High dimensional covariance matrix estimation with l_1-regularized factor models[a]

Maurizio Daniele[b] — University of Konstanz — GSDS

Winfried Pohlmeier[c] — University of Konstanz — CoFE, RCEA

Aygul Zagidulina[d] — University of Konstanz — QEF

this version: February 16, 2018

Abstract

Estimates of high dimensional covariance matrices suffer from great instability. In this paper a novel approach of robust covariance matrix estimation based on sparse factor modeling is presented. Sparsity is obtained by both factor modeling and l_1-regularization of the factor loadings matrix that shrinks single factor loadings to zero. The positive aspect of this framework is the ability to consider also weak factors that affect only a subset of the available time series. Hence, our sparse factor model enhances the modeling flexibility in contrast to the standard approximate factor model that allows only for strong factors, affecting the entire set of time series. In the theoretical part of the paper, we derive the consistency of the factors and factor loadings estimators and establish various risk bounds for the covariance matrix estimator based on our sparse factor model.

The new approach applied to portfolio modeling shows superior properties compared to alternative shrinkage strategies that are commonly used in the literature. More specifically, our sparse factor model provides the lowest out-of-sample portfolio standard deviation across all considered portfolio sizes. Additionally, it offers the highest out-of-sample portfolio returns and hence is also superior in terms of certainty equivalent and sharpe ratio.

Keywords: Sparse Factor Model, l_1-Regularization, Covariance Matrix Estimation, Portfolio Allocation

JEL classification: G11, G17, C32, C38, C55

[a]Earlier versions of the paper were presented at the RCEA Macro-Money-Finance Workshop 2016 in Rimini and at the 3rd Konstanz-Lancaster Workshop on Finance and Econometrics 2017 in Lancaster. Financial support by the Graduate School of Decision Sciences (GSDS) and the German Science Foundation (DFG) is gratefully acknowledged. The usual disclaimer applies.

[b]Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-2657, fax: -4450, email: Maurizio.Daniele@uni-konstanz.de.

[c]Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-2660, fax: -4450, email: Winfried.Pohlmeier@uni-konstanz.de.

[d]Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-3753, fax: -4450, email: Aygul.Zagidullina@uni-konstanz.de.

1 Introduction

In the recent years, the importance of factor models as a dimension reduction technique has considerably increased in the macroeconomic and finance literature. This is mainly driven by the raising amount of available large datasets. The interesting aspect of factor models is their ability to concentrate the information contained in a large number of economic variables in a much smaller amount of latent factors. This dimension reduction has the great advantage that the number of parameters to estimate and therefore the estimation noise is highly reduced.

In the present paper, we propose a robust covariance matrix estimator based on a sparse factor model that allows for sparsity in the factor loadings matrix by shrinking single elements of the factor loadings matrix to zero. Essentially, the amount of parameters that have to be estimated in this setting decreases and the estimation noise is possibly reduced. On the other hand, this framework allows for weak factors that affect only a subset of the available time series. Hence, our sparse factor model enhances the modelling flexibility compared to the standard approximate factor model those assumptions generally only allow for strong or pervasive factors that affect the entire set of variables (see e.g. Bai and Ng (2002)). Furthermore, the strong factor assumption is unrealistic if we consider the characteristics of real datasets. In fact, strong factors imply a clear distinction of the eigenvalues of the common and idiosyncratic component, where the former diverge with a fast rate of N. However, the eigenvalues of the covariance matrix of real datasets do not show a clear separation, but decay in a smooth fashion. This properties can be better modelled by a weak factor framework and hence our sparse factor model allows for this flexibility.

In order to implement our sparse factor model, we refer to a penalized maximum likelihood estimation approach, where we incorporate a l_1 type of penalization on the factor loadings matrix to introduce sparsity in the loadings. The maximum likelihood approach has the advantage of estimating all parameters including the covariance matrix of the idiosyncratic error term Σ_u simultaneously, compared to a PCA-based method. In fact, the PCA setting needs a separate estimation of Σ_u, which on the other hand requires a consistent estimation of the factor loadings and factors. However, this is potentially not possible, if the cross-sectional dimension N is relatively small, as in this case a consistent estimation of the factors is no longer possible (see e.g. Bai and Liao (2016)).

In the theoretical part of the paper, we are able to show average consistency for the factor loadings and idiosyncratic error covariance matrix estimators. The factors estimated based on the GLS method are as well consistent. Furthermore, we derive several risk bounds for the covariance matrix estimated based on our sparse factor model.

In the empirical application, we utilize the new approach for the estimation of large covariance matrices of asset returns from stocks that are constituents of the S&P 500 index. Hereby, we consider an empirical horse race that compares the performance of the global minimum variance portfolio strategy based on our sparse factor model to popular portfolio strategies that are commonly used in the literature. The forecasting results reveal that our sparse factor model offers the lowest out-of-sample portfolio standard deviation across different portfolio sizes compared to the considered competing methods. Additionally, it generates the highest out-of-sample portfolio returns and hence provides the certainty equivalent and sharpe ratio.

The outline of the paper is as follows. In Section 2 we introduce the approximate factor model approach and show how sparsity can be obtained with respect to the factor loadings by l_1-regularization. In Section 3 we show that our approach yields consistent parameter estimates of the factor loadings for given regularity conditions and derive the risk bounds for the covariance estimates based on the factor model. Implementation issues are discussed in Section 4, while in Section 5 we show the performance of our approach when applied to the estimation of high dimensional portfolios. Section 6 summarizes the main findings and gives an outlook on future research.

Notation:

Let $\pi_{\max}(A)$ and $\pi_{\min}(A)$ denote the maximum and minimum eigenvalue of a matrix A. Further, $\|A\|$ and $\|A\|_F$ specify the spectral and Frobenius norms of A, respectively. They are defined as $\|A\| = \sqrt{\pi_{\max}(A'A)}$ and $\|A\|_F = \sqrt{\text{tr}\,(A)}$.

2 Model

2.1 The Approximate Factor Model

The following analysis is based on the approximate factor model to obtain a lower dimensional representation of a possibly high dimensional variance-covariance matrix to be estimated. Let x_{it} be the $i-th$ observable variable at time t for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, such that N and T denote the sample size in the cross-section and time series dimensions, respectively. The representation of the approximate factor model is given by:

$$x_{it} = \lambda_i' f_t + u_{it}\,, \tag{1}$$

where λ_i is a $(r \times 1)$-dimensional vector of factor loadings for variable i and f_t is a $(r \times 1)$-dimensional vector of latent factors at time t, where r denotes the number of factors common to all variables in the model. Typically we assume that r is much smaller than the number of variables N. Finally, the idiosyncratic component u_{it} accounts for variable-specific shocks which are not captured by the common component $\lambda_i' f_t$. Following Chamberlain and Rothschild (1983) represents an approximate factor model allowing for limited cross-sectional correlations among the idiosyncratic components. In matrix notation (1) is given by:

$$X = \Lambda F' + u\,, \tag{2}$$

where X denotes a $N \times T$ matrix containing T observations for N strict stationary time series. It is assumed that the time series are demeaned and standardized. $F = (f_1, \ldots, f_T)'$ is referred as a $T \times r$-dimensional matrix of unobserved factors, $\Lambda = (\lambda_1, \ldots, \lambda_N)'$ is a $N \times r$ matrix of corresponding latent factor loadings and u is a $(N \times T)$-dimensional matrix of idiosyncratic shocks.

There are several approaches to estimate a factor model as given by (2). The principal component analysis[1] and the maximum likelihood approach (see i.e. Bai and Li (2016)) are the two

[1] See i.e. Bai and Ng (2002) or Stock and Watson (2002b) for a detailed consideration of the PCA in the approximate factor model

most popular ones. In the following we pursue estimating the factor model by MLE. This allows us to introduce sparsity in the factor loadings by penalizing the likelihood function. Moreover, contrary to PCA all model parameters including the covariance matrix Σ_u can be estimated jointly, while PCA-based second stage estimates of Σ_u require consistent estimation of Λ and F in the first estimation stage. This, however, may be problematic for the case of a relatively small N, because F can no longer be estimated consistently (Bai and Liao (2016)).

The quasi log-likelihood function for the data covariance matrix in the approximate factor model is defined as:

$$\mathcal{L}(\Lambda, \Sigma_F, \Sigma_u) = \log \left|\det \left(\Lambda \Sigma_F \Lambda' + \Sigma_u\right)\right| + \operatorname{tr}\left[S_x \left(\Lambda \Sigma_F \Lambda' + \Sigma_u\right)^{-1}\right], \tag{3}$$

where $S_x = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(x_t - \bar{x})'$ denotes the sample covariance matrix based on the observed data, $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$ and Σ_F is a low dimensional covariance matrix of the factors. Along the lines of Lawley and Maxwell (1971)) we impose the following identification restrictions: $\Sigma_F = I_r$ and $\Lambda'\Sigma_u^{-1}\Lambda$ is diagonal. Moreover, the diagonal entries of $\Lambda'\Sigma_u^{-1}\Lambda$ are assumed to be distinct and arranged in a decreasing order.

The imposition of the identifying restrictions has the advantage, that the estimation of the covariance matrix of the factors becomes redundant. Hence, the log-likelihood function reduces to:

$$\mathcal{L}(\Lambda, \Sigma_u) = \log \left|\det \left(\Lambda \Lambda' + \Sigma_u\right)\right| + \operatorname{tr}\left[S_x \left(\Lambda \Lambda' + \Sigma_u\right)^{-1}\right]. \tag{4}$$

In a second step the factors f_t can be estimated by generalized least squares (GLS):

$$\hat{f}_t = \left(\hat{\Lambda}'\hat{\Sigma}_u\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Sigma}_u^{-1}x_t\,, \tag{5}$$

where the estimates $\hat{\Lambda}$ and $\hat{\Sigma}_u$ are obtained from (4).

2.2 The Sparse Approximate Factor Model

The sparse approximate factor model allows for sparsity in the factor loadings matrix Λ by shrinking single elements of the factor loading matrix Λ to zero. This is obtained by the l_1-norm penalized MLE of (4) based on the following optimization problem:

$$\min_{\Lambda,\Sigma_u} \left[\log \left| \det \left(\Lambda\Lambda' + \Sigma_u \right) \right| + \operatorname{tr} \left[S_x \left(\Lambda\Lambda' + \Sigma_u \right)^{-1} \right] + \mu \sum_{k=1}^{r} \sum_{i=1}^{N} |\lambda_{ik}| \right], \tag{6}$$

where $\mu \geq 0$ denotes regularization parameter. Note, that the number of factors r is predetermined and assumed to be non increasing with N. Sparsity is obtained by shrinking some elements of Λ to zero, such that not all r factors in (1) load on each x_{it}. Hence, this framework allows for weaker factors (see e.g. Onatski (2012)) that affect only a subset of the N time series. In comparison to that, the standard assumption in approximate factor models (see e.g. Bai and Ng (2002), Stock and Watson (2002a)), implies that the r largest eigenvalues of $\Lambda'\Lambda$ are growing with N, and is typically denoted as pervasiveness assumption. Intuitively, this allows only for strong factors that impact the entire set of time series. Consequently, the sparsity in the factor loadings matrix introduced by our model considerably weakens this assumption. In particular, we have:

Assumption 2.1. *There exists a $c > 0$ such that for all N,*

$$c^{-1} < \pi_{\min} \left(\frac{\Lambda'\Lambda}{N^{\beta}} \right) \leq \pi_{\max} \left(\frac{\Lambda'\Lambda}{N^{\beta}} \right) < c,$$

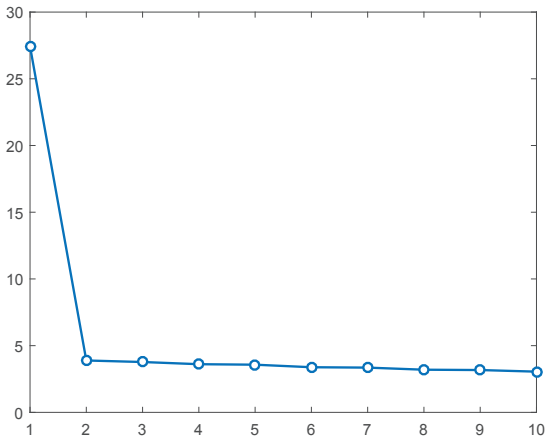
where $0 \leq \beta \leq 1$.

The previous assumption implies that the r largest eigenvalues of $\Lambda'\Lambda$ grow at $\mathcal{O}\left(N^{\beta}\right)$ that might be much slower than in the standard approximate factor model. On the other hand, it still allows for strong factors as β gets closer to 1. Hence, our sparse factor model offers a convenient generalization of the standard factor model. Further, Assumption 2.1 has a direct implication on the sparsity of Λ. In fact, this can be deduced by upper bounding the spectral norm of Λ according to the following expression:

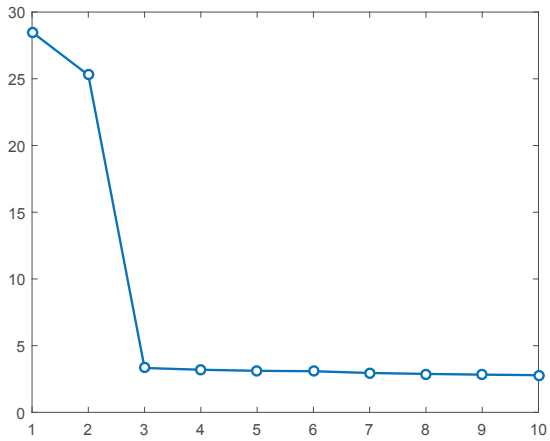
$$\|\Lambda\| \leq \sqrt{r} \|\Lambda\|_1 = \sqrt{r} \max_{k \leq r} \sum_{i=1}^{N} |\lambda_{ik}| = \mathcal{O}\left(N^{\beta/2}\right).$$

This result shows that imposing the weak factor assumption limits the amount of affected time series across all factors and hence requires a non-negligible amount of zero elements in each column of the factor loadings matrix. Nevertheless, the amount of non-zero factor loadings can be arbitrarily small as β increases.

