Sparse Approximate Factor Estimation for High-Dimensional Covariance Matrices*

Maurizio Daniele[a] | Winfried Pohlmeier[b] | Aygul Zagidullina[c]

University of Konstanz | University of Konstanz | University of Konstanz

GSDS | CoFE, RCEA | QEF

this version: June 12, 2018

Abstract

We propose a novel estimation approach for the covariance matrix based on the l_1-regularized approximate factor model. Our sparse approximate factor (SAF) covariance estimator allows for the existence of weak factors and hence relaxes the pervasiveness assumption generally adopted for the standard approximate factor model. We prove consistency of the covariance matrix estimator under the Frobenius norm as well as the consistency of the factor loadings and the factors.

Our Monte Carlo simulations reveal that the SAF covariance estimator has superior properties in finite samples for low and high dimensions and different designs of the covariance matrix. Moreover, in an out-of-sample portfolio forecasting application the estimator uniformly outperforms alternative portfolio strategies based on alternative covariance estimation approaches and modeling strategies including the $1/N$-strategy.

Keywords: Approximate Factor model, weak factors, l_1-regularization, high dimensional covariance matrix, portfolio allocation

JEL classification: C38, C55, G11, G17

* An earlier version of the paper was presented at the 3rd Konstanz-Lancaster Workshop on Finance and Econometrics 2017 in Lancaster. Financial support by the Graduate School of Decision Sciences (GSDS), the German Science Foundation (DFG) and the German Academic Exchange Service (DAAD) is gratefully acknowledged. For helpful comments on an earlier draft of the paper we would like to thank Lyudmila Grigoryeva and Karim Abadir. The usual disclaimer applies.

[a] Department of Economics, Universitätsstraße 1, D-78457 Konstanz, Germany. Phone: +49-7531-88-2657, email: Maurizio.Daniele@uni-konstanz.de.

[b] Department of Economics, Universitätsstraße 1, D-78457 Konstanz, Germany. Phone: +49-7531-88-2660, email: Winfried.Pohlmeier@uni-konstanz.de.

[c] Department of Economics, Universitätsstraße 1, D-78457 Konstanz, Germany. Phone: +49-7531-88-3753, email: Aygul.Zagidullina@uni-konstanz.de.

1 Introduction

The estimation of high-dimensional covariance matrices and their inverses (precision matrices) has recently received a great attention. In economics and finance it is central for portfolio allocation, risk measurement, asset pricing and graphical network analysis. The list of important applications from other areas of research includes, for example, the analysis of climate data, gene classification and image classification. What appears to be a trivial estimation problem for a large sample size T and a low dimensional vector of covariates, turns out to be demanding, if N is of the same order of magnitude or even larger than T. In these cases the sample covariance matrix becomes nearly singular and estimates the population covariance matrix poorly. Moreover, assumptions of standard asymptotic theory with $T \to \infty$, holding N fixed, turns out to be inappropriate and have to be replaced by assumptions allowing for both, T and N, approaching infinity.

In recent years numerous studies proposed alternative estimation approaches for high-dimensional covariance matrices, which differ in the way of bounding the dimensionality problem. Two major approaches are factor models imposing a lower dimensional factor structure for the underlying multivariate process and regularization strategies for the parameters of the covariance matrix or its eigenvalues (see Fan, Liao, and Liu (2016) for a recent survey on the estimation of large covariances and precision matrices). In this paper, we present an effective novel approach to the estimation of high-dimensional covariances, which profits from both branches of the literature. Our sparse approximate factor (SAF) approach to the estimation of high-dimensional covariance matrix is based on l_1-regularization of the factor loadings and thereby is able to account for weak factors and small but not necessary non-zero covariances among the covariates often found for economic or financial data.

Approaches to obtain consistent estimators by imposing a sparse structure on the covariance matrix directly include Bickel and Levina (2008a, 2008b), Cai and Liu (2011) and Cai and Zhou (2012). These thresholding approaches are shrinking small elements in the covariance matrix exactly to zero. While this may be a reasonable strategy e.g. for genetic data, this assumption may not be appropriate for economic or financial data, where variables are driven by common underlying factors. Such a feature may more appropriately captured by covariance matrices based factor representations.

In the literature on factor based covariance estimation Fan, Fan, and Lv (2008) consider the case of a strict factor representation with observed factors. This approach requires knowledge of additional observable variables (e.g. the Fama-French factors in the asset pricing framework), which may be an additional source of misspecification. Moreover, strict factor model representations impose the overly strong assumption of strictly uncorrelated idiosyncratic errors. This assumption was relaxed in Fan, Liao, and Mincheva (2011) and Fan, Liao, and Mincheva (2013), who propose a covariance estimator based on an approximate factor model representation. While Fan, Liao, and Mincheva (2011) shrink the entries of the covariance matrix of the idiosyncratic errors to zero using the adaptive thresholding technique by Cai and Liu (2011), the approach proposed in Fan, Liao, and Mincheva (2013) rests on the more general principal orthogonal complement thresholding method (POET) to allow for sparsity in the covariance matrix of the idiosyncratic errors.

Our SAF covariance matrix estimator extends the existing framework on factor based approaches by imposing sparsity on both, the factor loadings and the covariance matrix of the idiosyncratic errors. Unlike imposing sparsity for the covariance matrix directly by thresholding or l_1-norm regularization, the l_1-regularization of the factor loadings does not necessarily imply zero entities of the covariance matrix, but simply reduces the dimensionality problem in the estimation of the factor driven part of the covariance matrix. Moreover, the sparsity in the matrix of factor loadings allows for weak factors, which only affect a subset of the observed variables. Thus the SAF-approach relaxes the identifying assumption on the pervasiveness of the factors in the standard framework. This further implies that the eigenvalues of the covariance matrix corresponding to the common component are allowed to diverge at a slower rate than commonly considered (i.e. slower than $\mathcal{O}(N)$).

The weaker conditions on the eigenvalues allow us to derive the average consistency for the SAF covariance matrix estimator under the Frobenius norm under rather mild regularity conditions. To our knowledge this convergence result is new. Because of the fast diverging eigenvalues for estimators based on the approximate factor model, convergence has only be shown under the weaker weighted quadratic norm but not for the more general Frobenius norm (see e.g. Fan, Liao, and Mincheva (2013)). As a byproduct of our proof for the SAF covariance

matrix estimator, we also prove the consistency for the estimators of the sparse factor loadings, the factors and the covariance matrix of the idiosyncratic errors.

The favorable asymptotic properties of the SAF covariance matrix estimator are well supported by our Monte Carlo study based on different dimensions and alternative designs of the population covariance matrix. More precisely, the SAF covariance matrix estimator yields the lowest difference in the Frobenius norm to the true underlying covariance matrix compared to several competing estimation strategies.

Finally, in an a empirical study on the portfolio allocation problem we show, that the SAF covariance matrix estimator is a superior choice to construct the weights of the Global Minimum Variance Portfolio (GMVP) for low and large dimensional portfolios. Based on returns data from the S&P 500 the estimator uniformly outperforms portfolio strategies based on alternative covariance estimation approaches and modeling strategies including the $1/N$-strategy in terms of different popular out-of-sample portfolio performance measures.

The rest of the paper is organized 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 matrix by l_1-regularization. Section 3 discusses the theoretical setup and provides the convergence results. Implementation issues are discussed in Section 4. In Section 5, we present Monte-Carlo evidence on the finite sample properties of our new covariance estimator, while in Section 6 we show the performance of our approach when applied to the empirical portfolio allocation problem. Section 7 summarizes the main findings and gives an outlook on future research.

Throughout the paper we will use the following notation: $\pi_{\max}(A)$ and $\pi_{\min}(A)$ are the maximum and minimum eigenvalue of a matrix A. Further, $\|A\|$, $\|A\|_F$ and $\|A\|_1$ denote the spectral, Frobenius and the l_1-norm of A, respectively. They are defined as $\|A\| = \sqrt{\pi_{\max}(A'A)}$, $\|A\|_F = \sqrt{\mathrm{tr}\,(A'A)}$ and $\|A\|_1 = \max_j \sum_i |a_{ij}|$.

2 Factor Model Based Covariance Estimation

2.1 The Approximate Factor Model

The following analysis is based on the approximate factor model proposed by Chamberlain and Rothschild (1983) to obtain a lower dimensional representation of a possibly high dimensional covariance matrix. 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 in the time dimension, respectively. 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 $\lambda_i' f_t$. The approximate factor model allows for weak serial and cross-sectional correlations among the idiosyncratic components with a dense covariance matrix of the idiosyncratic error term vector, $\Sigma_u = \mathrm{V}\left[(u_{1t}, u_{2t}, \ldots u_{Nt})'\right]$. In matrix notation, (1) can be written as:

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

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

There are several estimation approaches for a factor model as given by (2). The principal component analysis (PCA) [1] and the quasi maximum likelihood estimation (QMLE) under

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

normality (see i.e. Bai and Li (2016)) are the two most popular ones. In the following, we pursue estimating the factor model by QMLE. 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 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 negative quasi log-likelihood function for the data 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 $S_x = \frac{1}{T} \sum_{t=1}^{T} x_t x_t'$ denotes the sample covariance matrix based on the observed data. Σ_F is the low dimensional covariance matrix of the factors. Within the framework of an approximate factor model the estimation of a full Σ_u is cumbersome, as the number of parameters to estimate is $\frac{N(N+1)}{2}$ which may exceed the sample size T. In order to overcome this problem, we treat Σ_u as a diagonal matrix in the first step and define $\Phi_u = \operatorname{diag}(\Sigma_u)$ denoting a diagonal matrix that contains only the elements of the main diagonal of Σ_u. Following Lawley and Maxwell (1971), we impose the following identification restrictions: $\Sigma_F = I_r$ and $\Lambda' \Phi_u^{-1} \Lambda$ is diagonal. Moreover, the diagonal entries of $\Lambda' \Phi_u^{-1} \Lambda$ are assumed to be distinct and arranged in a decreasing order.

