviews*, 31(5):483–531.

Stein, C. (1956). Inadmissability of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical and Statistical Probability. University of California Press*, 1:197–206.

Su, L. and Jin, S. (2012). Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics*, 169(1):34–47.

Tasche, D. (2002). Expected shortfall and beyond. *Journal of Banking & Finance*, 26(7):1519–1533.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133.

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2):428–441.

Westerlund, J., Karabiyik, H., Narayan, P. K., and Narayan, S. (2022). Estimating the speed of adjustment of leverage in the presence of interactive effects. *Journal of Financial Econometrics*, 20(5):942–960.

Westerlund, J., Petrova, Y., and Norkute, M. (2019). CCE in fixed-T panels. *Journal of Applied Econometrics*, 34(5):746–761.

Westerlund, J. and Urbain, J.-P. (2013). On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Economics Letters*, 119(3):247–250.

Westerlund, J. and Urbain, J.-P. (2015). Cross-sectional averages versus principal components. *Journal of Econometrics*, 185(2):372–377.

WRDS (2020). Wharton research data services. wrds.wharton.upenn.edu. accessed on 03.11.2020.