The pervasiveness assumption imposed by the standard approximate factor model, further implies a clear separation of the eigenvalues of the data covariance matrix into two groups, corresponding to the diverging eigenvalues of the common component and the bounded eigenvalues of the idiosyncratic errors. Those characteristics can be observed in Figure 1, where both panels illustrate the eigenvalue distribution of datasets that are simulated only based on strong factors for sample sizes $N = 200$ and $T = 450$. The panels differ solely in the amount of included factors, where the left panel includes one strong factor and the right one considers two factors. Both graphs reveal a clear partition in their respective eigenvalue distributions, into sets of diverging eigenvalues corresponding to the amount of included strong factors and sets of bounded eigenvalues associated to the idiosyncratic components.



(a) Eigenvalues for simulated data with 1 strong factor with $N = 200$ and $T = 450$



(b) Eigenvalues for simulated data with 2 strong factors with $N = 200$ and $T = 450$

Figure 1: Distribution of the eigenvalues based on strong factors

However, such a clear separation in the eigenvalue distribution of the covariance matrix is typically not found in real datasets. An example offers a dataset that contains the monthly asset returns of 200 stocks constituents of the S&P 500 stock index available for the entire period of 450 months[2], whose eigenvalue distribution is illustrated in Figure 2. The graph

[2] The identical dataset is also used in our empirical application and hence is introduced in more detail in Section 5.

shows a clear distinction of the first eigenvalue, however the remaining ones diverge at slower rates and a clear separation between the common and idiosyncratic component as implied by the standard approximate factor model is not possible. Hence, the weak factor framework that allows for slower divergence rates in the eigenvalues of the common component is more adequate for modelling the eigenvalue structures of real datasets. In fact, Figure 3 that shows the eigenvalue distribution of a dataset that was generated by one strong factor and three weak factors, closely models the decaying eigenvalue structure we observed for the S&P 500 asset returns.

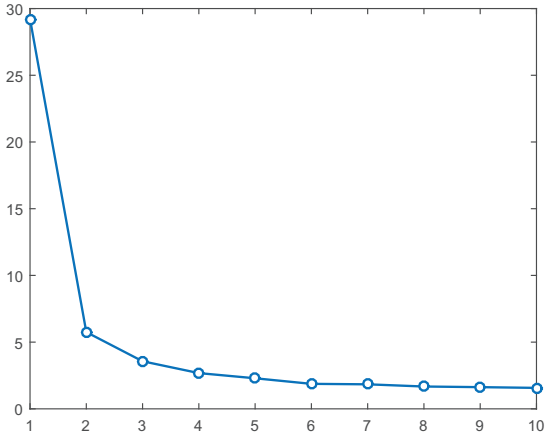


Figure 2: Eigenvalues for stock returns based on the S&P 500 index with $N = 200$ and $T = 450$

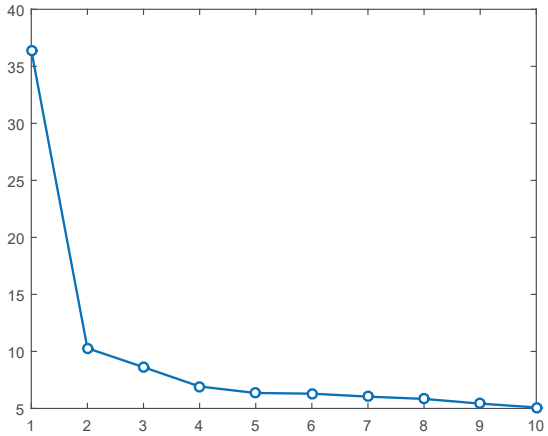


Figure 3: Eigenvalues for simulated data with 1 strong factor and 3 weak factors with $N = 200$ and $T = 450$

2.3 Estimation of the idiosyncratic error covariance matrix Σ_u

In order to allow for cross-sectional correlation in idiosyncratic error covariance matrix and hence to be in the context of an approximate factor model, we use a slight adaptation of the principal orthogonal complement thresholding (POET) estimator by Fan, Liao, and Mincheva (2013). The POET estimator is based on soft-thresholding the off-diagonal elements of the sample covariance matrix of the factor model residuals. Hence, it introduces sparsity in the idiosyncratic covariance matrix and offers an solution to the singularity problem, generated by sample covariance estimator, especially if we are in a high-dimensional setting, where N is close or even greater than T. More specifically, the estimated idiosyncratic error covariance matrix

$\hat{\Sigma}_u^\tau$ based on the POET method is defined as:

$$\hat{\Sigma}_u^\tau = \hat{\sigma}_{ij}^\tau, \quad \hat{\sigma}_{ij}^\tau = \begin{cases} \hat{u}_{ii}, & i = j \\ S(\hat{u}_{ij}, \tau_{ij}), & i \neq j \end{cases}$$

where $\hat{u}_{ij}$ is the ij-th element of the sample covariance matrix $\hat{U}$ of the estimated factor model residuals, $\tau_{ij} > 0$ is an entry-dependent threshold[3] and S denotes the soft-thresholding operator that is defined as:

$$S(Z, \beta)_{i,j} = \text{sign}(z_{i,j})(|z_{i,j}| - \beta)_+. \tag{7}$$

In comparison to the paper by Fan et al. (2013) that use the residuals of a static factor model based on the PCA estimator, we use the residuals obtained by our sparse factor model.

2.4 Data covariance matrix estimation based on the Sparse Approximate Factor Model

In this section we introduce the variance-covariance estimator according to our sparse approximate factor model. For this we consider the factor model representation in equation (1) and calculate the true covariance matrix for X. This leads to the following expression:

$$\begin{aligned} \Sigma_{\text{SF}} &= \mathbf{V}[X] = \mathbf{V}\left[\Lambda F'\right] + \mathbf{V}[u] \\ &= \Lambda \Sigma_F \Lambda' + \Sigma_u \end{aligned}$$

In order to obtain an estimate for Σ_{SF} we first need estimates for the factors F and factor loadings Λ according to our sparse factor model. These estimates are obtained by applying the estimation method introduced in the sections 2.2 and **??** for the dataset X. As an estimate for the idiosyncratic errors u, we use the regression residuals and apply the utilize the soft-thresholding method introduced in section 2.3.

[3] We fix $\tau_{ij} = \frac{1}{\sqrt{N}} + \sqrt{\frac{\log(N)}{T}}$ as specified by Fan et al. (2013) in all our applications.

Using this results, we construct the estimated covariance matrix according to:

$$\hat{\Sigma}_{\text{SF}} = \hat{\Lambda}\hat{\Sigma}_F\hat{\Lambda}' + \hat{\Sigma}_u^{\tau}, \tag{8}$$

where $\hat{\Sigma}_F$ denotes the sample estimator for the covariance matrix of the estimated dynamic factors.

3 Large Sample Properties and Risk Bounds

3.1 Consistency of the Sparse Factor Model Estimator

In order to establish the consistency estimators of the factor loading matrix Λ and the data variance covariance matrix Σ we adopt the following typical assumptions:

Assumption 3.1. *(i)* $\{u_t, f_t\}_{t\geq 1}$ *is strictly stationary. In addition,* $\mathbf{E}[u_{it}] = \mathbf{E}[u_{it}f_{kt}] = 0$, *for all* $i \leq N$, $k \leq r$ *and* $t \leq T$.

(ii) There exists $r_1, r_2 > 0$ *and* $b_1, b_2 > 0$, *such that for any* $s > 0$, $i \leq N$ *and* $k \leq r$,

$$\mathbf{P}(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1}), \quad \mathbf{P}(|f_{kt}| > s) \leq \exp(-(s/b_2)^{r_2})$$

(iii) Define the mixing coefficient:

$$\alpha(T) = \sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_T} |\mathbf{P}(A)\mathbf{P}(B) - \mathbf{P}(AB)|,$$

where $\mathcal{F}^0_{-\infty}$ *and* $\mathcal{F}^\infty_T$ *denote the* σ*-algebras generated by* $\{(f_t, u_t) : -\infty \leq t \leq 0\}$ *and* $\{(f_t, u_t) : T \leq t \leq \infty\}$

Strong mixing: There exist $r_3 > 0$ *and* $C > 0$ *satisfying: for all* $T \in \mathcal{Z}^+$,

$$\alpha(T) \leq \exp(-CT^{r_3})$$

(iv) There exist constants $c_1, c_2 > 0$ *such that* $c_2 \leq \pi_{\min}(\Sigma_{u0}) \leq \pi_{\max}(\Sigma_{u0}) \leq c_1$.

The assumptions in 3.1 impose the regularity conditions on the data generating process and are identical to those imposed by Bai and Liao (2016). Condition (i) imposes strict stationarity for u_t and f_t and requires that both terms are not correlated. Condition (ii) requires exponential-type tails, which allows to use large deviation theories for $\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - \sigma_{u,ij}$ and $\frac{1}{T}\sum_{t=1}^T f_{jt}u_{it}$. In order to allow for weakly serial dependence, we impose a strong mixing condition specified in Condition (iii). Further, Condition (iv) implies bounded eigenvalues of the idiosyncratic error covariance matrix, which is a common assumption on factor models.

Theorem 3.1 (Consistency)

Under Assumption 2.1 and Assumption 3.1, the sparse factor model in (6) satisfies for T and $N \to \infty$, the following properties:

$$\frac{1}{N}\left\|\hat{\Lambda} - \Lambda_0\right\|_F^2 = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N} + d_T\right)$$

and

$$\frac{1}{N}\left\|\hat{\Sigma}_u - \Sigma_{u0}\right\|_F^2 = \mathcal{O}_p\left(\frac{\log N}{T} + d_T\right),$$

where $d_T = \left(\frac{\log N^\beta}{N} + \sqrt{\frac{\log N}{T}}\right)$, for $0 \leq \beta \leq 1$ and D_N denotes the amount of non-zero elements in Λ.

Hence, for $\log(N) = o(T)$, we have

$$\frac{1}{N}\left\|\hat{\Lambda} - \Lambda_0\right\|_F^2 = o_p(1), \qquad \frac{1}{N}\left\|\hat{\Sigma}_u - \Sigma_{u0}\right\|_F^2 = o_p(1).$$

Furthermore, for all $t \leq T$:

$$\left\|\hat{f}_t - f_t\right\| = o_p(1)$$

Theorem 3.1 shows under some regularity conditions average consistency for the factor loadings and idiosyncratic error covariance matrix estimator based on our sparse factor model. Further, the factors f_t estimated based on GLS are also consistently estimated.

3.2 Consistency of the Data Covariance Matrix Estimator

This section takes a closer look to the asymptotic properties of the data covariance estimator based on the sparse factor model, elaborated in section 2.4. Hereby, we first introduce the relative matrix error norm, suggested by Fan et al. (2013), that is defined as

$$\|A\|_\Sigma = \frac{1}{\sqrt{N}}\left\|\Sigma^{-1/2} A \Sigma^{-1/2}\right\|_F \tag{9}$$

The following theorem gives the convergence rates of the data covariance matrix estimator and its inverse under different matrix norms.

Theorem 3.2 (Risk Bounds)

Under Assumption 2.1 and Assumption 3.1, the data covariance matrix estimator based on the sparse factor model satisfies for $T,\ N \to \infty$ and $d_T = \left(\frac{\log N^\beta}{N} + \sqrt{\frac{\log N}{T}}\right)$, for $0 \leq \beta \leq 1$, the following properties:

$$\frac{1}{N}\left\|\hat{\Sigma}_{SF} - \Sigma_0\right\|_F^2 = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N} + d_T + \left(\frac{1}{N} + \frac{\log N}{T}\right) m_N^2\right),$$

$$\frac{1}{N}\left\|\hat{\Sigma}_{SF} - \Sigma_0\right\|_\Sigma^2 = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N^2} + \frac{d_T}{N} + \left(\frac{1}{N} + \frac{\log N}{T}\right) \frac{m_N^2}{N^2}\right)$$

and

$$\frac{1}{N}\left\|\hat{\Sigma}_{SF}^{-1} - \Sigma_0^{-1}\right\|^2 = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N} + d_T + \left(\frac{1}{N} + \frac{\log N}{T}\right) \frac{m_N^2}{N}\right)$$

Theorem 3.2 shows that the covariance matrix estimator based on the sparse factor model is consistently estimated if we consider either the Frobenius norm or the weighted quadratic norm. In comparison to this result, a convergence under the Frobenius norm is hardly achievable based on a factor model estimated using PCA. Fan et al. (2013) show that this fact occurs because of the diverging eigenvalues of the factor loadings matrix.