Imposing these identifying restrictions has the advantage that the estimation of the covariance matrix of the factors becomes redundant. Hence, our objective function reduces to:

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

As the true covariance matrix of u_t allows for correlations of general form, but the previous objective function incorporates the error term structure of a strict factor model, (4) may be seen as a quasi-likelihood. Bai and Li (2016) show that the QML estimator based on (4) yields consistent parameter estimates. Hence, the consistency of Φ_u is not affected by the general form of cross-section and serial correlations in u_t.

The factors f_t can be estimated by generalized least squares (GLS):

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

where the estimates $\hat{\Lambda}$ and $\hat{\Phi}_u$ are the ones obtained from the optimization of the objective function in (4).

2.2 The Sparse Approximate Factor Model

The sparse approximate factor (SAF) 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:

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

where $\mu \geq 0$ denotes a regularization parameter. Note that the number of factors r is predetermined and assumed to be fixed. Sparsity is obtained by shrinking some elements of Λ to zero, such that not all r factors 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 contrast to the weak factor assumption introduced in the following, the pervasiveness assumption conventionally made for standard approximate factor models (e.g. Bai and Ng (2002), Stock and Watson (2002a)), implies that the r largest eigenvalues of $\Lambda'\Lambda$ diverge at the rate $\mathcal{O}(N)$. Intuitively, this means that all the factors are strong and a non-negligible fraction of the entire set of time series is affected. Consequently, the sparsity in the factor loadings matrix introduced in Assumption 2.1 below considerably relaxes the conventional pervasiveness assumption.

Assumption 2.1 (Weakness of Factor Loadings)**.**
There exists a constant $c > 0$ 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 $1/2 \leq \beta \leq 1$.[2]

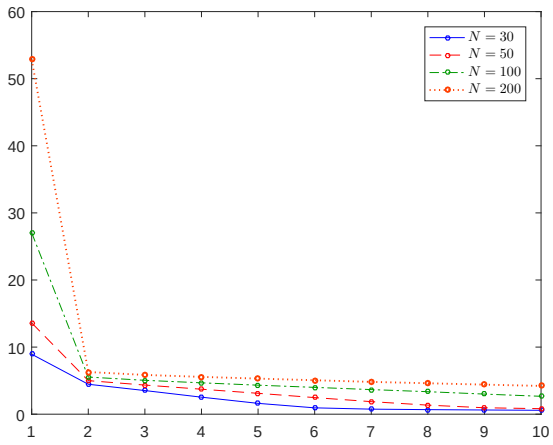
Assumption 2.1 implies that the r largest eigenvalues of $\Lambda'\Lambda$ diverge with the rate $\mathcal{O}\left(N^{\beta}\right)$, which can be much slower than in the standard approximate factor model. On the other hand, if $\beta = 1$, we are in the standard approximate factor model framework with strong factors (i.e. Fan, Liao, and Mincheva (2013), Bai and Liao (2016)). Hence, our sparse approximate factor model offers a convenient generalization of the standard one. Furthermore, 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:

$$\|\Lambda\| = \mathcal{O}\left(N^{\beta/2}\right) \leq \|\Lambda\|_1 \leq \sqrt{N}\,\|\Lambda\| = \mathcal{O}\left(N^{(1+\beta)/2}\right), \quad \text{for } 1/2 \leq \beta \leq 1. \tag{7}$$

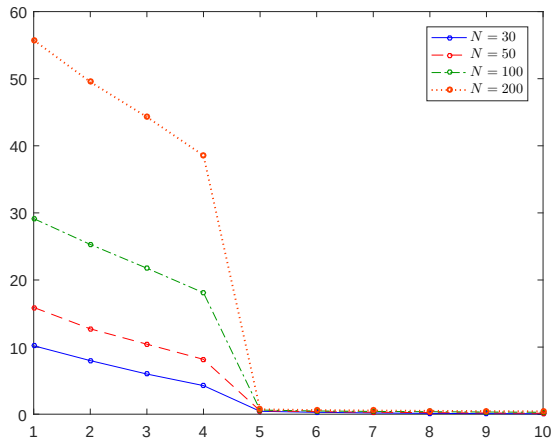
This result shows that imposing the weak factor assumption limits the amount of affected time series across all factors and hence requires a non-negligible amount of zero elements in each column of the factor loadings matrix. Nevertheless, the number of non-zero factor loadings can be arbitrarily small as β increases. Note, that the lower bound of equation (7) restricts the number of zero elements in each column of Λ, so that we can disentangle strong factors from the idiosyncratic component.

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 covariance matrix of the idiosyncratic errors. These characteristics can be observed in Figure 1, where both panels illustrate the eigenvalue structure of datasets, that are simulated only based on strong factors for $T = 450$ and different N. The panels differ solely in the number of factors included, where the left panel includes one strong factor and the right panel depicts the case of four strong factors. Both graphs reveal a clear partition in their respective eigenvalue structures, into sets of eigenvalues that diverge with the sample size N corresponding to the number of included strong factors and sets of bounded eigenvalues associated to the idiosyncratic components.

[2] The lower limit $1/2$ for β is necessary to consistently estimate the factors. See Lemma A.6 in Appendix A.1.



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



(b) Eigenvalues for simulated data with 4 strong factors with $T = 450$

Figure 1: Structure of the eigenvalues based on strong factors

However, such a clear separation in the eigenvalue structure of the covariance matrix cannot typically be found in real datasets. An example offers a dataset that contains the monthly asset returns of stocks constituents of the S&P 500 stock index available for the entire period of 450 months[3], whose eigenvalue distribution is illustrated in Figure 2. The graph shows a clear distinction of the first eigenvalue from the remaining eigenvalues. However, the remaining eigenvalues diverge at a slower rate and a clear separation between the common and idiosyncratic component as implied by the standard approximate factor model is impossible. Hence, the weak factor framework that allows for a slower divergence rate in the eigenvalues of the common component is more realistic for modeling the eigenvalue structure of real datasets. Figure 3 depicts the eigenvalue structure of a dataset, which is generated by one strong factor and three weak factors. This model with weak factors nicely mimics the decaying eigenvalue structure we observe for the S&P 500 asset returns.

[3] The same dataset is also used in our empirical application and is described in more detail in Section 6.

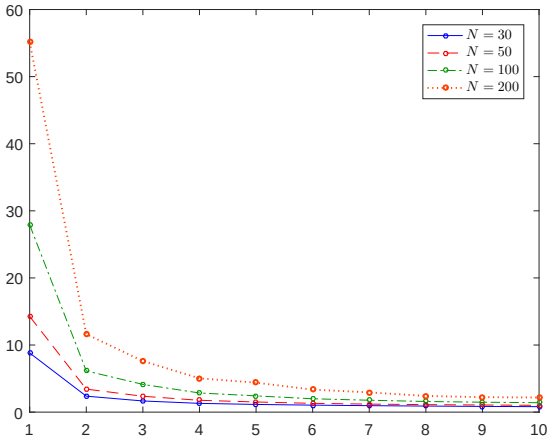


Figure 2: Eigenvalues for stock returns based on the S&P 500 index with $T = 450$

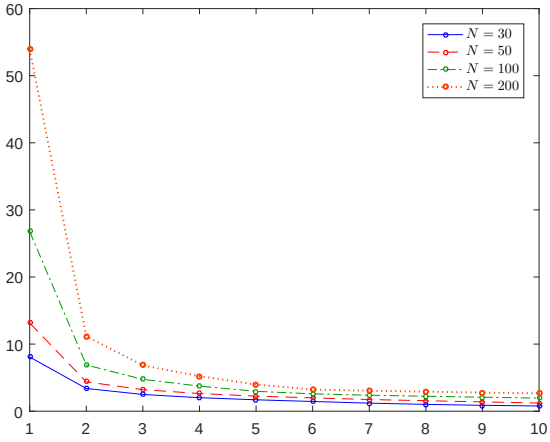


Figure 3: Eigenvalues for simulated data with 1 strong factor and 3 weak factors with $T = 450$

2.3 Estimation of the idiosyncratic error covariance matrix Σ_u

In order to relax the imposed diagonality assumption on Σ_u in the first step of our estimation, we re-estimate the covariance matrix of the idiosyncratic error term by means 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 residuals obtained from the estimation of an approximate factor model. Hence, it introduces sparsity in the idiosyncratic covariance matrix and offers a solution to the non-invertibility problem, generated using the sample covariance estimator, especially in high dimensional settings, where N is close or even larger than T. More specifically, the estimated idiosyncratic error covariance matrix $\hat{\Sigma}_u^\tau$ based on the POET method is defined as:

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

where $\hat{\sigma}_{u,ij}$ is the ij-th element of the sample covariance matrix $S_u = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{\Lambda}\hat{f}_t)(x_t - \hat{\Lambda}\hat{f}_t)'$ of the estimated factor model residuals, $\tau = \frac{1}{\sqrt{N}} + \sqrt{\frac{\log(N)}{T}}$ is a threshold[4] and $\mathcal{S}(\cdot)$ denotes the

[4] The threshold τ is based on the convergence rate of the idiosyncratic error covariance estimator specified in Lemma A.9. in Appendix A.2 .

soft-thresholding operator defined as:

$$\mathcal{S}(\sigma_{u,ij}, \tau) = \text{sign}(\sigma_{u,ij})(|\sigma_{u,ij}| - \tau)_+ \,. \tag{8}$$

In contrast to Fan, Liao, and Mincheva (2013), who use the residuals of a static factor model based on the PCA estimator, our estimates are based on the residuals obtained from our sparse factor model.

2.4 SAF covariance matrix estimation

The estimator of the data covariance matrix based on the approximate factor model is obtained according to $\Sigma = \text{Cov}[X] = \Lambda \Sigma_F \Lambda' + \Sigma_u$. Hereby, we first estimate the factors f_t and the factor loadings Λ according to our sparse factor model introduced in Section 2.2. Consistent estimates of Λ and f_t are obtained by MLE and GLS as given by (4) and (5), respectively. This yields the estimates of the common and idiosyncratic components of the approximate factor model defined in (1). The latter one is used as input to estimate Σ_u by the POET estimator introduced in Section 2.3. Hence, our SAF covariance matrix estimator is given by:

$$\hat{\Sigma}_{\text{SAF}} = \hat{\Lambda} \hat{\Sigma}_F \hat{\Lambda}' + \hat{\Sigma}_u^{\tau}, \tag{9}$$

where $\hat{\Sigma}_F$ denotes the sample estimator for the covariance matrix of the estimated factors, which is positive definite because the number of observations exceeds the number of factors. Further, using the convergence rate of the idiosyncratic error covariance matrix for the threshold τ also guarantees that $\hat{\Sigma}_u^{\tau}$ is positive definite with probability tending to one according to Bickel and Levina (2008a). Hence, the covariance matrix estimator $\hat{\Sigma}_{\text{SAF}}$ is positive definite by construction.

3 Large Sample Properties

3.1 Consistency of the Sparse Approximate Factor Model Estimator

In order to establish the consistency of the factor loadings matrix Λ and the data covariance matrix Σ estimators, we adapt the following standard assumptions:

Assumption 3.1 (Data generating process)**.**

(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* $k \leq r$*,*

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

(iii) Define the mixing coefficient:

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

where $\mathcal{F}_{-\infty}^{0}$ *and* $\mathcal{F}_{T}^{\infty}$ *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 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 the large deviation theory 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 identifying assumption in the factor model framework.

Assumption 3.2 (Sparsity)**.**

(i) $L_N = \sum_{i=1}^{N} \mathbb{1}\{\lambda_{ik} \neq 0\} = \mathcal{O}(N), \quad \forall k = 1, \ldots, r$

(ii) $S_N = \max_{i\leq N} \sum_{j=1}^{N} \mathbb{1}\{\sigma_{u,ij} \neq 0\}$,

where $\mathbb{1}\{\cdot\}$ *defines an indicator function that is equal to one if the boolean argument in braces is true.*

Assumptions 3.2 imposes sparsity conditions on Λ and Σ_u, where condition *(i)* defines the quantity L_N that reflects the number of non-zero elements in the factor loadings matrix Λ. As the number of factors r are assumed to be fixed, *(i)* restricts the number of non-zero elements in each column of Λ to be upper bounded by N. At the same time, this assumption allows for a sparse factor loadings matrix with less than N non-zero elements. Condition *(ii)* specifies S_N that quantifies the maximum number of non-zero elements in each row of Σ_u, following the definition of Bickel and Levina (2008a).

Theorem 3.1 (Consistency of the Sparse Approximate Factor Model Estimator)**.**
Under Assumptions 2.1, 3.1 and 3.2 the sparse factor model in (6) *satisfies the following properties, as* T *and* $N \to \infty$*:*

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

and

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

for $1/2 \leq \beta \leq 1$*.*
Hence, for $\log(N) = o(T)$ *and the regularization parameter* $\mu = o(1)$*, we have:*

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

Furthermore, for all $t \leq T$*:*

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

For the covariance matrix estimator of the idiosyncratic errors in the second step, specified in Section 2.3, we get:

$$\left\|\hat{\Sigma}_u^\tau - \Sigma_u\right\| = \mathcal{O}_p\left(S_N\sqrt{\mu^2 + \frac{N}{L_N}d_T}\right),$$

where $d_T = \frac{\log N^\beta}{N} + \frac{1}{N^\beta}\frac{\log N}{T}$.

The proof of Theorem 3.1 is given in Appendix A.1. Under the given regularity conditions this theorem establishes the average consistency in the Frobenius norm of the estimators for the factor loadings matrix and idiosyncratic error covariance matrix based on our sparse factor model. More specifically, we can see that Λ and Φ are estimated consistently, even though we impose the diagonality restriction on Σ_u in the first step of our estimation procedure. Consequently, the factors f_t estimated based on GLS are as well consistent. The lower limit $1/2$ on β is a necessary condition to achieve consistency. Intuitively this means that the factors should not be too weak such that there is still a clear distinction between the common and idiosyncratic component. Furthermore, if $S_N^2 d_T = o_p(1)$ and $S_N \mu = o(1)$, the second step estimator of Σ_u can be consistently estimated under the spectral norm.

3.2 Consistency of the Covariance Matrix Estimator

Finally, in this section we take a closer look on the asymptotic properties of the SAF covariance matrix estimator, given in Section 2.4. The following theorem gives the convergence rates of the covariance matrix estimator and of its inverse under different matrix norms.

Theorem 3.2 (Convergence Rates for the Covariance Matrix Estimator)**.**
Under Assumptions 2.1, 3.1 and 3.2, the covariance matrix estimator based on the sparse factor model in equation (9) *satisfies the following properties, as* $T, N \to \infty$ *and* $1/2 \leq \beta \leq 1$*:*

$$\frac{1}{N}\left\|\hat{\Sigma}_{\text{SAF}} - \Sigma\right\|_\Sigma^2 = \mathcal{O}_p\left(\left[\mu^2 + d_T\right]^2 + \left[\frac{N^\beta}{N} + \frac{S_N^2}{N}\right]\left[\mu^2 + d_T\right]\right), \tag{10}$$

$$\frac{1}{N}\left\|\hat{\Sigma}_{\text{SAF}} - \Sigma\right\|_F^2 = \mathcal{O}_p\left(N\left[\mu^2 + d_T\right]^2 + \left[N^\beta + S_N^2\right]\left[\mu^2 + d_T\right]\right) \tag{11}$$

and

$$\frac{1}{N}\left\|\hat{\Sigma}_{\text{SAF}}^{-1} - \Sigma^{-1}\right\|_F^2 = \mathcal{O}_p\left(\left[\frac{1}{N^\beta} + S_N^2\right]\left[\mu^2 + d_T\right]\right), \tag{12}$$

where $d_T = \frac{\log N^\beta}{N} + \frac{1}{N^\beta}\frac{\log N}{T}$ *and* $\|A\|_\Sigma = \frac{1}{\sqrt{N}}\left\|\Sigma^{-1/2} A \Sigma^{-1/2}\right\|_F$ *denotes the weighted quadratic norm introduced by Fan, Fan, and Lv (2008).*

The proof of Theorem 3.2 is given in Appendix A.3. Similar as for Theorem 3.1, we assume that the regularization parameter $\mu = o(1)$ and $\log(N) = o(T)$. Equation (10) in Theorem 3.2 shows that the covariance matrix estimator based on the sparse factor model in equation (9) is consistent if we consider the weighted quadratic norm for the entire set of possible values for β.

Generally, convergence under the Frobenius norm is hard to achieve because of the too fast diverging eigenvalues of the factor loadings matrix (see Fan, Liao, and Mincheva (2013)). However, Equation (11) shows that this is not the case for our SAF covariance matrix estimator if $\mu = o\left(N^{\beta/2}\right)$ and $1/2 \leq \beta \lessapprox 9/10$. Hence, the relaxation of the pervasiveness assumption in the standard approximate factor model to allow for weak factors leads to convergence of the covariance estimator under the Frobenius norm. The upper bound for β follows from the expression $\frac{N^\beta \log N^\beta}{N}$ in Equation (11) of Theorem 3.2.[5] Further, Equation (12) of Theorem 3.2 shows that the inverse of Σ_{SAF} is consistently estimated under the Frobenius norm.

4 Implementation Issues

For the implementation of the SAF model, we use a two-step estimation procedure that treats Σ_u in the first step as a diagonal matrix, denoted as Φ_u and re-estimates the idiosyncratic error covariance matrix in a second step by the method introduced in Section 2.3. Theorem 3.1 shows that this two-step procedure yields consistent estimates for Λ and Σ_u.