4 Implementation Issues

4.1 Majorized Log-Likelihood Function

In this section we elaborate the details regarding the implementation of our proposed sparse factor model. If we take a closer look at our objective function in (6), we notice that it is cumbersome to solve that function numerically. This issue arises from the fact that the first term in (6) $\log\left|\det\left(\Lambda\Lambda' + \Sigma_u\right)\right|$ is concave in Λ and Σ_u, whereas the second term $\operatorname{tr}\left[S_x\left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right]$ is convex. In order to solve this issue we are referring to a majorize-minimize EM algorithm introduced by Bien and Tibshirani (2011). The idea of this optimization approach is to approximate the numerically unstable concave part $\log\left|\det\left(\Lambda\Lambda' + \Sigma_u\right)\right|$ by its tangent plane, which corresponds to the following expression:

$$\log\left|\det\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Sigma}_{u,m}\right)\right| + \operatorname{tr}\left[2\hat{\Lambda}_m'\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Sigma}_{u,m}\right)^{-1}\left(\Lambda - \hat{\Lambda}_m\right)\right], \tag{10}$$

where the subscript m denotes the mth step in an iterative procedure that we elaborate in the following sections. If we replace the concave part in (4) by the convex expression in (10), we get the following majorized log-likelihood function:

$$\begin{aligned}\bar{\mathcal{L}}(\Lambda) = &\log\left|\det\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Sigma}_{u,m}\right)\right| + \operatorname{tr}\left[2\hat{\Lambda}_m'\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Sigma}_{u,m}\right)^{-1}\left(\Lambda - \hat{\Lambda}_m\right)\right] \\ &+ \operatorname{tr}\left[S_x\left(\Lambda\Lambda' + \hat{\Sigma}_u\right)^{-1}\right]\end{aligned} \tag{11}$$

If we augment the previous majorized log-likelihood by a L_1-regularization, we end up with the following optimization problem for our sparse approximate factor model:

$$\begin{aligned}\min_{\Lambda,\Sigma_u}\Bigg[&\log\left|\det\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Sigma}_{u,m}\right)\right| + \operatorname{tr}\left[2\hat{\Lambda}_m'\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Sigma}_{u,m}\right)^{-1}\left(\Lambda - \hat{\Lambda}_m\right)\right] \\ &+\operatorname{tr}\left[S_x\left(\Lambda\Lambda' + \hat{\Sigma}_u\right)^{-1}\right] + \mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik}|\Bigg]\end{aligned} \tag{12}$$

As all three components in (12) are convex, the optimization is much simpler compared to the original multi-modal problem.

4.2 Projection Gradient Algorithm

In order to solve the estimation problem in (12) efficiently, we utilize a fast projected gradient algorithm that has also been pointed out by Bien and Tibshirani (2011). More specifically, we can further approximate the majorized log-likelihood $\bar{\mathcal{L}}(\Lambda)$ in (11) by the following expression:

$$\tilde{\mathcal{L}}(\Lambda) = \frac{1}{2t} \left\| \Lambda - \hat{\Lambda}_m + t\hat{A} \right\|_F^2$$

where t is the depth of projection[4] and

$$\hat{A} = 2 \left[\left(\hat{\Lambda}_m \hat{\Lambda}_m' + \hat{\Sigma}_{u,m} \right)^{-1} - \left(\hat{\Lambda}_m \hat{\Lambda}_m' + \hat{\Sigma}_{u,m} \right)^{-1} S_x \left(\hat{\Lambda}_m \hat{\Lambda}_m' + \hat{\Sigma}_{u,m} \right)^{-1} \right] \hat{\Lambda}_m, \tag{13}$$

which corresponds to the first derivative of $\bar{\mathcal{L}}(\Lambda)$ with respect to Λ.

Hence, our final optimization problem corresponds to:

$$\min_{\lambda_{ik}} \frac{1}{2t} \sum_{k=1}^{r} \sum_{i=1}^{N} \left(\lambda_{ik} - \hat{\lambda}_{ik,m} + t\hat{A}_{ik,m} \right)^2 + \mu \sum_{k=1}^{r} \sum_{i=1}^{N} |\lambda_{ik}| \,. \tag{14}$$

4.3 Solution for the Sparse Approximate Factor Model

In this section we propose an iterative procedure to find a solution for our sparse approximate factor model.

The optimization of the objective function in equation (14) can be carried out by computing its first derivative with respect to λ_{ik} and setting it to zero. This yields the following result:

$$\begin{aligned} \frac{\partial}{\lambda_{ik}} & \left[\frac{1}{2t} \sum_{k=1}^{r} \sum_{i=1}^{N} \left(\lambda_{ik} - \hat{\lambda}_{ik,m} + t\hat{A}_{ik,m} \right)^2 + \mu \sum_{k=1}^{r} \sum_{i=1}^{N} |\lambda_{ik}| \right] \\ & = \frac{1}{t} \sum_{k=1}^{r} \sum_{i=1}^{N} \left(\hat{\lambda}_{ik} - \hat{\lambda}_{ik,m} + t\hat{A}_{ik,m} \right) + \mu \sum_{k=1}^{r} \sum_{i=1}^{N} \nu_{ik} = 0, \end{aligned}$$

[4] We set $t = 0.01$ in all our studies

where ν_{ik} denotes the subgradient of $|\lambda_{ik}|$.

If we now solve for a specific $\hat{\lambda}_{ik}$, we get:

$$\begin{aligned} \hat{\lambda}_{ik} + t \cdot \mu \nu_{ik} &= \hat{\lambda}_{ik,m} - t\hat{A}_{ik,m} \\ \hat{\lambda}_{ik} &= S\left(\hat{\lambda}_{ik,m} - t\hat{A}_{ik,m}, \quad t \cdot \mu\right), \end{aligned} \tag{15}$$

where S denotes the soft-thresholding operator and is defined as in equation (7). The result in equation (15) can be used to obtain an updated estimate for the factor loadings $\hat{\lambda}_{ik,m+1}$ given the estimate in the previous step $\hat{\lambda}_{ik,m}$.

In order to obtain an update for the estimate of the covariance matrix of the idiosyncratic error Σ_u, we can rely on the EM algorithm elaborated by Bai and Li (2012), which yield the following specification:

$$\hat{\Sigma}_{u,m+1} = \text{diag}\left[S_x - \hat{\Lambda}_{m+1}\hat{\Lambda}'_m\left(\hat{\Lambda}_m\hat{\Lambda}'_m + \hat{\Sigma}_{u,m}\right)^{-1} S_x\right]$$

Hence, in order to obtain an estimate of our sparse approximate factor model we can utilize the following iterative procedure:

Iterative Algorithm

Step 1: Obtain an initial consistent estimate for the factor loading matrix Λ_0 and the idiosyncratic error covariance matrix Σ_u , i.e. by using the unpenalized maximum likelihood approach pointed out in Section **??** and set $m = 1$.

Step 2: Update $\hat{\lambda}_{ik,m-1}$, by

$$\hat{\lambda}_{ik,m} = S\left(\hat{\lambda}_{ik,m-1} - t\hat{A}_{ik,m-1}, \quad t \cdot \mu\right)$$

Step 3: Update $\hat{\Sigma}_u$ using the EM algorithm in Bai and Li (2012), according the following formula

$$\hat{\Sigma}_{u,m} = \text{diag}\left[S_x - \hat{\Lambda}_m\hat{\Lambda}'_{m-1}\left(\hat{\Lambda}_{m-1}\hat{\Lambda}'_{m-1} + \hat{\Sigma}_{u,m-1}\right)^{-1} S_x\right]$$

Step 4: If $\left\|\hat{\Lambda}_m - \hat{\Lambda}_{m-1}\right\|$ is small enough, stop the procedure.
Otherwise set $m = m + 1$ and go back to Step 2.

Step 5: Get an estimate for the factors according to the following function:

$$\hat{f}_t = \left(\hat{\Lambda}'\hat{\Sigma}_u\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Sigma}_u^{-1}x_t\,,$$

where $\hat{\Lambda}$ and $\hat{\Sigma}_u$ denote the converged parameter estimates.

Step 6: Re-estimate the covariance matrix of the idiosyncratic errors based on the procedure introduced in Section 2.3.

4.4 Selecting the number of factors

In order to obtain an estimate for the number of latent factors r we use the method by Onatski (2010). This approach utilizes the difference in subsequent eigenvalues and chooses the largest $\hat{r}$ that belongs to the following set:

$$\{\hat{r} \leq r_{\max} : \pi_{\hat{r}}((X'X)/T) - \pi_{\hat{r}+1}((X'X)/T) > \xi\},$$

where ξ is a fixed positive constant, $r_{\max}$ is an upper bound for the possible number of factors and $\pi_r((X'X)/T)$ denotes the $\hat{r}^{\text{th}}$ largest eigenvalue of the covariance matrix of X.
The author takes for the choice of ξ the empirical distribution of the eigenvalues of the data sample covariance matrix into account[5]. Hence, the method is also a robust estimator for the case if we are confronted with correlated error terms. On the other hand, the estimation of the number of factors based on the empirical distribution of the eigenvalues of the sample covariance matrix, requires a clear separation of the eigenvalues from the common and idiosyncratic component. Hence, the performance may depend on the degree of differentiability of the two components. Nevertheless, the main reason of choosing the method by Onatski (2010) arises from the fact that it is to the best of our knowledge the only method that does not explicitly require that all factors are strong and have eigenvalues that diverge with a rate of N. Therefore, it is suitable for our setting that allows also for weak factors with a slower divergence rate.

[5] We refer to the paper of Onatski (2010) for the detailed procedure to determine ξ.

4.5 Tuning parameter choice

It is great importance to select an optimal value for the shrinkage parameter μ, as it controls the amount of sparsity in the factor loadings matrix and impacts the efficiency of our estimator. In our case we estimate μ based on a bayesian type of information criterion, according to the following formula:

$$IC(\mu) = \mathcal{L}\left(\hat{\Lambda}, \hat{F}\right) + 2\kappa_\mu \frac{N+T}{N \cdot T} \log\left(\frac{N \cdot T}{N+T}\right) \tag{16}$$

where κ_μ denotes the number of non-zero elements in the factor loading matrix Λ for a given value of μ and $\mathcal{L}\left(\hat{\Lambda}, \hat{F}\right)$ is the value of the log-likelihood function in equation (3), evaluated at the estimated factors and factor loadings. The penalty function in (16) is identically specified as in the IC_{p1} criterion by Bai and Ng (2002) and has the property of converging to zero as both N and T increase to infinity. Hence, the penalization vanishes as the sample sizes increase and a smaller value for μ is selected. The characteristics of our information criterion are therefore convenient with respect to the asymptotic properties we require for the regularization parameter μ. In fact, we need $\mu = o(1)$ in order to achieve estimation consistency, as elaborated in Section 3. To select the optimal μ, we estimate the criterion in (16) for a grid of different values for μ and choose the one that minimizes our information criterion. For the grid of the shrinkage parameter we consider the interval $\mu = (0, \mu_{\max})$, where $\mu_{\max}$ denotes the highest value for the shrinkage parameter such that all imposed model restrictions are still fulfilled.

5 An Application to Portfolio Choice

In this section we apply our new approach for the estimation of large covariances of asset returns. In an empirical horse race we compare the performance of the global minimum variance portfolio strategy based on our sparse factor model approach to popular alternative portfolio strategies with regularized variance-covariance estimators.

5.1 Data and Description of the Forecasting Experiment

The dataset comprises the monthly excess return data of stocks of the S&P 500 index, that were constituents of the index in April, 2015. The excess returns are obtained by subtracting the corresponding one-month T-Bill rate from the asset returns. We consider the time period from January, 1974 until April, 2015, which yields $T = 496$ monthly returns for each of the 205 available stocks[6]. As different investment dimension we use the following portfolio sizes: $N \in \{30, 50, 100, 150, 200\}$. Out of the 205 stocks, we then select the individual subsamples N at random and keep the selected asset fixed for the entire forecasting experiment.

In order to estimate the portfolio weights for each strategy, we apply a rolling window approach with $h = 60$ months, corresponding to 5 years of past data. Thus, in time t we use the last 60 months from $t-59$ until t for our estimation. Using the estimated portfolio weights we compute the out-of-sample portfolio return $\hat{r}_{s,t+1}$ for the period $t+1$ for method s. All portfolios are rebalanced on a monthly basis. This generates a series of $T-h$ out-of-sample portfolio returns. These results are used to estimate the mean μ_s and variance σ_s^2 of the portfolio returns for each strategy, according to the following formulas:

$$\hat{\mu}_s = \frac{1}{T}\sum_{t=h+1}^{T} \hat{r}_{t,s} \qquad \text{and} \qquad \hat{\sigma}_s^2 = \frac{1}{T-1}\sum_{t=h+1}^{T} (\hat{r}_{t,s} - \hat{\mu}_s)^2 \tag{17}$$

We repeat this procedure 100 times to avoid that the out-of-sample results depend on the initially randomly selected stocks and make sure that only distinct portfolios are considered for each forecasting experiment. Hence, all results that we later report are average outcomes across the 100 forecasting experiments.

[6] The return data are retained from Thompson Reuters Datastream.

5.2 Criteria for Performance Evaluation

For our analysis, we consider several evaluation criteria in order to compare the performance of the previously introduced models.

1. **Standard Deviation (SD):** The out-of-sample standard deviation is defined as the square root of the variance $\hat{\sigma}^2$ in equation (17).
   In terms of the GMV portfolio is designed to minimize the portfolio variance, therefore this measure is of great importance.

2. **Average Return (AV):** The out-of-sample average return is expressed as $\hat{\mu}$ in equation (17).

3. **Certainty Equivalent (CE):** The CE is defined as

$$\widehat{\mathrm{CE}}_s = \hat{\mu}_s - \frac{\gamma}{2} \cdot \hat{\sigma}_s^2, \tag{18}$$

   where γ specifies the risk aversion of the investor. Similar as in DeMiguel, Garlappi, Nogales, and Uppal (2009a) we set $\gamma = 2$. The CE determines the risk-free rate that an investor is willing to accept instead of investing into a particular risky portfolio strategy.

4. **Sharpe Ratio (SR):** The Sharpe ratio is calculated using the following formula:

$$\widehat{\mathrm{SR}}_s = \frac{\hat{\mu}_s}{\hat{\sigma}_s}. \tag{19}$$

To analyse the properties of the estimated portfolio weights over time, we consider the following measures:

5. **Min:** Minimum portfolio weight

6. **Max:** Maximum portfolio weight

7. **SD-W:** Standard deviation of the portfolio weights computed as

$$\sum_{i=1}^{N} \left(\hat{\omega}_{i,t,s} - \hat{\mu}_{\hat{\omega}_{t,s}}\right)^2,$$

where $\hat{\mu}_{\hat{\omega}_{t,s}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\omega}_{i,t,s}$.