4.1 Majorized Log-Likelihood Function

The numerical minimization of the objective function (6) is cumbersome as it is not globally convex. This problem arises because the first term in (6) $\log\left|\det\left(\Lambda\Lambda' + \Phi_u\right)\right|$ is concave in Λ and Φ_u, whereas the second term $\text{tr}\left[S_x\left(\Lambda\Lambda' + \Phi_u\right)^{-1}\right]$ is convex. For our implementation we employ the majorize-minimize EM algorithm introduced by Bien and Tibshirani (2011). The idea of this optimization approach is to approximate the numerically unstable concave part $\log\left|\det\left(\Lambda\Lambda' + \Phi_u\right)\right|$ by its tangent plane, which corresponds to the following expression:

$$\log\left|\det\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)\right| + \text{tr}\left[2\hat{\Lambda}_m'\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)^{-1}\left(\Lambda - \hat{\Lambda}_m\right)\right], \tag{13}$$

[5] A closed form solution for the upper bound of β is not feasible, hence we numerically approximate the maximum value of β in the neighbourhood of one such that the expression $\frac{N^\beta \log N^\beta}{N}$ is converging to zero.

where the subscript m denotes the m-th step in an iterative procedure outlined in Section 4.2. Replacing the concave part in (4) by the convex expression in (13), yields the following majorized log-likelihood function:

$$\begin{aligned}\bar{\mathcal{L}}_m(\Lambda) = &\log\left|\det\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)\right| + \operatorname{tr}\left[2\hat{\Lambda}_m'\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)^{-1}\left(\Lambda - \hat{\Lambda}_m\right)\right] \\ &+ \operatorname{tr}\left[S_x\left(\Lambda\Lambda' + \hat{\Phi}_u\right)^{-1}\right]\end{aligned} \tag{14}$$

Augmenting the majorized log-likelihood by the l_1-penalty term, leads to the following optimization problem for our SAF model:

$$\begin{aligned}\min_{\Lambda,\Phi_u}\Bigg\{&\log\left|\det\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)\right| + \operatorname{tr}\left[2\hat{\Lambda}_m'\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)^{-1}\left(\Lambda - \hat{\Lambda}_m\right)\right] \\ &+\operatorname{tr}\left[S_x\left(\Lambda\Lambda' + \hat{\Phi}_u\right)^{-1}\right] + \mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik}|\Bigg\}\end{aligned} \tag{15}$$

As all three components in (15) are convex, the optimization problem simplifies considerably compared to the original problem in equation (6).

4.2 Projection Gradient Algorithm

In order to minimize (15) efficiently, we apply the fast projected gradient algorithm proposed by Bien and Tibshirani (2011). More specifically, we approximate the majorized log-likelihood $\bar{\mathcal{L}}_m(\Lambda)$ in (14) 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[6] and

$$\hat{A} = 2\left[\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)^{-1} - \left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)^{-1} S_x\left(\hat{\Lambda}_m\hat{\Lambda}_m' + \hat{\Phi}_{u,m}\right)^{-1}\right]\hat{\Lambda}_m, \tag{16}$$

[6] We set $t = 0.01$ for all our applications below.

which corresponds to the first derivative of $\bar{\mathcal{L}}(\Lambda)$ with respect to Λ. Hence, our final optimization problem corresponds to:

$$\min_{\lambda_{ik}} \frac{1}{2t} \sum_{k=1}^{r} \sum_{i=1}^{N} \left(\lambda_{ik} - \hat{\lambda}_{ik,m} + t\hat{A}_{ik,m} \right)^2 + \mu \sum_{k=1}^{r} \sum_{i=1}^{N} |\lambda_{ik}| \,. \tag{17}$$

The minimization of the objective function (17) can be carried out by computing its gradient with respect to λ_{ik} and setting it to zero, which yields:

$$\begin{aligned} \frac{\partial}{\lambda_{ik}} & \left[\frac{1}{2t} \sum_{k=1}^{r} \sum_{i=1}^{N} \left(\lambda_{ik} - \hat{\lambda}_{ik,m} + t\hat{A}_{ik,m} \right)^2 + \mu \sum_{k=1}^{r} \sum_{i=1}^{N} |\lambda_{ik}| \right] \\ & = \frac{1}{t} \sum_{k=1}^{r} \sum_{i=1}^{N} \left(\hat{\lambda}_{ik} - \hat{\lambda}_{ik,m} + t\hat{A}_{ik,m} \right) + \mu \sum_{k=1}^{r} \sum_{i=1}^{N} \nu_{ik} = 0 \,, \end{aligned} \tag{18}$$

where ν_{ik} denotes the subgradient of $|\lambda_{ik}|$. Solving (18) for a specific $\hat{\lambda}_{ik}$, gives:

$$\begin{aligned} \hat{\lambda}_{ik} + t \cdot \mu\nu_{ik} &= \hat{\lambda}_{ik,m} - t\hat{A}_{ik,m} \\ \hat{\lambda}_{ik} &= S\left(\hat{\lambda}_{ik,m} - t\hat{A}_{ik,m},\, t \cdot \mu \right), \end{aligned} \tag{19}$$

where S denotes the soft-thresholding operator and it is defined in equation (8). Equation (19) can be used to update the estimated factor loadings $\hat{\lambda}_{ik,m+1}$ given the estimate from the previous step $\hat{\lambda}_{ik,m}$.

In order to obtain an update for the estimate of the covariance matrix of the idiosyncratic error Φ_u, we use the EM algorithm suggested by Bai and Li (2012):

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

Our iterative estimation procedure for the SAF model can be briefly summarized as given below.

Iterative Algorithm

Step 1: Obtain an initial consistent estimate for the factor loading matrix Λ and for the diagonal idiosyncratic error covariance matrix Φ_u , i.e. by using unpenalized MLE and set $m = 1$.

Step 2: Update $\hat{\lambda}_{ik,m-1}$, by $\hat{\lambda}_{ik,m} = S\left(\hat{\lambda}_{ik,m-1} - t\hat{A}_{ik,m-1},\, t \cdot \mu\right)$

Step 3: Update $\hat{\Phi}_u$ using the EM algorithm in Bai and Li (2012), according to
$\hat{\Phi}_{u,m} = \text{diag}\left[S_x - \hat{\Lambda}_m \hat{\Lambda}'_{m-1}\left(\hat{\Lambda}_{m-1}\hat{\Lambda}'_{m-1} + \hat{\Phi}_{u,m-1}\right)^{-1} S_x\right]$

Step 4: If $\left\|\hat{\Lambda}_m - \hat{\Lambda}_{m-1}\right\|$ and $\left\|\hat{\Phi}_{u,m} - \hat{\Phi}_{u,m-1}\right\|$ are sufficiently small, stop the procedure, otherwise set $m = m + 1$ and return to *Step 2*.

Step 5: Estimate the factors by $\hat{f}_t = \left(\hat{\Lambda}'\hat{\Phi}_u^{-1}\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Phi}_u^{-1}x_t$, where $\hat{\Lambda}$ and $\hat{\Phi}_u$ are the parameter estimates after convergence.

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

For the high dimensional case of $N > T$, the sample covariance matrix S_x is not of full rank and hence leads to inconsistent parameter estimates. To overcome this problem, we adopt the solution proposed by Bien and Tibshirani (2011), who suggest augmenting the diagonal elements of S_x by an arbitrarily small $\varepsilon > 0$, when S_x is not of full rank. This augmentation stabilizes S_x and yields a non-degenerate solution for our sparse factor model.

4.3 Selecting the number of factors

In order to select the number of latent factors r, we follow Onatski (2010). To the best of our knowledge Onatski's method is the only one, which does not explicitly require that all factors are strong and have eigenvalues that diverge at rate N. Therefore, it is suitable for our setting, which allows as well for weak factors with a slower divergence rate. The method uses the difference in subsequent eigenvalues and chooses the largest $\hat{r}$ such that:

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

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 $\hat{r}$-th largest eigenvalue of the covariance matrix of X. For the choice of ξ, the empirical distribution of the eigenvalues of the data sample covariance matrix is taken into account[7]. However, the estimation of the number of factors based on the empirical

[7] We refer to Onatski (2010) for the detailed description on the determination of ξ.

distribution of the eigenvalues of the sample covariance matrix still requires a clear separation of the eigenvalues from the common and idiosyncratic component. Therefore, its selection accuracy depends on the degree of differentiability between the two components. Nevertheless, even if the selection method of Onatski (2010) overestimates the true number of factors, the sparsity assumption in our setting would allow us to disentangle the informative factors from those that are too weak. Thus, compared to the standard approximate factor model we avoid including redundant factors that amplify the misspecification error. Moreover, to further support the above argument, we refer to Yu and Samworth (2013), who show that in the weak factor setting the true number of factors is not asymptotically overestimated.

4.4 Choosing the tuning parameter

As for any penalized estimation approach, the selection of the tuning parameter μ is crucial, as it controls the degree of sparsity in the factor loadings matrix and it affects the efficiency of our estimator. In our case we select μ based on a type of Bayesian information criterion, according to:

$$IC(\mu) = \mathcal{L}\left(\hat{\Lambda}, \hat{F}\right) + 2\kappa_\mu \frac{N+T}{N \cdot T} \log\left(\frac{N \cdot T}{N+T}\right) \tag{20}$$

where κ_μ denotes the number of non-zero elements in the factor loading matrix $\hat{\Lambda}$ 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 term in (20) is identical to the IC_{p1}-criterion by Bai and Ng (2002) and has the property of converging to zero as both N and T approach infinity. Hence, the penalization vanishes as the sample size increases 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 (20) 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 Monte Carlo Evidence

In the following, we present Monte Carlo evidence on the finite sample properties of our new covariance estimator. In particular, we focus on the accuracy of the covariance matrix estimates depending on the dimensionality as well as on the strength of correlations in the true covariance matrix to be estimated. The simulation results for our SAF estimator are compared to the ones obtained from seven competing estimators that are popular in the literature.

5.1 Monte Carlo Design

For our Monte Carlo experiments we use two different designs of the true covariance matrix Σ. In the first case, we consider the uniform covariance matrix design used in Abadir, Distaso, and Žikeš (2014), which takes the following form:

$$\sigma_{ii}^{\mathrm{u}} = 1 \text{ and } \sigma_{ij}^{\mathrm{u}} = \eta \, \mathcal{U}_{(0,1)}, \quad \text{for } i \neq j, \tag{21}$$

where $\mathcal{U}_{(0,1)}$ denotes a standard uniform random variable, and we set $\eta \in \{0, 0.025, 0.05, 0.075\}$. In this setting, η controls for the correlations among the variables, where an increase in η amplifies the strength of the dependencies among the covariates.

For the second design, we use the sparse covariance matrix suggested by Bien and Tibshirani (2011), which contains zero entries for the off-diagonals with a certain probability. More specifically, the ij-th element of the covariance matrix $\sigma_{ij} = \sigma_{ji}$ is assigned to be non-zero with probability p, where $p \in \{0, 0.05, 0.1, 0.2\}$. Similar as in the uniform design, the diagonal elements are set to 1. The non-zero off-diagonal are independently drawn from the uniform distribution $\mathcal{U}_{(0,0.25)}$.

For both covariance matrix designs, we draw a time independent random data series X from a multivariate normal distribution with zero population mean. The time dimension T is set to 60, which relates to a dataset with 5 years of monthly data. The number of replications is 500. Further, we consider several dimensions for X and set $N \in \{30, 50, 100, 200\}$. As goodness of fit criterion for the difference between the true and the estimated covariance matrix we use the Frobenius norm.

5.2 Alternative covariance estimation strategies

A. *Factor Models*

1. *Approximate Factor Model* (AFM)

In our comparative study we include the standard approximate factor model with a dense factor loadings matrix introduced by Chamberlain and Rothschild (1983). The estimate of the covariance matrix from this model, denoted as $\hat{\Sigma}_{\text{AFM}}$, is similarly obtained as described in Section 2.4. However, the factors F and factor loadings Λ are estimated by PCA. Similar to SAF, we use the number of factors selected by Onatski (2010).

2. *Dynamic Factor Model* (DFM)

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:

$$x_{it} = B_i'(L) f_t + \varepsilon_{it}, \tag{22}$$

where $B_i(L) = \left(b_{i1} + b_{i2}L + \cdots + b_{ip}L^p\right)$ and L corresponds to the lag operator such that, $\forall p$, $L^p f_t = f_{t-p}$. In this setup $f_t = \left(f_{1t}, f_{2t}, \ldots, f_{qt}\right)'$ is a $(q \times 1)$-dimensional vector of dynamic factors following a VAR process and $b_{ij}, j = 1, \ldots, p$ denote the corresponding q-dimensional factor loadings. In order to estimate the dynamic factor model in (22), we use the two step procedure of Doz, Giannone, and Reichlin (2011). The estimation requires that the number of dynamic factors is given ex-ante. We use the consistent method by Bai and Ng (2007) to determine q.

B. *Covariance Matrix Shrinking Strategies*

Within the class of covariance matrix shrinkage strategies, we consider the method proposed by Ledoit and Wolf (2003) and the design-free estimator by Abadir, Distaso, and Žikeš (2014).

1. *Ledoit and Wolf (2003)* (LW)

The LW approach shrinks the sample covariance matrix S_x towards the covariance matrix of a

single index model that is well-conditioned. This yields the following definition:

$$\hat{\Sigma}_{\text{LW}} = \alpha^* S_x + (1 - \alpha^*)\hat{\Sigma}_{\text{SIM}},$$

where $\alpha^* \in (0, 1)$ is a constant, which corresponds to the shrinkage intensity. Ledoit and Wolf (2003) propose the following estimator to be used in practice $\hat{\alpha}^* = \frac{1}{T}\frac{\tau-\rho}{\gamma}$, where τ denotes the error on the sample covariance matrix, ρ measures the covariance between the estimation errors of $\hat{\Sigma}_{\text{SIM}}$ and S_x and γ accounts for the misspecification of the shrinkage target $\hat{\Sigma}_{\text{SIM}}$.

2. *Abadir, Distaso, and Žikeš (2014)* (ADZ)

The design-free estimator for the covariance matrix by Abadir, Distaso, and Žikeš (2014) aims to improve the estimation of the eigenvalues $\hat{P}$ of S_x, that is a possible source of ill-conditioning compared to the orthogonal eigenvectors $\hat{\Gamma}$ that are not affected by this problem by construction. The authors consider the following spectral decomposition of S_x:

$$S_x = \hat{\Gamma}\hat{P}\hat{\Gamma}'. \tag{23}$$

In order to obtain an improved estimator for P, X is split into two subsamples $X = \underset{N\times n}{X_1}\ ,\ \underset{N\times(T-n)}{X_2}$. Calculating the sample covariance matrix for the first n observations yields:

$$S_1 = \frac{1}{n}X_1 M_n X_1' = \hat{\Gamma}_1 \hat{P}_1 \hat{\Gamma}_1', \tag{24}$$

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 S_1 provides the matrix of eigenvectors $\hat{\Gamma}_1$ and the diagonal matrix of eigenvalues $\hat{P}_1$.

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

$$\tilde{P} = \text{diag}\left(\mathbf{V}\left[\hat{\Gamma}_1' X_2\right]\right) = \text{diag}\left(\hat{\Gamma}_1' S_2 \hat{\Gamma}_1\right). \tag{25}$$

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

$$\hat{\Sigma}_{\text{AZD}} = \hat{\Gamma}\tilde{P}\hat{\Gamma}'. \tag{26}$$

C. *Sparse Covariance Estimators*

The following estimators are explicitly designed to provide sparse covariance matrices. Hence, these models are appropriate for empirical settings that are reflected by our second simulation design.

1. *Rothman, Levina, and Zhu (2009) (ST)*

As a special case of the generalized thresholding estimators studied by Rothman, Levina, and Zhu (2009), we use the soft-thresholding method as a sparse covariance estimator and obtain:

$$\hat{\Sigma}_{\text{ST}} = \hat{\sigma}_{\text{ST},ij}, \quad \hat{\sigma}_{\text{ST},ij} = \begin{cases} \hat{\sigma}_{s,ij}, & i = j \\ \mathcal{S}(\hat{\sigma}_{s,ij}, c), & i \neq j \end{cases}$$

where $\hat{\sigma}_{s,ij}$ is the ij-th element of the sample covariance matrix and $\mathcal{S}$ denotes the soft-thresholding operator defined in (8). The authors suggest to select the thresholding parameter c, by minimizing the difference between $\hat{\Sigma}_{\text{ST}}$ and S_X in Frobenius norm based on cross-validation.

2. *Bien and Tibshirani (2011) (BT)*

The authors propose a penalized maximum likelihood estimator based on a lasso penalty in order to allow for sparsity in the covariance matrix and to reduce the effective number of parameters. More specifically, the following objective function is optimized:

$$\min_{\Sigma \succ 0} \quad \log\det(\Sigma) + \text{tr}\left(\Sigma^{-1}S_x\right) + \alpha \sum_{i=1}^{N}\sum_{j=1}^{N} \left|h_{ij}\sigma_{ij}\right|,$$

where α is a regularization parameter selected based on 5-fold cross-validation. The ij-th element of the selection matrix H is defined as $h_{ij} = \mathbb{1}\{i \neq j\}$ and enables an equal penalization of the off-diagonal elements and leaves the diagonal elements unaffected. Further, Bien and Tibshirani (2011) show that the estimated sparse covariance matrix is positive definite.

5.3 Simulation results

Table 1 below contains the Monte Carlo results for the uniform design of the true variance covariance matrix, while Table 2 gives the results based on the sparse covariance matrix design. Interestingly, we find a very similar and clear picture. In terms of the goodness of fit, our sparse approximate factor model approach provides the smallest Frobenius norm, i.e. the SAF fits the true covariance matrix best. This results hold for the two rather different designs, all dimensions and degrees of correlation between the variables. Note that the advantage of our SAF model in accurately estimating the true covariance matrix is even more pronounced when N increases, especially for the two high dimensional settings with $N = 100, 200$.

Concerning the alternative approaches, ST, which is rather similar to our approach, performs second best in most of the scenarios. Only for the uniform covariance matrix design for high dimensions and very strong dependencies ($N = 100, 200, \eta = .0075$), ADZ performs slightly better than ST. It is also interesting to note that direct l_1-norm penalization of the covariance matrix as suggested by Bien and Tibshirani (2011) does not do nearly as well as our approach, which profits from sparsity in the factor loadings matrix and thresholding of the covariance matrix of the idiosyncratic component.

Table 1: Simulation results - Uniform Covariance Matrix Design

N	Model	η			
		0	0.025	0.05	0.075
30	Sample	15.82	15.72	15.93	15.75
	AFM	6.53	6.66	6.83	6.90
	DFM	6.27	6.42	6.63	6.71
	SAF	**0.45**	**0.62**	**1.09**	**1.95**
	ST	1.02	1.20	1.77	2.86
	BT	2.99	3.18	3.71	4.78
	LW	3.15	3.36	3.90	4.38
	ADZ	2.02	2.24	2.77	3.54
50	Sample	42.96	43.31	42.79	43.32
	AFM	10.73	11.24	11.81	12.41
	DFM	10.60	11.04	11.60	12.26
	SAF	**0.50**	**1.00**	**2.37**	**4.48**
	ST	1.72	2.23	3.69	6.30
	BT	4.99	5.54	7.03	9.67
	LW	4.90	5.94	7.14	8.68
	ADZ	3.33	3.93	5.36	7.33
100	Sample	1.72E+02	1.71E+02	1.72E+02	1.72E+02
	AFM	23.99	25.91	27.97	32.62
	DFM	23.82	25.61	27.60	32.17
	SAF	**0.64**	**2.65**	**8.16**	**14.88**
	ST	3.36	5.46	11.56	21.88
	BT	10.25	12.31	18.58	28.83
	LW	10.68	15.17	19.98	26.64
	ADZ	6.82	8.75	14.34	21.02
200	Sample	6.81E+02	6.83E+02	6.84E+02	6.84E+02
	AFM	58.44	65.56	77.33	101.78
	DFM	58.03	64.66	76.90	100.87
	SAF	**0.81**	**8.92**	**28.73**	**53.48**
	ST	6.74	15.09	39.93	81.63
	BT	21.26	29.56	54.49	96.56
	LW	26.44	44.57	65.88	92.75
	ADZ	13.89	21.86	42.03	67.39

The table gives the mean goodness of fit in terms of the Frobenius norm (**??**) for $T = 60$. The sparse approximate factor model (SAF) is compared to the approximate factor model (AFM), the dynamic factor model (DFM), the soft-thresholding estimator (ST) of Rothman, Levina, and Zhu (2009), the sparse covariance estimator by Bien and Tibshirani (2011) (BT), the shrinkage estimator by Ledoit and Wolf (2003) (LW), and the design-free estimator by Abadir, Distaso, and Žikeš (2014) (ADZ).