8. **MAD**: Mean absolute deviance from the $1/N$ portfolio

$$\text{MAD}_{t,s} = \sum_{i=1}^{N}\left|\hat{\omega}_{i,t,s} - \frac{1}{N}\right|$$

5.3 Considered estimation models

This section discusses the different strategies in portfolio optimization, that are used and compared to our sparse factor model.

1. **Naive portfolio** $(1/N)$

The first strategy that we consider is the naive portfolio, which comprises an identical portfolio weight of $\omega_{1/N} = \frac{1}{N}$ in each of the N risky assets. Therefore, it is not necessary to estimate any moments of the data to gather the portfolio weights. Since this moments estimation causes in many application a high risk and error, it is difficult to beat the naive portfolio. DeMiguel, Garlappi, and Uppal (2009b) find that the mean-variance portfolio and most of its extensions cannot significantly outperform this portfolio. For this reason we use this strategy as good benchmark for the following portfolio allocation models.

2. **Sample-based GMV portfolio**

The portfolio weights based on the sample covariance matrix of the asset returns are also considered. The main problem using this approach arises for the case when the number of assets N is larger or close to the time series dimension T, since the resulting sample covariance is very ill-conditioned, singular and therefore not invertible.

3. **Approximate Factor Model**

In order to have a comparison to a standard factor model with a dense factor loadings matrix, we utilize the approximate factor model as introduced by Chamberlain and Rothschild (1983). An estimate for the covariance matrix using this model, denoted as $\hat{\Sigma}_{\text{AFM}}$, is similarly obtained as pointed out in section 2.4. The main difference lies in the fact that the necessary estimates for the factors F and factor loadings Λ are obtained by principal component analysis. A precise

description of the estimation of an approximate factor model can be found in Bai and Ng (2002). Similar as in our sparse approximate factor model, we use the number of factors selected by Onatski (2010) to estimate this model.

4. **Dynamic Factor Model**

To allow for some dynamics in the latent factors, we consider also a dynamic factor model originally proposed by Geweke (1977). Specifically, the dynamic factor model is represented by the following equation:

$$x_{it} = B_i'(L) f_t + \varepsilon_{it} \tag{20}$$

where $B_i(L) = (b_{i1} + b_{i2}L + \cdots + b_{ip}L^p)$ and L corresponds to the lag operator such that, $\forall p$, $L^p f_t = f_{t-p}$.

In this setup $\mathbf{f}_t$ is a $q \times 1$-dimensional vector of dynamic factors, i.e. $f_t = (f_{1t}, f_{2t}, \ldots, f_{qt})'$ that offer their own dynamic structure according to a VAR process and b_{ij}, where $j = 1, \ldots, p$ are the corresponding q-dimensional factor loadings.

In order to estimate the dynamic factor model in (20) we use the two step procedure by Doz, Giannone, and Reichlin (2011). To be able to estimate the model it is necessary to know the number of dynamic factor beforehand. We use the consistent method by Bai and Ng (2007) to determine q.

Factor models with observed factors

Further, we consider two additional factor model approaches that are common in the finance literature. The difference to the previously presented methodology lies in the specification and estimation of the factors. In comparison to the approximate factor case, the models in this section are not estimating the factors from the time series data, but they utilize a concrete specification for the factors. A positive aspect of this methodology compared to the latent factor model case is the reduction of the estimation noise as the factor parameters have not to be estimated. On the other hand, there is the sever possibility that the factor specification is not capturing properly the data generating process and leads to a model misspecification.

The first model we are focusing on is the single index model by Sharpe (1963) and the second

one is the 3-factor model by Fama and French (1993).

5. **The Single Index Model**

The single index model by Sharpe (1963) is defined by the following equation:

$$x_{it} = \alpha + \beta_{i1} f_{1t} + \varepsilon_{it}, \tag{21}$$

where $t = 1, \ldots, T$ and $i = 1, \ldots, N$.

In equation (21) the factor f_{1t} represents the excess return on the proxy for the US equity market portfolio at time t. This proxy is defined as a value-weighted return of all Center for Research in Security Prices (CRSP) firms incorporated in the US and listed on the AMEX, NASDAQ, or the NYSE. In order to obtain excess returns the corresponding one-month T-Bill rate is subtracted from all returns. The estimator for the variance-covariance matrix of the single index model is obtained by the following expression:

$$\hat{\Sigma}_{\text{SFM}} = \hat{\beta}_1 \hat{\Sigma}_{f_1} \hat{\beta}_1' + \hat{D}$$

where $\hat{\Sigma}_{f_1}$ denotes the sample variance-covariance matrix of the market excess returns, $\hat{\beta}_1$ represents the OLS estimates of the factor loadings and $\hat{\mathbf{D}}$ is a diagonal matrix of the OLS residual variances of regression model (21).

6. **Fama and French 3-Factor Model (FF)**

The Fama and French 3-factor model offers an extension to the single index model by Sharpe (1963) and is defined by the following expression:

$$X_t = \beta_1 f_{1t} + \beta_2 f_{2t} + \beta_3 f_{3t} + \varepsilon_t \tag{22}$$

The first factor f_1 is identically defined as in equation (21) and corresponds to the excess returns on the proxy for the US equity market portfolio, i.e. the market premium.

The second factor f_2 defined as SMB is composed as the average returns on the three small portfolios minus the average returns on the three big portfolios. In particular, it defines a zero-cost portfolio that is long in stocks with a small market capitalization and short in stocks with a

large market capitalization[7]. The third factor f_3 denoted as HML comprises a zero-cost portfolio that is long in stocks with a high book-to-market value and short in low book-to-market stocks [8].

If we put equation 22 into matrix notation we get the following expression:

$$X = \beta F' + \varepsilon \tag{23}$$

where $F = [f_1, f_2, f_3]$ with dimension $T \times 3$ and $\beta = [\beta_1, \beta_2, \beta_3]$ with dimension $N \times 3$.
The estimator for the variance-covariance matrix for the 3-factor model by Fama and French (1993) Σ_{FF} is equal to the following equation:

$$\hat{\Sigma}_{\text{FF}} = \hat{\beta}\hat{\Sigma}_F\hat{\beta}' + \hat{D}$$

where $\hat{\Sigma}_F$ denotes the covariance matrix of the three factors and $\hat{D}$ represents a diagonal matrix that contains the variances of the OLS residuals covariance matrix on its main diagonal.

Covariance Matrix Estimation Strategies

In the following sections we draw our attention to methods that introduce efficient variance-covariance matrix estimators. Hereby, we focus on two shrinkage strategies by Ledoit and Wolf (2003) and Kourtis, Dotsis, and Markellos (2012) and the design-free estimator by Abadir, Distaso, and Žikeš (2014).

7. **Ledoit and Wolf (2003)**

The authors propose a shrinkage approach that offers a combination of sample covariance matrix $\hat{\Sigma}_X$ with the covariance matrix of a target estimator $\hat{\Sigma}_{\text{target}}$ that is well-conditioned. They create their covariance estimator according to the following definition:

$$\hat{\Sigma}_{\text{LW}} = \alpha\hat{\Sigma}_X + (1-\alpha)\hat{\Sigma}_{\text{target}} \tag{24}$$

[7] It is important to note that securities with a long position in a portfolio are expected to rise in value and on the other hand securities with short positions in a portfolio are expected to decline in value.

[8] A detailed definition of the factors can be found on the website of Kenneth R. French. See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

where $\alpha \in (0,1)$ is a constant, which corresponds to the shrinkage parameter. The authors show that using this method can be interpreted as shrinking the sample covariance matrix towards the estimator $\hat{\Sigma}_{\text{target}}$.

The covariance matrix from a 1-factor model conduces as a target estimator.

From equation (24) we can see that it is essential to determine an estimator for the shrinkage intensity α. Hereby, Ledoit and Wolf (2003) come up with a procedure that relies on minimizing the Frobenius norm of the difference between the shrinkage estimator $\hat{\Sigma}_{\text{LW}}$ and the true covariance matrix.

8. **Kourtis et al. (2012)**

The estimation method by Kourtis et al. (2012) is directly shrinking the inverse of the sample-based variance-covariance matrix $\hat{\Sigma}_X$ towards a well-conditioned target matrix $\hat{\Omega}$, according to the following equation:

$$\hat{\Sigma}_{\text{K}}^{-1} = c_1 \hat{\Sigma}_X^{-1} + c_2 \hat{\Omega} \tag{25}$$

The authors consider three different targets for $\hat{\Omega}$:

1. The identity matrix I, which offers a very simple constant structure
2. The inverse of the covariance matrix resulting from a 1-factor model by Sharpe (1963)
3. A combination of both targets, which yields

$$\hat{\Sigma}_{\text{K}}^{-1} = c_1 \hat{\Sigma}_X^{-1} + c_2 I_N + c_3 \hat{\Sigma}_{f_1}^{-1} \tag{26}$$

The authors show that the resulting weights are a three-fund strategy, i.e. a linear combination of the sample-based weights $\hat{\omega}$, the equally weighted portfolio weights $\hat{\omega}_{1/N}$ and those of the 1-factor model $\hat{\omega}_f$. In order to select an optimal quantity for the shrinkage coefficients in the equations 25 respectively 26, the authors are minimizing the out-of-sample portfolio variance using the non-parametric method of cross-validation.

It is important to mention that this portfolio strategy is also applicable for the case when

$N > T$. In order to obtain reliable results for the inverse of $\hat{\Sigma}_X$ the authors use the Moore-Penrose pseudo-inverse. For our empirical application we report the results for the three fund strategy, which offered the best performance in terms of certainty equivalent and Sharpe ratio across all datasets, compared to the first two shrinkage targets individually.

9. **Abadir et al. (2014)**

The design-free estimator for the variance-covariance matrix by Abadir et al. (2014) yields to improve the estimation of the eigenvalues P for the sample covariance matrix of X, that is a possible source of ill-conditioning compared to the orthogonal eigenvectors Γ that are not affected by this problem by construction. The authors consider the following spectral decomposition of $\hat{\Sigma}_X$

$$\mathbf{V}[X] = \hat{\Sigma}_X = \hat{\Gamma}\hat{P}\hat{\Gamma}' \tag{27}$$

In order to obtain an improved estimator for $\mathbf{P}$ the data series X is split into two subsample as in (28)

$$X = \underset{N\times n}{\left(X_1\right.}, \underset{N\times(T-n)}{\left.X_2\right)} \tag{28}$$

Calculating the sample covariance matrix for the first n observations yields

$$\mathbf{V}[X_1] = \hat{\Sigma}_1 = \frac{1}{n} X_1 M_n X_1' = \hat{\Gamma}_1 \hat{P}_1 \hat{\Gamma}_1' \tag{29}$$

where $M_n = I_n - \frac{1}{n} 1_n 1_n'$ is the de-meaning matrix of dimension n and 1_n denotes a $n \times 1$ vector of ones. The spectral decomposition of $\hat{\Sigma}_1$ yields the matrix of eigenvectors $\hat{\Gamma}_1$ and the diagonal matrix of eigenvalues $\hat{P}_1$.

In the second step an improved estimator for the P is computed from the remaining orthogonalized observations

$$\tilde{P} = \operatorname{diag}\left(\mathbf{V}\left[\hat{\Gamma}_1' X_2\right]\right) = \operatorname{diag}\left(\hat{\Gamma}_1' \Sigma_2 \hat{\Gamma}_1\right) \tag{30}$$

The new estimator for the variance-covariance matrix is now obtained according to

$$\hat{\Sigma}_{\mathrm{A}} = \hat{\Gamma}\tilde{P}\hat{\Gamma}' \tag{31}$$

Weight Shrinking Strategies

The following asset allocation methods are focusing on directly shrinking the estimated portfolio weights.

10. **Frahm and Memmel (2010)**

The first portfolio strategy that we are considering, comprises the method by Frahm and Memmel (2010). The authors propose a shrinkage procedure that shrinks the sample-base GMVP weights $\hat{\omega}$ towards the weights of the $1/N$ portfolio, according to the following equation:

$$\hat{\omega}_{\mathrm{FM}} = c\,\omega_{1/N} + (1-c)\hat{\omega}$$

In the previous equation c is the shrinkage parameter and is defined as $c = \min(c_s, 1)$, where $c_s = \frac{N-3}{T-N+2}\frac{1}{\hat{\phi}}$. Hereby, $\hat{\phi} = \frac{\hat{\omega}'_{1/N}\hat{\Sigma}\hat{\omega}_{1/N} - \hat{\omega}'\hat{\Sigma}\hat{\omega}}{\hat{\omega}'\hat{\Sigma}\hat{\omega}}$ and defines the relative portfolio variance difference between the 1/N and the GMV portfolio. Since this portfolio method is considering the GMV portfolio using the sample based covariance matrix estimate $\hat{\Sigma}_X$ it is not computable for the case when the number of assets N is larger than the number of observations T, since for this case $\hat{\Sigma}_X$ will be singular.

11. **Pollak (2011)**

Finally, we use the portfolio-choice approach proposed by Pollak (2011). He considers also a method that shrinks the estimated GMVP weights $\hat{\omega}$ towards the portfolio weights of the $1/N$ approach $\omega_{1/N}$. The combined portfolio weights are constructed according to

$$\hat{\omega}_P = g\left(D\left(\hat{\omega}, \omega_{1/N}\right)\right)\omega_{1/N} + \left[1 - g\left(D\left(\hat{\omega}, \omega_{1/N}\right)\right)\right]\hat{\omega}, \tag{32}$$

where $g(x) = \frac{1}{(1+b\cdot x)}$ and $D\left(\hat{\omega}, \omega_{1/N}\right) = \sqrt{\left(\hat{\omega} - \omega_{1/N}\right)' S \left(\hat{\omega} - \omega_{1/N}\right)}$. The variable b is a positive constant that takes the value 0.5 in the original specification of the article. In our analysis we

let b vary from 0 to 10 and select those value that is maximizing the insample Sharpe Ratio. The statistical distance $D\left(\hat{\omega}, \omega_{1/N}\right)$ has the form of a regular Euclidean distance, only that it is extended by the matrix S. In this case, S is a diagonal matrix that contains the reciprocal of the maximum sample variance of all components of $\hat{\omega}$, that is obtained using a bootstrap procedure, on its main diagonal. S is designed to reduce the distance between the portfolio weights if there is a significant uncertainty in the GMVP estimate and to increase the distance if the estimation uncertainty is small.