Table 2: Simulation results - Sparse Covariance Matrix Design

N	Model	p			
		0	0.05	0.1	0.2
30	Sample	15.82	15.89	16.14	15.86
	AFM	6.53	11.01	11.45	12.61
	DFM	6.27	11.38	11.18	12.29
	SAF	**0.45**	**3.63**	**4.77**	**6.68**
	ST	1.02	5.60	6.45	8.97
	BT	2.99	7.44	8.27	9.83
	LW	3.15	6.86	6.84	7.39
	ADZ	2.02	6.49	6.45	7.21
50	Sample	42.96	43.62	42.98	43.53
	AFM	10.73	19.42	24.08	25.26
	DFM	10.60	19.76	24.47	25.22
	SAF	**0.50**	**8.21**	**13.17**	**14.85**
	ST	1.72	10.48	15.49	17.06
	BT	4.99	13.81	18.93	20.43
	LW	4.90	12.54	15.43	16.21
	ADZ	3.33	11.44	15.00	15.67
100	Sample	1.72E+02	1.72E+02	1.71E+02	1.72E+02
	AFM	23.99	45.50	47.07	52.28
	DFM	23.82	46.81	47.05	52.08
	SAF	**0.64**	**20.57**	**22.14**	**28.49**
	ST	3.36	24.45	25.46	32.05
	BT	10.25	31.55	32.52	39.13
	LW	10.68	31.09	31.70	37.51
	ADZ	6.82	27.13	27.53	32.92
200	Sample	6.81E+02	6.84E+02	6.84E+02	6.84E+02
	AFM	58.44	111.01	110.47	113.82
	DFM	58.03	112.33	110.58	113.26
	SAF	**0.81**	**48.93**	**48.57**	**52.89**
	ST	6.74	55.54	55.07	59.45
	BT	21.26	70.50	69.88	74.28
	LW	26.44	78.12	78.72	85.55
	ADZ	13.89	61.36	60.73	64.37

The figure gives the mean goodness of fit in terms of the Frobenius norm (**??**) for $T = 60$. The sparse approximate factor model (SAF) is compared to the approximate factor model (AFM), the dynamic factor model (DFM), the soft-thresholding estimator (ST) of Rothman, Levina, and Zhu (2009), the sparse covariance estimator by Bien and Tibshirani (2011) (BT), the shrinkage estimator by Ledoit and Wolf (2003) (LW), and the design-free estimator by Abadir, Distaso, and Žikeš (2014) (ADZ).

6 An Application to Portfolio Choice

Empirical portfolio models, particularly when applied to large asset spaces, suffer from a high degree of instability. The estimation of N mean and $N(N+1)/2$ variance-covariance parameters yields extremely noisy estimates of portfolio weights with large standard errors. It is well-documented that these estimated portfolios show poor out-of-sample performance, extreme short positions and no diversification (e.g. Jobson and Korkie (1980) and Michaud (1989)). In order to mitigate these shortcomings and to improve portfolio estimates against extreme estimation noise, a range of alternative strategies have been proposed including the shrinkage estimation of the covariance matrix of asset returns (Ledoit and Wolf (2003); Ledoit and Wolf (2017) and Kourtis, Dotsis, and Markellos (2012)).

In the following, we investigate to what extend our SAF model can be used to obtain robustified estimates of high dimensional covariance matrices of asset returns as input for empirical portfolio models. In an out-of-sample portfolio forecasting experiment, we compare the performance of the global minimum variance portfolio (GMVP) strategy based on a covariance matrix estimated by our sparse factor model to popular alternative portfolio strategies with regularized covariance estimators. As in many other studies, we restrict our analysis to the GMVP, because its vector of portfolio weights, $\omega = \frac{\Sigma^{-1}1_N}{1_N'\Sigma^{-1}1_N}$, is solely a function of the covariance matrix of the asset returns. Thus, for estimating the GMVP the mean vector of asset returns is redundant and its empirical performance solely depends on the quality of the covariance matrix estimator.

6.1 Data and Design 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 treasury 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[8]. In order to check the performance of our estimator with respect to the dimensionality of the asset space, we consider the following portfolio sizes: $N \in \{30, 50, 100, 150, 200\}$. Out of the 205 stocks, we select at random individual subsets from the overall number of assets and work with the selected assets for the entire forecasting experiment.

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

Since by construction a theoretical portfolio built on a subset of assets from a larger portfolio cannot outperform the larger one, an observed inferiority of the larger empirical portfolio can only be the consequence of higher estimation noise due to the larger dimensionality, which over-compensates for the ex-ante theoretical superiority. Therefore, this selection strategy provides us with insights into the impact of estimation noise on the performance of empirical portfolios.

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 historic data. Thus, at 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 $\hat{r}(s)_{t+1} = \hat{\omega}(s)' r_{t+1}$ for the period $t+1$ for the 12 different estimation strategies $s = 1, \ldots, 12$. All portfolios are rebalanced on a monthly basis. This generates a series of $T-h$ out-of-sample portfolio returns. The results are then used to estimate the mean $\mu(s)$ and variance $\sigma^2(s)$ of the portfolio returns for each strategy by their empirical counterparts:

$$\hat{\mu}(s) = \frac{1}{T} \sum_{t=h+1}^{T} \hat{r}_t(s) \qquad \text{and} \qquad \hat{\sigma}^2(s) = \frac{1}{T-1} \sum_{t=h+1}^{T} \big(\hat{r}_t(s) - \hat{\mu}(s)\big)^2 \,. \tag{27}$$

We repeat this procedure 100 times to avoid that the out-of-sample results depend on the initially randomly selected stocks. Hence, all results reported below are average outcomes across the 100 forecasting experiments.

6.2 Competing Estimation Strategies

For our empirical portfolio application, we consider factor based models (with latent and observable factors) as well as models based on direct shrinkage of the covariances besides two fundamental baseline strategies: the naive $1/N$ strategy and the simple plug-in estimator for the GMVP.[9]

- *Equally Weighted Portfolio ($1/N$)*

The equally weighted or $1/N$-portfolio strategy comprises identical portfolio weights of size $1/N$, for each of the risky assets. By ignoring any type of optimizing portfolio strategy it often serves

[9] We also included in our extended comparative study the approaches by Frahm and Memmel (2010) and Pollak (2011), which are based on direct shrinkage of the portfolio weights. The performance of these two models was clearly inferior, so that we refrained from giving the results here. However, they can be obtained from the authors on request.

as a benchmark case to be outperformed in empirical performance comparisons. As the weights have not to be estimated, the $1/N$-strategy strategy is free from any estimation risk. Moreover, the $1/N$ portfolio weights can be considered as the outcome for portfolio weights under extreme l_2-penalization. DeMiguel, Garlappi, and Uppal (2009) find that the mean-variance portfolio and most of its extensions cannot significantly outperform the $1/N$ portfolio.[10]

- *Plug-in GMVP*

As the extreme alternative to the $1/N$-strategy, we consider the plug-in estimator of the GMVP based on the sample covariance matrix of the asset returns. The plug-in estimator is free from any type of regularization. The plug-in approach yields unbiased estimates of the true weights (Okhrin and Schmid (2006)), but the weight estimates are extremely unstable when the asset space is large relative to the time series dimension. For some of our empirical designs with $N = 100, 150, 200,$ the asset dimension exceeds the sample size, $T = 60$. For these cases the plug-in estimator is infeasible, because the sample covariance matrix is singular.

Factor models with observable factors

In addition, we consider two factor models that have been frequently used in the empirical finance literature. Contrary to the approximate factor models, the factors in these models are not latent but observable time series variables. In this respect, these type of models incorporate more information than approaches, which solely use the information on the return process itself to estimate the covariance matrix of returns. However, the inclusion of additional time series information may give rise to an additional source of misspecification, if the factor specification fails to describe the true data generating process properly.

- *The Single Index Model* (SIM)

The single index model by Sharpe (1963) is based on a single observable factor, f_{1t}, representing the excess market return:

$$x_{it} = \alpha + \beta_{i1} f_{1t} + \varepsilon_{it} \,. \tag{28}$$

[10] Kazak and Pohlmeier (2018) show, however, that conventional portfolio performance tests suffer from very low power, so that the rejection of null hypothesis of equal performance of a given data-based strategy and the $1/N$-strategy is very unlikely.

In our study, we use as a proxy for the market return, the value-weighted returns of all Center for Research in Security Prices (CRSP) firms incorporated in the US and listed on the AMEX, NASDAQ, or the NYSE. The one-month treasury bill rate serves as the risk free rate to construct the excess market returns. The estimator for the covariance matrix of the single index model is given by:

$$\hat{\Sigma}_{\text{SIM}} = \hat{\beta}_1 \hat{\sigma}_{f_1} \hat{\beta}_1' + \hat{D}\,,$$

where $\hat{\Sigma}_{f_1}$ denotes the sample variance of the market excess returns. $\hat{\beta}_1$ represents the OLS estimates of the factor loadings and $\hat{D}$ is a diagonal matrix of the OLS residual variances of regression model (28) assuming that the observed factor picks up the cross-correlations of the returns completely.

- *Fama and French 3-Factor Model (FF3F)*

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

$$X_t = \beta_1 f_{1t} + \beta_2 f_{2t} + \beta_3 f_{3t} + \varepsilon_t\,. \tag{29}$$

The first factor f_1 is identical to the one of the one-factor model in (28). The second factor f_{2t} often denoted by the acronym 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[11]. The third factor f_{3t}, 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 [12]. In matrix notation (29) is given by:

$$X = \beta F' + \varepsilon\,, \tag{30}$$

[11] 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.

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

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

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

Covariance Matrix Estimation Strategies

Further, we consider from the group of covariance matrix estimators introduced in Section 5.2 the plug-in estimation approaches for the GMVP weights based on the shrinkage covariance estimator by Ledoit and Wolf (2003) (LW) and the design free estimator by Abadir, Distaso, and Žikeš (2014) (ADZ). We refrain from considering the soft thresholding estimator (ST), because, as mentioned earlier, this estimator does not necessarily yield estimates of the covariance matrix, which are a positive definite, hence its inverse, needed for the computation of the GMVP weights, might be ill-conditioned. However, in addition to the estimators considered in the Monte Carlo study in Section 5, we consider the shrinkage approach by Kourtis, Dotsis, and Markellos (2012) (KDM), which targets directly on the inverse of the covariance matrix and is particularly designed for portfolio applications.

- *Kourtis, Dotsis, and Markellos (2012) (KDM)*

The estimation method by Kourtis, Dotsis, and Markellos (2012) directly shrinks the inverse of the sample-based covariance matrix S_x towards the identity matrix I_N and the inverse of the covariance matrix resulting from a single index model by Sharpe (1963), according to the following equation:

$$\hat{\Sigma}^{-1}_{\text{KDM}} = c_1 S_x^{-1} + c_2 I_N + c_3 \hat{\Sigma}^{-1}_{\text{SIM}}. \tag{31}$$

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 single index model model $\hat{\omega}_{\text{SIM}}$. In order to select the optimal shrinkage coefficients in (31),

the authors suggest minimizing the out-of-sample portfolio variance using cross-validation. It is important to note that this portfolio strategy is also applicable for the case when $N > T$. In order to obtain reliable results for the inverse of S_x the authors use the Moore-Penrose pseudo-inverse.

6.3 Criteria for Performance Evaluation

For our analysis, we consider the following four different evaluation criteria 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 $\hat{\sigma}^2(s)$ given in Equation (27). This measure yields an estimate of the performance criterion the GMVP strategy is designed for. Moreover, for the GMVP-strategy a clear ranking concerning portfolios of different dimensions exists, i.e. $\sigma^2(N) \leq \sigma^2(N')$ for $N \leq N'$, while the variance of the equally weighted portfolio is independent of the portfolio dimension.

2. *Average Return (AV):* The out-of-sample average return is expressed as $\hat{\mu}(s)$ from (27).

3. *Certainty Equivalent (CE):* The CE is defined as $\widehat{\mathrm{CE}}(s) = \hat{\mu}(s) - \frac{\gamma}{2} \cdot \hat{\sigma}^2(s)$, where γ specifies the risk aversion of the investor. Following DeMiguel, Garlappi, Nogales, and Uppal (2009) we set $\gamma = 2$. The CE is defined as the risk-free rate that an investor is willing to accept to make him indifferent to an investment based on the risky portfolio strategy s in terms of expected utility.

4. *Sharpe Ratio (SR):* The Sharpe ratio is given by $\widehat{\mathrm{SR}}(s) = \hat{\mu}(s)/\hat{\sigma}(s)$.

6.4 Out-of-Sample Portfolio Performance

Table 3 contains the results of our comparative study on the out-of-sample performance of different portfolio estimation approaches. The results represent average outcomes across the 100 different forecasting experiments for each of the four performance measures. Our sparse approximate factor model (SAF) yields the lowest out-of-sample portfolio standard deviation for all portfolio dimensions, i.e. it is performing best for the performance criterion the GMVP-strategy is designed for.

In theory, the GMVP-strategy may not necessarily outperform the $1/N$-strategy in terms of the remaining three performance criteria, since it completely disregards optimization with respect to the expected portfolio return. Nevertheless, our SAF model also outperforms the $1/N$-strategy and the other estimation approaches in terms of AV, CE and SR, which depend on the expected return. In the portfolio forecasting experiment our regularization method does best for the expected out-of-sample portfolio return.

It is of utmost importance to note that the superiority of our approach does not only hold for different performance measures, but also for all portfolio dimensions. The SAF model performs best for low, but also for high dimensional portfolios, for which the sample size is much smaller than the portfolio dimension, i.e. $T \ll N$. This indicates, at least for this specific application, that the selection of the penalty parameter is reasonable.

Table 3: Estimation results for the Portfolio Application

Model	1/N	GMVP	**SAF**	AFM	DFM	SIM	FF3F	LW	KDM	ADZ
					N = 30					
SD	0.0477	0.0665	**0.0458**	0.0498	0.0484	0.0478	0.0469	0.0489	0.0483	0.0474
AV	0.0084	0.0075	**0.0086**	0.0074	0.0075	0.0085	0.0081	0.0077	0.0081	0.0079
CE	0.0062	0.0031	**0.0065**	0.0049	0.0052	0.0062	0.0059	0.0053	0.0058	0.0056
SR	0.1766	0.1123	**0.1868**	0.1475	0.1553	0.1768	0.1727	0.1581	0.1678	0.1658
					N = 50					
SD	0.0467	0.1165	**0.0446**	0.0496	0.0479	0.0468	0.0458	0.0486	0.0479	0.0470
AV	0.0084	0.0082	**0.0088**	0.0075	0.0075	0.0084	0.0079	0.0080	0.0081	0.0081
CE	0.0062	-0.0055	**0.0068**	0.0050	0.0052	0.0062	0.0058	0.0057	0.0058	0.0059
SR	0.1797	0.0706	**0.1973**	0.1510	0.1564	0.1799	0.1736	0.1651	0.1687	0.1717
					N = 100					
SD	0.0462	-	**0.0435**	0.0479	0.0468	0.0463	0.0448	0.0466	0.0463	0.0456
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
					N = 150					
SD	0.0460	-	**0.0429**	0.0466	0.0459	0.0460	0.0444	0.0453	0.0452	0.0445
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
					N = 200					
SD	0.0459	-	**0.0426**	0.0459	0.0454	0.0459	0.0440	0.0444	0.0450	0.0440
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

Note: The sparse approximate factor model (SAF) in the third column 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) (SIM), the Three-Factor Model by Fama and French (1993) (FF3F), the estimators by Ledoit and Wolf (2003) (LW), Kourtis, Dotsis, and Markellos (2012) (KDM), Abadir, Distaso, and Žikeš (2014) (ADZ).

As mentioned earlier, increasing the portfolio dimension does not necessarily improve the out-of-sample performance of an empirical portfolio as the theoretical gains maybe overcompensated by the increase in estimation noise due to the increase in the number of parameters to be estimated. It is not too surprising that this phenomenon is most dramatically pronounced for the plug-in estimator of the GMVP, but we also find it to some extend for the DFM. Moreover, for the SIM and FF3F, we do not find a strict monotonicity between portfolio dimension and portfolio performance, while the performance of our SAF model strictly increases with N.

While in the portfolio forecasting experiment for any performance measure and any portfolio dimension our sparse factor model shows the best performance, there is no clear further ranking regarding the other approaches. FF3F is performing second best in terms of the minimization of portfolio risk for all portfolio dimensions, but it is outperformed by other estimation approaches, when performance measures other than the portfolio risk are considered.

Our comparative study also confirms the findings of DeMiguel, Garlappi, and Uppal (2009) that the $1/N$ portfolio is a strong competitor for many alternative portfolio strategies. 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 equally weighted portfolio, although it is very close to it. In terms of the portfolio risk, only our method and FF3F reveal performance superior to the $1/N$ portfolio for low dimensions of the asset space. The picture slightly changes, when higher asset dimensions ($N > 50$) are considered. For higher dimensions, the method by Abadir, Distaso, and Žikeš (2014) is a serious competitor to the $1/N$ portfolio. This mirrors our finding from the simulation study in Section 5, where the ADZ estimator performs comparatively well in high dimensional settings with strong linear dependencies.

Table 4 in Appendix B provides additional insights into the quality of the weight estimates. The summary statistics indicates that the outstanding performance of the SAF model results from effectively stabilizing the estimated portfolio weights by avoiding extreme positions (moderate minima and maxima in the weight estimates) and by the low standard deviations, while the relative good performance of SIM and FF3F result from very low variation in the portfolio weights, which come for the SIM with $N = 200$ close to the constant weights of the equally weighted portfolio.