5.4 Results of the Forecasting Experiment

The results for our first simulation study are reported in Table 1 and represent the average outcomes for each performance measure across the 1000 different forecasting experiments. If we look at the results we can see that our sparse factor model offers the lowest out-of-sample portfolio standard deviation across all subsamples. At the same time it generates the highest portfolio returns and consequently offers also the best performance in terms of CE and SR compared to all other methods. Furthermore, we can see that the CE results for our method do not depend on the imposed value for the risk aversion factor γ.

Furthermore, we can observe that an increase of the investment dimension causes a high improvement of the out-of-sample portfolio performance of our sparse factor model. Hereby, we can see that increasing N from 30 to 200 yields a reduction of the portfolio standard deviation of about 6% and an increase in the CE of approximately 18%. This result shows that our model is capable to manage the problems in estimation resulting from the cause of dimensionality. In comparison to that, we observe that the CE of the approach by Fama and French (1993) which offers the second lowest standard deviation decreases by 4%. The reason for that lies in the decrease of the average portfolio return which turns out to be higher than the improvement in the standard deviation. In comparison to that, the generated portfolio returns from our method increase with the asset space.

Also in our simulation study we can confirm the statement by DeMiguel et al. (2009b) that the equally weighted portfolio is hardly to beat. For low dimensions ($N = 30$ and $N = 50$) we can see that apart from our estimator only the single factor model generates a higher average SR compared to the $1/N$ portfolio, although it is very close to it. In terms of standard devia-

tion, only our method and 3 factor model by Fama and French (1993) offer throughout lower quantities compared to the $1/N$ portfolio. The picture changes slightly if we move to higher asset dimensions ($N > 50$). We can observe that the method by Abadir et al. (2014) has also advantages in terms of all performance measures compared to the equally weighted portfolio.
It is also interesting to compare our method to the general approximate factor model with a dense factor loadings matrix and to the dynamic factor model. The simulation results show that none of both methods provide better estimation results compared to the $1/N$ portfolio.
Furthermore, Table 2 shows several properties of the estimated portfolio weights for the different

For our main empirical analysis we are considering the period from January 1974 until April 2015, but it is also important to verify during which subperiods of the entire time span our FLasso estimator is improving most in terms of portfolio standard deviation compared to the competing methods. In our analysis we are especially focusing on the periods before and after the recent financial crisis in 2007.
The results are illustrated in Figure 4, where the portfolio standard deviation at time t incorporates the out-of-sample portfolio returns until t (e.g. the out-of-sample portfolio standard deviation in January 1995 incorporates the out-of-sample portfolio returns from January 1979 until January 1995). The graphs indicate that our FLasso estimator offers also for different subperiods the lowest portfolio standard deviation compared to the FF3F and A estimators. Furthermore, the difference is more pronounced if the recent financial crisis period is included. Hence, in comparison to our sparse factor model the FF3F and A estimators are more severely affected by the crisis and as a result they provide more volatile portfolio estimates. Overall, we also observe a slightly decreasing portfolio standard deviation for an increasing sample size N.

6 Conclusions

This paper takes a closer look at a novel variance-covariance matrix estimator based on a sparse factor model that allows for sparsity in the factor loadings matrix, by shrinking single factor loadings to zero. Hence, this setting reduces the amount of parameters that have to be estimated and might lead to a reduction of the estimation noise. Further, the sparse factor model framework allows for weak factors that affect only a subset of the available time series. This framework is therefore more general compared to the standard approximate factor model that allows only for strong factors, affecting a big contingent of available time series. In the theoretical part of the paper, we are able to show average consistency under the frobenius norm for the factor loadings and idiosyncratic error covariance matrix estimators based on our sparse factor model. The factors estimated using the GLS method are as well consistent. Furthermore, we derive several risk bounds for the covariance matrix estimator based on our sparse factor model.

In an empirical horse race we compare the performance of the global minimum variance portfolio based on our sparse factor model to alternative portfolio strategies that are commonly used in the literature. The forecasting results reveal that our sparse factor model offers the lowest average out-of-sample portfolio standard deviation across different portfolio sizes. At the same time it generates the highest certainty equivalent and sharpe ratio compared to all considered portfolio strategies. The performance gains of our sparse factor model are especially pronounced during the recent financial crisis. Hence, it has a stabilizing impact on the portfolio weights, during highly volatile periods.

At this stage, we focus on a static factor model representation and hence do not incorporate any dynamics in the factors. However, our portfolio forecasting experiment shows especially for small samples sizes, a lower portfolio standard deviation for the dynamic factor model compared to the approximate factor model. This leads to the conclusion that there might be efficiency gains, by modelling directly dynamics in the factors. Further, an extension of our sparse factor model to consider dynamic factors might not be very cumbersome and could be obtained using the Kalman smoother estimator. The efficiency of any factor model based estimator is also highly influenced by the amount of included factors in the model. In a strong factor setting this issue is not severe as many consistent estimators for the number of factors are provided

in the literature. However, in many real data applications the strong factor assumption is hardly observed and hence the number of factors might be inconsistently estimated. Therefore, it would interesting to analyse the performance of our sparse factor model if the number of included factors is lower or higher than the true one. Both issues are left for future research.

References

Abadir, K. M., W. Distaso, and F. Žikeš (2014): "Design-free estimation of variance matrices," Journal of Econometrics, 181, 165–180.

Bai, J. and K. Li (2012): "Statistical analysis of factor models of high dimension," The Annals of Statistics, 436–465.

——— (2016): "Maximum likelihood estimation and inference for approximate factor models of high dimension," Review of Economics and Statistics, 98, 298–309.

Bai, J. and Y. Liao (2016): "Efficient estimation of approximate factor models via penalized maximum likelihood," Journal of Econometrics, 191, 1–18.

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

——— (2007): "Determining the number of primitive shocks in factor models," Journal of Business & Economic Statistics, 25.

Bien, J. and R. J. Tibshirani (2011): "Sparse estimation of a covariance matrix," Biometrika, 98, 807.

Chamberlain, G. and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," Econometrica, 51, 1281–304.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009a): "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," Management Science, 55, 798–812.

DeMiguel, V., L. Garlappi, and R. Uppal (2009b): "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" Review of Financial Studies, 22, 1915–1953.

Doz, C., D. Giannone, and L. Reichlin (2011): "A two-step estimator for large approximate dynamic factor models based on Kalman filtering," Journal of Econometrics, 164, 188–205.

Fama, E. F. and K. R. French (1993): "Common risk factors in the returns on stocks and bonds," Journal of Financial Economics, 33, 3–56.

Fan, J., Y. Fan, and J. Lv (2008): "High dimensional covariance matrix estimation using a factor model," Journal of Econometrics, 147, 186–197.

Fan, J., Y. Liao, and M. Mincheva (2011): "High Dimensional Covariance Matrix Estimation in Approximate Factor Models." Annals of Statistics, 39, 3320–3356.

——— (2013): "Large covariance estimation by thresholding principal orthogonal complements," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75, 603–680.

Frahm, G. and C. Memmel (2010): "Dominating estimators for minimum-variance portfolios," Journal of Econometrics, 159, 289–302.

Geweke, J. (1977): "The dynamic factor analysis of economic timeseries models," Latent variables in socio-economic models, 365–383.

James, W. and C. Stein (1961): "Estimation with quadratic loss," in Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, vol. 1, 361–379.

Kourtis, A., G. Dotsis, and R. N. Markellos (2012): "Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix," Journal of Banking & Finance, 36, 2522–2531.

Lawley, D. and A. Maxwell (1971): Factor Analysis as a Statistical Method, second ed., Butterworths, London.

Ledoit, O. and M. Wolf (2003): "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," Journal of Empirical Finance, 10, 603–621.

Lütkepohl, H. (2005): New introduction to multiple time series analysis, Springer Science & Business Media.

Onatski, A. (2010): "Determining the Number of Factors from Empirical Distribution of Eigenvalues," Review of Economics and Statistics, 92, 1004–1016.

——— (2012): "Asymptotics of the principal components estimator of large factor models with weakly influential factors," Journal of Econometrics, 168, 244–258.

Pollak, I. (2011): "Weight shrinkage for portfolio optimization," in Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on, IEEE, 37–40.

Sharpe, W. F. (1963): "A simplified model for portfolio analysis," Management Science, 9, 277–293.

Stock, J. H. and M. W. Watson (2002a): "Forecasting using principal components from a large number of predictors," Journal of the American Statistical Association, 97, 1167–1179.

——— (2002b): "Macroeconomic forecasting using diffusion indexes," Journal of Business & Economic Statistics, 20, 147–162.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

A Appendix

A.1 Proofs

Proof: Theorem 3.1 (Consistency)

Define the penalized likelihood

$$L_p(\Lambda, \Sigma_u) = Q_1(\Lambda, \Sigma_u) + Q_2(\Lambda, \Sigma_u) + Q_3(\Lambda, \Sigma_u), \tag{33}$$

where

$$\begin{aligned}
Q_1(\Lambda, \Sigma_u) &= \frac{1}{N}\log|\Sigma_u| + \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right) - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\text{tr}\left(S_u\Sigma_{u0}^{-1}\right) \\
&\quad + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik}| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}| \\
Q_2(\Lambda, \Sigma_u) &= \frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda_0 - \Lambda_0'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0\right) \\
Q_3(\Lambda, \Sigma_u) &= \frac{1}{N}\log\left|\Lambda\Lambda' + \Sigma_u\right| + \frac{1}{N}\text{tr}\left(S_x\left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right) - Q_2(\Lambda, \Sigma_u) \\
&\quad - \frac{1}{N}\log|\Sigma_u| - \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)
\end{aligned}$$

Therefore, we can see that equation (33) can be written as

$$\begin{aligned}
L_p(\Lambda, \Sigma_u) &= \frac{1}{N}\log\left|\Lambda\Lambda' + \Sigma_u\right| + \frac{1}{N}\text{tr}\left(S_x\left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right) \\
&\quad - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\text{tr}\left(S_u\Sigma_{u0}^{-1}\right) \\
&\quad + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik}| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|
\end{aligned} \tag{34}$$

Define the set,

$$\begin{aligned}
\Psi_\delta = \{(\Lambda, \Sigma_u) : \delta^{-1} &< \pi_{\min}\left(\frac{\Lambda'\Lambda}{N^\beta}\right) \leq \pi_{\max}\left(\frac{\Lambda'\Lambda}{N^\beta}\right) < \delta, \quad \text{for } 0 \leq \beta \leq N \\
\delta^{-1} &< \pi_{\min}\left(\Sigma_u\right) \leq \pi_{\max}\left(\Sigma_u\right) < \delta\}
\end{aligned}$$

We introduce a lemma that will be necessary for the forthcoming derivations

Lemma A.1 (i) $\max_{i,j\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt} - \mathbf{E}\left[u_{it}u_{jt}\right]\right| = \mathcal{O}_p\left(\sqrt{(\log N)/T}\right)$

(ii) $\max_{i\leq r, j\leq N}\left|\frac{1}{T}\sum_{t=1}^{T} f_{it}u_{jt}\right| = \mathcal{O}_p\left(\sqrt{(\log N)/T}\right)$

Proof. See Lemmas A.3 and B.1 in Fan, Liao, and Mincheva (2011). □

Lemma A.2

$$\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta} |Q_3(\Lambda,\Sigma_u)| = \mathcal{O}\left(\frac{1}{N} + \sqrt{\frac{\log N}{T}}\right)$$

Proof. The unpenalized likelihood

$$L(\Lambda,\Sigma_u) = \frac{1}{N}\log\left|\Lambda\Lambda' + \Sigma_u\right| + \frac{1}{N}\mathrm{tr}\left(S_x\left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right), \tag{35}$$

can be decomposed in a similar fashion as in *Lemma A.2.* in Bai and Liao (2016).
The first term in equation (35) can be written as:

$$\frac{1}{N}\log\left|\Lambda\Lambda' + \Sigma_u\right| = \frac{1}{N}\log|\Sigma_u| + \frac{1}{N}\log\left|I_r + \Lambda'\Sigma_u^{-1}\Lambda\right|.$$

Hence, we have

$$\frac{1}{N}\log\left|\Lambda\Lambda' + \Sigma_u\right| = \frac{1}{N}\log|\Sigma_u| + \mathcal{O}\left(\frac{1}{N}\right) \tag{36}$$

Now, we consider the second term $\frac{1}{N}\mathrm{tr}\left(S_x\left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right)$. Hereby, S_x is defined as:

$$S_x = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(x_t - \bar{x})' = \Lambda_0\Lambda_0' + S_u + \Lambda_0\frac{1}{T}\sum_{t=1}^{T} f_t u_t' + \left(\Lambda_0\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\right)' - \bar{u}\bar{u}',$$

where $S_u = \frac{1}{T}\sum_{t=1}^{T} u_t u_t'$ and the identification condition $\frac{1}{T}\sum_{t=1}^{T} f_t f_t' = I_r$ is used.
By the matrix inversion formula we have:

$$\left(\Lambda\Lambda' + \Sigma_u\right)^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}\Lambda\left(I_r + \Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}$$

Hence, we get:

$$\begin{aligned}\frac{1}{N}\text{tr}\left(S_x\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right) = \frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda_0\right) + \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right) \\ - A_1 + A_2 + A_3 - A_4 - A_5,\end{aligned} \tag{37}$$

where $A_1 = \frac{1}{N}\text{tr}\left(\Lambda_0\Lambda_0'\Sigma_u^{-1}\Lambda\left(I_r+\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right)$,
$A_2 = \frac{1}{N}\text{tr}\left(\frac{1}{T}\sum_{t=1}^{T}\Lambda_0 f_t u_t'\left(\Lambda\Lambda+\Sigma_u\right)^{-1}\right)$, $A_3 = \frac{1}{N}\text{tr}\left(\frac{1}{T}\sum_{t=1}^{T}u_t f_t'\Lambda_0'\left(\Lambda\Lambda+\Sigma_u\right)^{-1}\right)$,
$A_4 = \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\Lambda\left(I_r+\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right)$ and $A_5 = \frac{1}{N}\text{tr}\left(\bar{u}\bar{u}'\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right)$.

Subsequently, we look at the terms $A_1 - A_5$, respectively.

Since $\pi_{\max}(\Sigma_u)$ and $\pi_{\min}^{-1}(\Lambda'\Lambda)$ are bounded from above uniformly in Ψ_δ, we have:

$$\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}\pi_{\max}\left[\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\right] \leq \sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}\frac{\pi_{\max}(\Sigma_u)}{\pi_{\min}(\Lambda'\Lambda)} = \mathcal{O}(1) \tag{38}$$

$$\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}\pi_{\max}\left[\left(I_r+\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\right] \leq \sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}\pi_{\max}\left[\left(\Lambda'\Sigma_u^{-1}\Lambda\right)\right] = \mathcal{O}(1) \tag{39}$$

By applying the matrix inversion formula we have,

$$\begin{aligned}A_1 =& \frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0\right) \\ &- \frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\left(I_r+\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0\right),\end{aligned}$$

where the second term can be bounded using (38) and (39), by the following:

$$\begin{aligned}&\frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\left(I_r+\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0\right)) \\ &\quad\leq \frac{1}{N}\left\|\Lambda_0'\Sigma_u^{-1}\Lambda\right\|_F^2\pi_{\max}\left[\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\right]\pi_{\max}\left[\left(I_r+\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\right] \\ &\quad\leq r\left\|\Lambda_0'\Sigma_u^{-1}\Lambda\right\|^2\mathcal{O}\left(\frac{1}{N}\right) = \mathcal{O}\left(\frac{1}{N}\right)\end{aligned}$$

Hence,

$$A_1 = \frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0\right) + \mathcal{O}\left(\frac{1}{N}\right)$$

Using Lemma A.1 and the fact that $\pi_{\max}\left[\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right] \leq \pi_{\max}\left[\left(\Sigma_u\right)^{-1}\right] = \mathcal{O}(1)$, we have:

$$\begin{aligned}
\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta} |A_2| &\leq \frac{1}{N}\left\|\Lambda_0'\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right\|_F \left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\right\|_F \\
&\leq \frac{\sqrt{r}}{N}\left\|\Lambda_0\right\| \left\|\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right\| \sqrt{rN}\left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\right\|_{\max} \\
&\leq \mathcal{O}(1)\,\mathcal{O}_p\left(\frac{r}{\sqrt{N}}\sqrt{\frac{\log N}{T}}\right) = \mathcal{O}_p\left(\sqrt{\frac{\log N}{NT}}\right)
\end{aligned}$$

Similarly, $\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}|A_3| = \mathcal{O}_p\left(\sqrt{\frac{\log N}{NT}}\right)$.

By the matrix inversion formula, we have for A_4 the following:

$$\begin{aligned}
A_4 =& \frac{1}{N}\mathrm{tr}\left(S_u\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right) \\
&- \frac{1}{N}\mathrm{tr}\left(S_u\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\left(I_r+\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right)
\end{aligned}$$

By the equations (38) and (39), we can see that the second term on the right hand side is uniformly of the same order as the first term. The first term of A_4 is bounded by:

$$\begin{aligned}
A_4 &\leq \frac{1}{N}\left\|S_u\Sigma_u^{-1}\Lambda\right\|_F \left\|\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\right\| \left\|\Lambda'\Sigma_u^{-1}\right\| \\
&\leq \frac{\sqrt{r}}{N}\left\|S_u\right\|\mathcal{O}(1) \\
&\leq \mathcal{O}\left(N^{-1}\right)\pi_{\max}\left(S_u\right) = \mathcal{O}\left(\sqrt{\frac{\log N}{T}}+\frac{1}{N}\right)
\end{aligned}$$

Finally, $A_5 = \mathcal{O}_p\left(\frac{\log N}{T}\right)$ by a similar argument as in Bai and Liao (2016).

Hence, the unpenalized likelihood can be bounded by:

$$\begin{aligned}
L(\Lambda,\Sigma_u) &= \frac{1}{N}\mathrm{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda_0\right) + \frac{1}{N}\mathrm{tr}\left(S_u\Sigma_u^{-1}\right) + \frac{1}{N}\log\left|\Sigma_u\right| \\
&\quad - \frac{1}{N}\mathrm{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0\right) + \mathcal{O}\left(\frac{1}{N}+\sqrt{\frac{\log N}{T}}\right) \\
&= \frac{1}{N}\log\left|\Sigma_u\right| + \frac{1}{N}\mathrm{tr}\left(S_u\Sigma_u^{-1}\right) + Q_2(\Lambda,\Sigma_u) + \mathcal{O}\left(\frac{1}{N}+\sqrt{\frac{\log N}{T}}\right)
\end{aligned}$$

By the definition of $Q_3(\Lambda, \Sigma_u)$ we have

$$\sup |Q_3(\Lambda, \Sigma_u)| = \mathcal{O}\left(\frac{1}{N} + \sqrt{\frac{\log N}{T}}\right)$$

□

Lemma A.3
For $d_T = \left(\frac{1}{N} + \sqrt{\frac{\log N}{T}}\right)$

$$Q_1(\hat{\Lambda}, \hat{\Sigma}_u) + Q_2(\hat{\Lambda}, \hat{\Sigma}_u) = \mathcal{O}_p\left(d_T\right)$$

Proof. If we consider equation (34) at the true parameter values, we get

$$\begin{aligned} L_p(\Lambda_0, \Sigma_{u0}) &= \frac{1}{N}\log\left|\Lambda_0\Lambda_0' + \Sigma_{u0}\right| + \frac{1}{N}\mathrm{tr}\left(S_x\left(\Lambda_0\Lambda_0' + \Sigma_{u0}\right)^{-1}\right) \\ &\quad - Q_2(\Lambda_0, \Sigma_{u0}) - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\mathrm{tr}\left(S_u\Sigma_{u0}^{-1}\right) \\ &\quad + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}| \\ &= Q_3(\Lambda_0, \Sigma_{u0}) \end{aligned} \tag{40}$$

Hence, by (33) and (40), we have

$$\begin{aligned} Q_1(\hat{\Lambda}, \hat{\Sigma}_u) + Q_2(\hat{\Lambda}, \hat{\Sigma}_u) &= L_c(\hat{\Lambda}, \hat{\Sigma}_u) - Q_3(\hat{\Lambda}, \hat{\Sigma}_u) \\ &\leq L_c(\Lambda_0, \Sigma_{u0}) - Q_3(\hat{\Lambda}, \hat{\Sigma}_u) \\ &= Q_3(\Lambda_0, \Sigma_{u0}) - Q_3(\hat{\Lambda}, \hat{\Sigma}_u) \\ &= 2\sup|Q_3(\Lambda, \Sigma_u)| \end{aligned}$$

Therefore, by Lemma A.2 we have

$$Q_1(\hat{\Lambda}, \hat{\Sigma}_u) + Q_2(\hat{\Lambda}, \hat{\Sigma}_u) \leq d_T, \tag{41}$$

□

Lemma A.4

$$\frac{1}{N}\left\|\hat{\Sigma}_u - \Sigma_{u0}\right\|_F^2 = \mathcal{O}_p\left(\frac{\log N}{T} + d_T\right) = o_p(1)$$

Proof. By equation (41) and the definition of $Q_1(\hat{\Lambda}, \hat{\Sigma}_u)$ and $Q_2(\hat{\Lambda}, \hat{\Sigma}_u)$, we get

$$B_1 + B_2 \leq d_T, \tag{42}$$

where B_1 and B_2 are defined as

$$\begin{aligned}
B_1 &= \frac{1}{N}\log\left|\hat{\Sigma}_u\right| + \frac{1}{N}\mathrm{tr}\left(S_u\hat{\Sigma}_u^{-1}\right) - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\mathrm{tr}\left(S_u\Sigma_{u0}^{-1}\right) \\
B_2 &= \frac{1}{N}\mathrm{tr}\left(\Lambda_0'\hat{\Sigma}_u^{-1}\Lambda_0 - \Lambda_0'\hat{\Sigma}_u^{-1}\hat{\Lambda}\left(\hat{\Lambda}'\hat{\Sigma}_u^{-1}\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Sigma}_u^{-1}\Lambda_0\right) \\
&\quad + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}\right| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|
\end{aligned}$$

In the following we will first analyse the term B_2 which can be further decomposed into

$$\begin{aligned}
&\frac{1}{N}\mathrm{tr}\left[\left(\hat{\Lambda} - \Lambda_0\right)'\hat{\Sigma}_u^{-1}\left(\hat{\Lambda} - \Lambda_0\right)\right] - \underbrace{\mathrm{tr}\left[\frac{1}{N}JH^{-1}J'\right]}_{B_3} \\
&\quad + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}\right| - |\lambda_{ik0}|,
\end{aligned} \tag{43}$$

where $J = \left(\hat{\Lambda} - \Lambda_0\right)'\hat{\Sigma}_u^{-1}\hat{\Lambda}\left(\hat{\Lambda}'\hat{\Sigma}_u^{-1}\hat{\Lambda}\right)^{-1}$ and $H^{-1} = \hat{\Lambda}'\hat{\Sigma}_u^{-1}\hat{\Lambda}$. Subsequently, we are looking at the term B_3.

$$\begin{aligned}
B_3 &\leq \frac{1}{N}\mathrm{tr}\left[JJ'\right] \\
&= \frac{1}{N}\mathrm{tr}\left[\left(\hat{\Lambda} - \Lambda_0\right)'\hat{\Sigma}_u^{-1}\hat{\Lambda}\left(\hat{\Lambda}'\hat{\Sigma}_u^{-1}\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Sigma}_u^{-1}\left(\hat{\Lambda} - \Lambda_0\right)\right] \\
&\leq \frac{1}{N}\left\|\hat{\Lambda} - \Lambda_0\right\|_F^2\left\|\hat{\Lambda}'\hat{\Sigma}_u\right\|^2\left\|\left(\hat{\Lambda}'\hat{\Sigma}_u^{-1}\hat{\Lambda}\right)^{-1}\right\| = \mathcal{O}_p\left(\frac{1}{N}\right)\left\|\hat{\Lambda} - \Lambda_0\right\|_F^2
\end{aligned} \tag{44}$$

Hence, by equation (44), we can see that B_3 is of the same order as the first term in equation (43). In the next step we want to take a closer look to the third term in (43), which can be

similarly expressed as

$$\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}\right|-|\lambda_{ik0}|=-\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|-\left|\hat{\lambda}_{ik}\right|$$

Hence, an upper bound on the expression $\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|-\left|\hat{\lambda}_{ik}\right|$ can be obtained by

$$\begin{aligned}\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|-\left|\hat{\lambda}_{ik}\right| &\leq \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}-\lambda_{ik0}\right| \\ &\leq \frac{1}{N}\mu\sqrt{D_N}\left\|\hat{\Lambda}-\Lambda_0\right\|_F\end{aligned}$$

Hence, this result shows the positivity of term B_2 and by equation (42), we can see that

$$\frac{1}{N}\log\left|\hat{\Sigma}_u\right|+\frac{1}{N}\mathrm{tr}\left(S_u\hat{\Sigma}_u^{-1}\right)-\frac{1}{N}\log|\Sigma_{u0}|-\frac{1}{N}\mathrm{tr}\left(S_u\Sigma_{u0}^{-1}\right)\leq d_T \tag{45}$$

Using the same argument as in the proof of *Lemma B.1.* in Bai and Liao (2016), we get

$$\begin{aligned}c\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F^2-\mathcal{O}_p\left(\sqrt{\frac{\log N}{T}}\right)\sum_{ij}\left|\Sigma_{u0,ij}-\hat{\Sigma}_{u,ij}\right| &\leq Nd_T \\ c\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F^2-\mathcal{O}_p\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N}\left\|\Sigma_{u0}-\hat{\Sigma}_u\right\|_F &\leq Nd_T \\ c\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F^2-\mathcal{O}_p\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N}\left\|\hat{\Sigma}_u\left(\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right)\Sigma_{u0}\right\|_F &\leq Nd_T \\ c\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F^2-\mathcal{O}_p\left(\sqrt{\frac{\log N}{T}}\right)\sqrt{N}\left\|\hat{\Sigma}_u\right\|\|\Sigma_{u0}\|\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F &\leq Nd_T\end{aligned}$$

Solving for $\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F$ yields

$$\begin{aligned}\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F &= \mathcal{O}_p\left(\sqrt{\frac{N\log N}{T}}+\sqrt{Nd_T}\right) \\ \frac{1}{N}\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F^2 &= \mathcal{O}_p\left(\frac{\log N}{T}+d_T\right)=o_p(1)\end{aligned}$$

Hence, we can conclude the proof by the following derivation:

$$\begin{aligned}\frac{1}{N}\left\|\hat{\Sigma}_u^{-1}-\Sigma_{u0}^{-1}\right\|_F^2 &= \frac{1}{N}\left\|\hat{\Sigma}_u^{-1}\left(\Sigma_{u0}-\hat{\Sigma}_u\right)\Sigma_{u0}^{-1}\right\|_F^2 \\ &\leq \frac{1}{N}\left\|\hat{\Sigma}_u^{-1}\right\|^2\left\|\Sigma_{u0}^{-1}\right\|^2\left\|\Sigma_{u0}-\hat{\Sigma}_u\right\|_F^2\end{aligned}$$

□

Lemma A.5

$$\frac{1}{N}\left\|\hat{\Lambda}-\Lambda_0\right\|_F^2 = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N}+d_T\right)$$

Proof. If we consider equation (42) and Lemma A.4, we have

$$\frac{1}{N}\mathrm{tr}\left[\Lambda_0'\hat{\Sigma}_u^{-1}\Lambda_0-\Lambda_0'\hat{\Sigma}_u^{-1}\hat{\Lambda}\left(\hat{\Lambda}'\hat{\Sigma}_u^{-1}\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Sigma}_u^{-1}\Lambda_0\right]+\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}\right|-|\lambda_{ik0}|\leq d_T,$$

Hence, the result follows by equations (43) and (44)

$$\begin{aligned}\frac{1}{N}\mathrm{tr}\left[\left(\hat{\Lambda}-\Lambda_0\right)'\hat{\Sigma}_u^{-1}\left(\hat{\Lambda}-\Lambda_0\right)\right]+\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}\right|-|\lambda_{ik0}| &\leq d_T \\ \frac{1}{N}\left\|\hat{\Lambda}-\Lambda_0\right\|_F^2-\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|-\left|\hat{\lambda}_{ik}\right| &\leq d_T \\ \frac{1}{N}\left\|\hat{\Lambda}-\Lambda_0\right\|_F^2-\frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}-\lambda_{ik0}\right| &\leq d_T \\ \frac{1}{N}\left\|\hat{\Lambda}-\Lambda_0\right\|_F^2-\frac{1}{N}\mu\sqrt{D_N}\left\|\hat{\Lambda}-\Lambda_0\right\|_F &\leq d_T\end{aligned}$$

Solving for $\left\|\hat{\Lambda}-\Lambda_0\right\|_F$ yields

$$\begin{aligned}\left\|\hat{\Lambda}-\Lambda_0\right\|_F &\leq \mu\sqrt{D_N}+\sqrt{\mu^2 D_N+4Nd_T} \\ &\leq \mu\sqrt{D_N}+\sqrt{Nd_T}\end{aligned}$$

□

Proof: Theorem 3.2 (Risk Bounds)

$$\Sigma = \Lambda_0 \Lambda_0' + \Sigma_{u0} \tag{46}$$

$$\hat{\Sigma}_{SF} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Sigma}_u^{\tau}, \tag{47}$$

where $\hat{\Sigma}_u^{\tau}$ corresponds to the POET estimator of Fan et al. (2013). Similar as in Fan et al. (2013), we consider the weighted quadratic norm introduced by Fan, Fan, and Lv (2008) and which is defined as:

$$\|A\|_{\Sigma} = N^{-1/2} \left\| \Sigma^{-1/2} A \Sigma^{-1/2} \right\|_F$$

Lemma A.6

$$\frac{1}{N} \left\| \hat{\Sigma}_{SF} - \Sigma \right\|_{\Sigma}^2 = \mathcal{O}_p \left(\frac{\mu^2 D_N}{N^2} + \frac{d_T}{N} + \left(\frac{1}{N} + \frac{\log N}{T} \right) \frac{m_N^2}{N^2} \right)$$

Proof. The weighted quadratic norm of the difference between the estimated covariance matrix $\hat{\Sigma}_{SF}$ and the true one Σ can be expressed as:

$$\left\| \hat{\Sigma}_{SF} - \Sigma \right\|_{\Sigma}^2 \leq \left\| \hat{\Lambda}\hat{\Lambda}' - \Lambda_0 \Lambda_0' \right\|_{\Sigma}^2 + \left\| \hat{\Sigma}_u^{\tau} - \Sigma_{u0} \right\|_{\Sigma}^2 \tag{48}$$

If we consider $C = \hat{\Lambda} - \Lambda_0$ we can introduce the following definitions:

$$C\hat{\Lambda}' = \hat{\Lambda}\hat{\Lambda}' - \Lambda_0 \hat{\Lambda}'$$

$$\Lambda_0 C' = \Lambda_0 \hat{\Lambda}' - \Lambda_0 \Lambda_0'$$

Using the previous definitions, we can rewrite the first term in (48) in the following form

$$\begin{aligned} \left\| \hat{\Lambda}\hat{\Lambda}' - \Lambda_0 \Lambda_0' \right\|_{\Sigma}^2 &= \left\| C\hat{\Lambda}' + \Lambda_0 C' \right\|_{\Sigma}^2 \\ &\leq \left\| C\hat{\Lambda}' \right\|_{\Sigma}^2 + \left\| \Lambda_0 C' \right\|_{\Sigma}^2 \end{aligned}$$

Hence, equation (48) can be expressed as:

$$\left\|\hat{\Sigma}_{SF}-\Sigma\right\|_{\Sigma}^{2} \leq \left\|C\hat{\Lambda}'\right\|_{\Sigma}^{2} + \left\|\Lambda_0 C'\right\|_{\Sigma}^{2} + \left\|\hat{\Sigma}_u^{\tau}-\Sigma_u\right\|_{\Sigma}^{2} \tag{49}$$

Now we analyse each term in (49) separately:

$$\begin{aligned}
\left\|C\hat{\Lambda}'\right\|_{\Sigma}^{2} &= N^{-1}\mathrm{tr}\left(\Sigma^{-1/2}C\hat{\Lambda}'\Sigma^{-1/2}\Sigma^{-1/2}\hat{\Lambda}C'\Sigma^{-1/2}\right) \\
&= N^{-1}\mathrm{tr}\left(C'\Sigma^{-1}C\hat{\Lambda}'\Sigma^{-1}\hat{\Lambda}\right) \\
&\leq N^{-1}\left\|\hat{\Lambda}'\Sigma^{-1}\hat{\Lambda}\right\|\left\|\Sigma^{-1}\right\|\left\|C\right\|_F^2 = \mathcal{O}_p\left(N^{-1}\left\|C\right\|_F^2\right)
\end{aligned}$$

Similarly, we get $\left\|\Lambda_0 C'\right\|_{\Sigma}^{2} == \mathcal{O}_p\left(N^{-1}\left\|C\right\|_F^2\right)$. Hence, by Lemma C.12. in Fan et al. (2013) we get:

$$\begin{aligned}
\left\|\hat{\Sigma}_{SF}-\Sigma\right\|_{\Sigma}^{2} &\leq \mathcal{O}_p\left(N^{-1}\left\|C\right\|_F^2\right) + \mathcal{O}_p\left(\left\|\hat{\Sigma}_u^{\tau}-\Sigma_u\right\|_{\Sigma}^{2}\right) \\
&= \mathcal{O}_p\left(N^{-1}\left[\mu^2 D_N + N\left(\frac{1}{N}+\sqrt{\frac{\log N}{T}}\right)\right]\right) + \mathcal{O}_p\left(\left\|\hat{\Sigma}_u^{\tau}-\Sigma_u\right\|_{\Sigma}^{2}\right) \\
&= \mathcal{O}_p\left(\frac{\mu^2 D_N}{N}+\frac{1}{N}+\sqrt{\frac{\log N}{T}}+\left(\frac{1}{N}+\frac{\log N}{T}\right)\frac{m_N^2}{N}\right)
\end{aligned}$$

□

Under the **Frobenius norm** we have:

Lemma A.7

$$\frac{1}{N}\left\|\hat{\Sigma}_{SF}-\Sigma\right\|_{F}^{2} = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N}+d_T+\left(\frac{1}{N}+\frac{\log N}{T}\right)m_N^2\right)$$

Proof.

$$\left\|\hat{\Sigma}_{SF}-\Sigma\right\|_{F}^{2} \leq \left\|C\hat{\Lambda}'\right\|_{F}^{2} + \left\|\Lambda_0 C'\right\|_{F}^{2} + \left\|\hat{\Sigma}_u^{\tau}-\Sigma_u\right\|_{F}^{2}, \tag{50}$$

where the second term can be bounded by

$$\begin{aligned}\left\|\Lambda_0 C'\right\|_F^2 &= \operatorname{tr}\left(\Lambda_0'\Lambda_0 C'C\right) \\ &\leq \|\Lambda_0\|^2 \|C\|_F^2 = \mathcal{O}_p\left(\|C\|_F^2\right)\end{aligned}$$

Furthermore, the first term in (50) has the same upper bound. Hence, again by using Lemma C.12. in Fan et al. (2013) we get:

$$\begin{aligned}\left\|\hat{\Sigma}_{SF} - \Sigma\right\|_F^2 &\leq \mathcal{O}_p\left(\|C\|_F^2\right) + \mathcal{O}_p\left(\left\|\hat{\Sigma}_u^\tau - \Sigma_u\right\|_F^2\right) \\ &\leq \mathcal{O}_p\left(\mu^2 D_N + N d_T\right) + \mathcal{O}_p\left(N\left(\frac{1}{N} + \frac{\log N}{T}\right) m_N^2\right)\end{aligned}$$

□

Inverse of the covariance matrix

Define,

$$\begin{aligned}\hat{G} &= \left(I_r + \hat{\Lambda}'\left(\hat{\Sigma}_u^\tau\right)^{-1}\hat{\Lambda}\right)^{-1} \\ G_0 &= \left(I_r + \Lambda_0'\Sigma_{u0}^{-1}\Lambda_0\right)^{-1}\end{aligned}$$

Lemma A.8 (i) $\left\|\hat{G}\right\| = \mathcal{O}_p(1)$

(ii) $\left\|\hat{G}^{-1} - G_0^{-1}\right\| = \mathcal{O}_p\left(\|C\| + \left\|\left(\hat{\Sigma}_u^\tau\right)^{-1} - \Sigma_u^{-1}\right\|\right)$

Proof.

(i) Theorem 3.1 in Fan et al. (2013) implies $\left\|\left(\hat{\Sigma}_u^\tau\right)^{-1}\right\| = \mathcal{O}_p(1)$. Then, by the definition of $\hat{G}$ we have $\left\|\hat{G}\right\| \leq \left(\mathcal{O}_p(1)\left\|\hat{\Lambda}\right\|^2\right)^{-1} = \mathcal{O}_p(1)$.

(ii) By the definition of $\hat{G}$ and G_0, we have: $\hat{G}^{-1} - G_0^{-1} = \hat{\Lambda}'\left(\hat{\Sigma}_u^\tau\right)^{-1}\hat{\Lambda} - \Lambda_0'\Sigma_{u0}^{-1}\Lambda_0$. Hence, the previous quantitiy can be decomposed according to:

$$\hat{G}^{-1} - G_0^{-1} = C'\left(\hat{\Sigma}_u^\tau\right)^{-1}\hat{\Lambda} + \Lambda_0'\Sigma_{u0}^{-1}C + \Lambda_0'\left(\left(\hat{\Sigma}_u^\tau\right)^{-1} - \Sigma_{u0}^{-1}\right)\hat{\Lambda} \qquad (51)$$

If we bound all three terms on the right hand side of equation (51), we get:

$$\begin{aligned}\left\|\hat{G}^{-1} - G_0^{-1}\right\| &\leq \|C\|\,\mathcal{O}_p(1) + \left\|\left(\hat{\Sigma}_u^{\tau}\right)^{-1} - \Sigma_{u0}^{-1}\right\|\mathcal{O}_p(1) \\ &= \mathcal{O}_p\left(\mu\sqrt{D_N} + \sqrt{N d_T} + \left(\frac{1}{\sqrt{N}} + \sqrt{\frac{\log N}{T}}\right) m_N\right)\end{aligned}$$

□

Lemma A.9

$$\frac{1}{N}\left\|\hat{\Sigma}_{SF}^{-1} - \Sigma^{-1}\right\|^2 = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N} + d_T + \left(\frac{1}{N} + \frac{\log N}{T}\right)\frac{m_N^2}{N}\right)$$

Proof. Using the Sherman-Morrison-Woodbury inverse formula, we get

$$\left\|\hat{\Sigma}_{SF}^{-1} - \Sigma^{-1}\right\|^2 = \sum_{i=1}^{6} L_i,$$

where

$$\begin{aligned}
L_1 &= \left\|\left(\hat{\Sigma}_u^{\tau}\right)^{-1} - \Sigma_{u0}^{-1}\right\|^2 \\
L_2 &= \left\|\left[\left(\hat{\Sigma}_u^{\tau}\right)^{-1} - \Sigma_{u0}^{-1}\right]\hat{\Lambda}\left[I_r + \hat{\Lambda}'\left(\hat{\Sigma}_u^{\tau}\right)^{-1}\hat{\Lambda}\right]^{-1}\hat{\Lambda}'\left(\hat{\Sigma}_u^{\tau}\right)^{-1}\right\|^2 \\
L_3 &= \left\|\left[\left(\hat{\Sigma}_u^{\tau}\right)^{-1} - \Sigma_{u0}^{-1}\right]\hat{\Lambda}\left[I_r + \hat{\Lambda}'\left(\hat{\Sigma}_u^{\tau}\right)^{-1}\hat{\Lambda}\right]^{-1}\hat{\Lambda}'\Sigma_{u0}^{-1}\right\|^2 \\
L_4 &= \left\|\Sigma_{u0}^{-1}\left(\hat{\Lambda} - \Lambda_0\right)\left[I_r + \hat{\Lambda}'\left(\hat{\Sigma}_u^{\tau}\right)^{-1}\hat{\Lambda}\right]^{-1}\hat{\Lambda}'\Sigma_{u0}^{-1}\right\|^2 \\
L_5 &= \left\|\Sigma_{u0}^{-1}\left(\hat{\Lambda} - \Lambda_0\right)\left[I_r + \hat{\Lambda}'\left(\hat{\Sigma}_u^{\tau}\right)^{-1}\hat{\Lambda}\right]^{-1}\Lambda_0'\Sigma_{u0}^{-1}\right\|^2 \\
L_6 &= \left\|\Sigma_{u0}^{-1}\Lambda_0\left(\left[I_r + \hat{\Lambda}'\left(\hat{\Sigma}_u^{\tau}\right)^{-1}\hat{\Lambda}\right]^{-1} - \left[I_r + \Lambda_0'\Sigma_u^{-1}\Lambda_0\right]^{-1}\right)\Lambda_0'\Sigma_{u0}^{-1}\right\|^2
\end{aligned}$$

In the following, we bound each of the six terms, separately.

$$L_2 \leq \left\| \left(\hat{\Sigma}_u^{\tau}\right)^{-1} - \Sigma_{u0}^{-1} \right\|^2 \left\| \hat{\Lambda}\hat{G}\hat{\Lambda}' \right\|^2 \left\| \left(\hat{\Sigma}_u^{\tau}\right)^{-1} \right\|^2$$

By Lemma A.8 follows that $L_2 \leq \mathcal{O}_p(L_1)$. Similarly, L_3 is also $\mathcal{O}_p(L_1)$.

Further,

$$L_4 \leq \left\|\Sigma_{u0}^{-1}\right\|^2 \|C\|^2 \left\|\hat{G}\right\|^2 \left\|\hat{\Lambda}'\Sigma_{u0}^{-1}\right\|^2$$

Hence, also by Lemma A.8

$$L_4 \leq \|C\|^2 \mathcal{O}_p(1) \mathcal{O}_p(1) = \mathcal{O}_p\left(\|C\|^2\right)$$

Similarly, $L_5 = \mathcal{O}_p(L_4)$. Finally,

$$L_6 \leq \left\|\Sigma_{u0}^{-1}\Lambda_0\right\|^4 \left\|\hat{G} - G_0\right\|^2$$

By Lemma A.8 we have,

$$\begin{aligned} L_6 &\leq \mathcal{O}_p(1) \left\|\hat{G}^{-1} - G_0^{-1}\right\|^2 \\ &= \mathcal{O}_p\left(\mu\sqrt{D_N} + \sqrt{N d_T} + \left(\frac{1}{\sqrt{N}} + \sqrt{\frac{\log N}{T}}\right) m_N\right) \end{aligned}$$

Adding up the terms $L_1 - L_6$ gives

$$\frac{1}{N}\left\|\hat{\Sigma}_{SF}^{-1} - \Sigma^{-1}\right\|^2 = \mathcal{O}_p\left(\frac{\mu^2 D_N}{N} + d_T + \left(\frac{1}{N} + \frac{\log N}{T}\right)\frac{m_N^2}{N}\right)$$

□

A.2 Tables

Table 1: Estimation results for the Portfolio Application

Model	1/N	GMVP	FLasso	AFM	DFM	SFM	FF3F	LW	K	A	FM	P
						N = 30						
AV	0.0084	0.0075	**0.0086**	0.0074	0.0075	0.0085	0.0081	0.0077	0.0081	0.0079	0.0079	0.0081
CE	0.0062	0.0031	**0.0065**	0.0049	0.0052	0.0062	0.0059	0.0053	0.0058	0.0056	0.0054	0.0056
SR	0.1766	0.1123	**0.1868**	0.1475	0.1553	0.1768	0.1727	0.1581	0.1678	0.1658	0.1586	0.1619
SD	0.0477	0.0665	**0.0458**	0.0498	0.0484	0.0478	0.0469	0.0489	0.0483	0.0474	0.0500	0.0498
						N = 50						
AV	0.0084	0.0082	**0.0088**	0.0075	0.0075	0.0084	0.0079	0.0080	0.0081	0.0081	0.0080	0.0084
CE	0.0062	-0.0055	**0.0068**	0.0050	0.0052	0.0062	0.0058	0.0057	0.0058	0.0059	0.0029	0.0016
SR	0.1797	0.0706	**0.1973**	0.1510	0.1564	0.1799	0.1736	0.1651	0.1687	0.1717	0.1127	0.1026
SD	0.0467	0.1165	**0.0446**	0.0496	0.0479	0.0468	0.0458	0.0486	0.0479	0.0470	0.0712	0.0817
						N = 100						
AV	0.0084	-	**0.0092**	0.0078	0.0070	0.0084	0.0078	0.0082	0.0081	0.0081	-	-
CE	0.0063	-	**0.0073**	0.0055	0.0048	0.0063	0.0058	0.0060	0.0059	0.0061	-	-
SR	0.1825	-	**0.2112**	0.1627	0.1500	0.1827	0.1740	0.1753	0.1740	0.1785	-	-
SD	0.0462	-	**0.0435**	0.0479	0.0468	0.0463	0.0448	0.0466	0.0463	0.0456	-	-
						N = 150						
AV	0.0084	-	**0.0094**	0.0082	0.0066	0.0084	0.0077	0.0082	0.0080	0.0083	-	-
CE	0.0062	-	**0.0075**	0.0060	0.0045	0.0063	0.0057	0.0062	0.0060	0.0063	-	-
SR	0.1817	-	**0.2182**	0.1747	0.1438	0.1819	0.1730	0.1811	0.1774	0.1864	-	-
SD	0.0460	-	**0.0429**	0.0466	0.0459	0.0460	0.0444	0.0453	0.0452	0.0445	-	-
						N = 200						
AV	0.0084	-	**0.0095**	0.0086	0.0063	0.0084	0.0076	0.0083	0.0078	0.0083	-	-
CE	0.0063	-	**0.0077**	0.0065	0.0043	0.0063	0.0057	0.0063	0.0058	0.0063	-	-
SR	0.1822	-	**0.2238**	0.1883	0.1394	0.1824	0.1728	0.1864	0.1732	0.1883	-	-
SD	0.0459	-	**0.0426**	0.0459	0.0454	0.0459	0.0440	0.0444	0.0450	0.0440	-	-

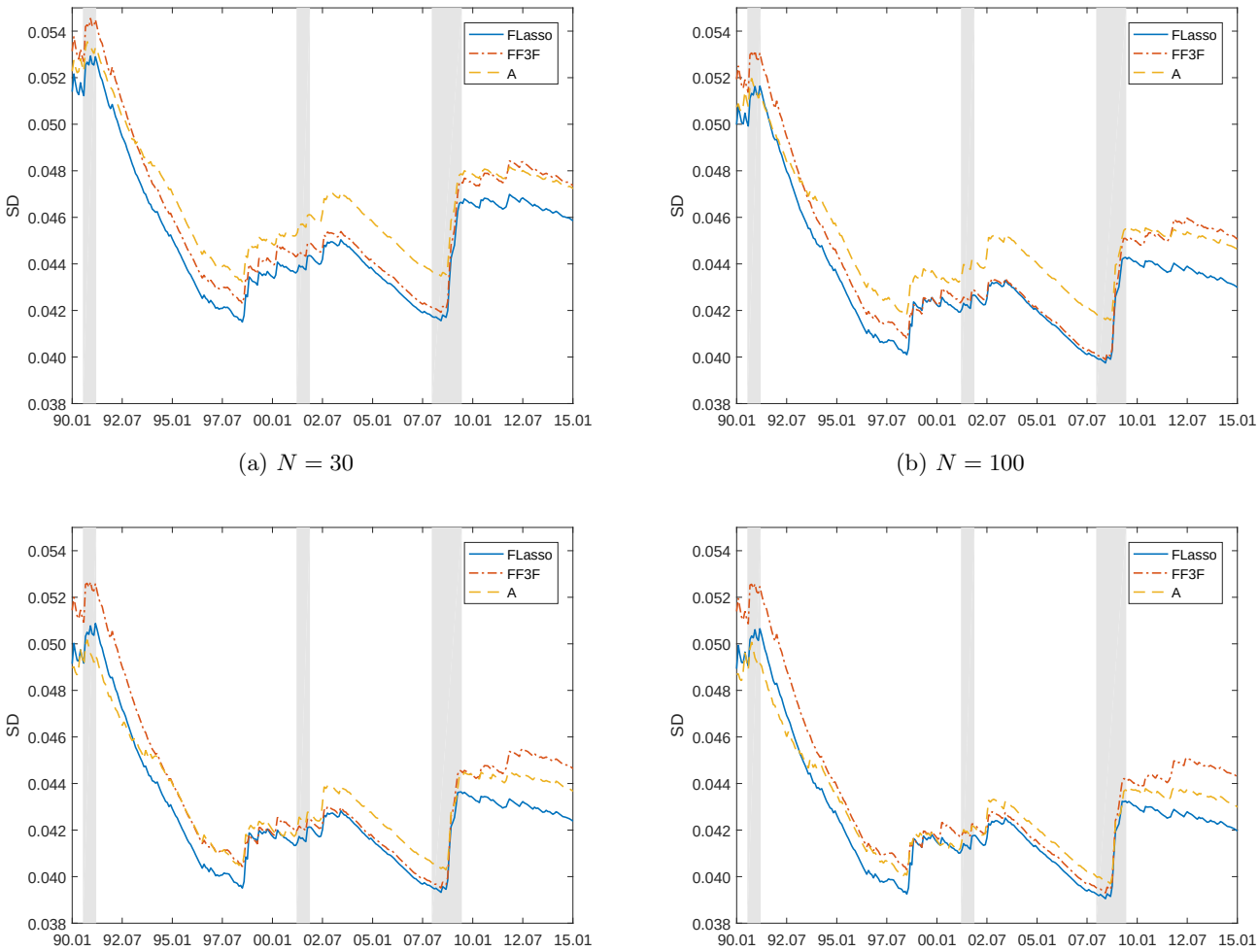
Note: Our sparse approximate factor model (FLasso) is compared to the equally weighted portfolio ($1/N$), the GMVP, the approximate factor model (AFM), the dynamic factor model (DFM), the single factor model by Sharpe (1963) (SFM), the 3 factor model by Fama and French (1993), the estimators by Ledoit and Wolf (2003) (LW), Kourtis et al. (2012) (K), Abadir et al. (2014) (A), Frahm and Memmel (2010) (FM) and Pollak (2011) (P).

Table 2: Estimated weights for the Portfolio Application

Model	1/N	GMVP	FLasso	AFM	DFM	SFM	FF3F	LW	K	A	FM	P
						N = 30						
Min	0.0333	-1.0988	-0.1907	-0.5123	-0.4828	0.0295	-0.0669	-0.5176	-0.6166	-0.1776	-0.7379	-0.7162
Max	0.0333	1.0253	0.1996	0.4584	0.3353	0.0448	0.1121	0.3198	0.5534	0.3392	0.6503	0.6694
SD	0.0000	0.1366	0.0197	0.0670	0.0543	0.0010	0.0088	0.0648	0.0249	0.0500	0.0502	0.0597
MAD	0.0000	0.1075	0.0157	0.0522	0.0421	0.0007	0.0069	0.0506	0.0196	0.0397	0.0395	0.0469
						N = 50						
Min	0.0200	-2.7950	-0.1461	-0.4305	-0.4127	0.0176	-0.0701	-0.4237	-1.0228	-0.1492	-1.8535	-2.3299
Max	0.0200	2.7290	0.1634	0.3798	0.2158	0.0270	0.0884	0.2629	1.0182	0.3004	2.2202	2.3992
SD	0.0000	0.2255	0.0158	0.0505	0.0394	0.0006	0.0073	0.0513	0.0214	0.0400	0.1016	0.1405
MAD	0.0000	0.1777	0.0126	0.0394	0.0307	0.0004	0.0058	0.0399	0.0169	0.0318	0.0800	0.1106
						N = 100						
Min	0.0100	-	-0.1626	-0.3204	-0.2557	0.0089	-0.0558	-0.3158	-0.2538	-0.1020	-	-
Max	0.0100	-	0.1146	0.2179	0.1806	0.0134	0.0612	0.1886	0.2371	0.2450	-	-
SD	0.0000	-	0.0105	0.0315	0.0238	0.0003	0.0052	0.0339	0.0171	0.0269	-	-
MAD	0.0000	-	0.0083	0.0246	0.0185	0.0002	0.0041	0.0263	0.0136	0.0213	-	-
						N = 150						
Min	0.0067	-	-0.1401	-0.1778	-0.2016	0.0060	-0.0431	-0.2203	-0.1768	-0.0752	-	-
Max	0.0067	-	0.0920	0.1482	0.1753	0.0089	0.0435	0.1320	0.2003	0.1696	-	-
SD	0.0000	-	0.0078	0.0231	0.0171	0.0002	0.0040	0.0255	0.0139	0.0204	-	-
MAD	0.0000	-	0.0063	0.0180	0.0133	0.0001	0.0032	0.0198	0.0111	0.0161	-	-
						N = 200						
Min	0.0050	-	-0.1084	-0.1430	-0.1733	0.0045	-0.0325	-0.1561	-0.1053	-0.0562	-	-
Max	0.0050	-	0.0603	0.1048	0.1169	0.0067	0.0349	0.1027	0.1136	0.1372	-	-
SD	0.0000	-	0.0062	0.0183	0.0132	0.0002	0.0033	0.0203	0.0115	0.0164	-	-
MAD	0.0000	-	0.0050	0.0142	0.0103	0.0001	0.0026	0.0157	0.0091	0.0129	-	-

Note: Our sparse approximate factor model (FLasso) is compared to the equally weighted portfolio ($1/N$), the GMVP, the approximate factor model (AFM), the dynamic factor model (DFM), the single factor model by Sharpe (1963) (SFM), the 3 factor model by Fama and French (1993), the estimators by Ledoit and Wolf (2003) (LW), Kourtis et al. (2012) (K), Abadir et al. (2014) (A), Frahm and Memmel (2010) (FM) and Pollak (2011) (P).

A.3 Figures



(a) $N = 30$ (b) $N = 100$

(c) $N = 150$ (d) $N = 200$

Figure 4: SD for different subperiods