In order to check the robustness of our findings, which are based on data from January

1974 until April 2015, we also consider forecasts based on subperiods. We restrict our attention to the standard deviation of the out-of-sample portfolio returns and consider how a gradual increase of the evaluation sample affects the performance of the competing estimators. 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). Special attention is given to the periods before and after the financial crisis in 2007. The graphs indicate that the SAF estimator also provides for different subperiods the lowest portfolio standard deviation compared to FF3F and ADZ. Note, that the difference is more pronounced, when the recent financial crisis period is included. Hence, in comparison to our SAF model both, FF3F and ADZ, fail to pick up the changing risk during the crisis and, as a result, they provide more volatile portfolio estimates.

7 Conclusions

In this paper we propose a novel approach for the estimation of high dimensional covariance matrices based on a sparse approximate factor model. The estimator allows for sparsity in the factor loadings matrix by shrinking single elements of the factor loadings matrix to zero. Hence, this setting reduces the number of parameters to be estimated and therefore leads to a reduction in estimation noise. Furthermore, the sparse factor model framework allows for weak factors, that only affect a subset of the available time series. Thus, our framework offers a convenient generalization to the pervasiveness assumption in the standard approximate factor model that solely leads to strong factors.

We prove average consistency under the Frobenius norm for the factor loadings matrix estimator and consistency in the spectral norm for the idiosyncratic component covariance matrix estimator based on our sparse approximate factor model. The factors estimated using the GLS method are also shown to be consistent. Furthermore, we derive average consistency for our factor model based covariance matrix estimator under the Frobenius norm for a particular rate of divergence for the eigenvalues of the covariance matrix corresponding to the common component. To the best of our knowledge, this result has not been shown in the existing literature because of the fast diverging eigenvalues (see e.g. Fan, Liao, and Mincheva (2013)). Additionally, we

provide consistency results of our covariance matrix estimator under the weighted quadratic norm (Fan, Fan, and Lv (2008)).

In our Monte Carlo study we analyze the finite sample properties of our covariance matrix estimator for different simulation designs for the true underlying covariance matrix. The results show that our estimator offers the lowest difference in Frobenius norm to the true covariance matrix compared to the competing estimators. Further, the benefit of the covariance matrix estimator based on our sparse factor model is even more pronounced if the dimensionality of the problem increases.

In an out-of-sample portfolio forecasting experiment we compare the performance of the global minimum variance portfolio based on the covariance matrix estimator of our sparse approximate factor model to alternative estimation approaches frequently used in the literature. The forecasting results reveal that our estimator yields the lowest average out-of-sample portfolio standard deviation across different portfolio dimensions. At the same time, it generates the highest Certainty Equivalent and Sharpe Ratio compared to all considered portfolio strategies. The performance gains of our SAF model are especially pronounced during the recent financial crisis. Hence, our estimator has a stabilizing impact on the portfolio weights, especially during highly volatile periods.

The results of our out-of-sample portfolio forecasting study show a substantial reduction of the portfolio standard deviation of the dynamic factor model compared to the standard approximate factor model, especially for small asset dimensions. Hence, it would be interesting to analyze if a possible extension of our SAF model by considering dynamic factors, would as well lead to a more efficient estimation of the covariance matrix. We leave this for future research.

References

Abadir, K. M., W. Distaso, and F. Žikeš (2014): "Design-free estimation of variance matrices," *Journal of Econometrics*, 181(2), 165–180.

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

——— (2016): "Maximum likelihood estimation and inference for approximate factor models of high dimension," *Review of Economics and Statistics*, 98(2), 298–309.

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

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

——— (2007): "Determining the number of primitive shocks in factor models," *Journal of Business & Economic Statistics*, 25(1).

Bernstein, D. S. (2009): *Matrix Mathematics: Theory, Facts, and Formulas Ed. 2*. Princeton University Press.

Bickel, P. J., and E. Levina (2008a): "Covariance regularization by thresholding," *The Annals of Statistics*, pp. 2577–2604.

——— (2008b): "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227.

Bien, J., and R. J. Tibshirani (2011): "Sparse estimation of a covariance matrix," *Biometrika*, 98(4), 807.

Cai, T., and W. Liu (2011): "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 106(494), 672–684.

Cai, T. T., and H. H. Zhou (2012): "Optimal rates of convergence for sparse covariance matrix estimation," *Ann. Statist.*, 40(5), 2389–2420.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51(5), 1281–304.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009): "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," *Management Science*, 55(5), 798–812.

DeMiguel, V., L. Garlappi, and R. Uppal (2009): "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?," *Review of Financial Studies*, 22(5), 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(1), 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(1), 3–56.

Fan, J., Y. Fan, and J. Lv (2008): "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 147(1), 186–197.

Fan, J., Y. Liao, and H. Liu (2016): "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, 19(1), C1–C32.

Fan, J., Y. Liao, and M. Mincheva (2011): "High Dimensional Covariance Matrix Estimation in Approximate Factor Models.," *Annals of Statistics*, 39(6), 3320–3356.

——— (2013): "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 603–680.

Frahm, G., and C. Memmel (2010): "Dominating estimators for minimum-variance portfolios," *Journal of Econometrics*, 159(2), 289–302.

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

Jobson, J. D., and B. Korkie (1980): "Estimation for Markowitz Efficient Portfolios," *Journal of the American Statistical Association*, 75(371), 544–554.

Kazak, E., and W. Pohlmeier (2018): "Testing Out-of-Sample Portfolio Performance," Working paper, Center for Finance and Econometrics, University of Konstanz.

Kourtis, A., G. Dotsis, and R. N. Markellos (2012): "Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix," *Journal of Banking & Finance*, 36(9), 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(5), 603–621.

——— (2017): "Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks," *The Review of Financial Studies*, 30(12), 4349–4388.

Michaud, R. O. (1989): "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?," *Financial Analysts Journal*, 45(1), 31–42.

Okhrin, Y., and W. Schmid (2006): "Distributional Properties of Portfolio Weights," *Journal of Econometrics*, 134(1), 235–256.

Onatski, A. (2010): "Determining the Number of Factors from Empirical Distribution of Eigenvalues," *Review of Economics and Statistics*, 92(4), 1004–1016.

——— (2012): "Asymptotics of the principal components estimator of large factor models with weakly influential factors," *Journal of Econometrics*, 168(2), 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*, pp. 37–40. IEEE.

Rothman, A. J., E. Levina, and J. Zhu (2009): "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, 104(485), 177–186.

Sharpe, W. F. (1963): "A simplified model for portfolio analysis," *Management Science*, 9(2), 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(460), 1167–1179.

——— (2002b): "Macroeconomic forecasting using diffusion indexes," *Journal of Business & Economic Statistics*, 20(2), 147–162.

Yu, Y., and R. J. Samworth (2013): "Discussion on Fan, Liao and Mincheva "Large covariance estimation by thresholding principal orthogonal complements"," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 650–652.

Appendix

A Proofs

A.1 Consistency of the Sparse Approximate Factor Model Estimator

Proof. Theorem 3.1 (Consistency of the Sparse Approximate Factor Model Estimator)

Define the penalized log-likelihood

$$\mathcal{L}_p(\Lambda, \Sigma_u) = Q_1(\Lambda, \Sigma_u) + Q_2(\Lambda, \Sigma_u) + Q_3(\Lambda, \Sigma_u), \tag{32}$$

where

$$\begin{aligned}
Q_1(\Lambda, \Sigma_u) &= \frac{1}{N}\log|\Sigma_u| + \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right) - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\text{tr}\left(S_u\Sigma_{u0}^{-1}\right) \\
&\quad + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik}| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}| \\
Q_2(\Lambda) &= \mathcal{O}\left(\frac{L_N}{N}\right)\max_{i\leq N}\sum_{k=1}^{r}|\lambda_{ik} - \lambda_{ik0}|^2 = \mathcal{O}\left(\frac{L_N}{N}\right)\max_{i\leq N}\|\lambda_i - \lambda_{i0}\|^2 \\
Q_3(\Lambda, \Sigma_u) &= \frac{1}{N}\log\left|\Lambda\Lambda' + \Sigma_u\right| + \frac{1}{N}\text{tr}\left(S_x\left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right) - Q_2(\Lambda) \\
&\quad - \frac{1}{N}\log|\Sigma_u| - \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)
\end{aligned}$$

Therefore, we can see that equation (32) can be written as

$$\begin{aligned}
\mathcal{L}_p(\Lambda, \Sigma_u) &= \frac{1}{N}\log\left|\Lambda\Lambda' + \Sigma_u\right| + \frac{1}{N}\text{tr}\left(S_x\left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right) \\
&\quad - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\text{tr}\left(S_u\Sigma_{u0}^{-1}\right) \\
&\quad + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik}| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|
\end{aligned} \tag{33}$$

Define the set,

$$\begin{aligned}
\Psi_\delta = \Big\{(\Lambda, \Sigma_u) : \delta^{-1} &< \pi_{\min}\left(\frac{\Lambda'\Lambda}{N^\beta}\right) \leq \pi_{\max}\left(\frac{\Lambda'\Lambda}{N^\beta}\right) < \delta \\
&\delta^{-1} < \pi_{\min}\left(\Sigma_u\right) \leq \pi_{\max}\left(\Sigma_u\right) < \delta\Big\}, \quad \text{for } 1/2 \leq \beta \leq 1.
\end{aligned}$$

Further, $\Phi_u = \text{diag}\,(\Sigma_u)$ and denotes a covariance matrix that contains only the elements of the main diagonal of Σ_u.

We impose the following sparsity assumptions on Λ and Σ_u:

$$L_N = \sum_{i=1}^{N} \mathbb{1}\left\{\lambda_{ik} \neq 0\right\} = \mathcal{O}\left(N\right), \quad \forall k = 1, \ldots, r$$
$$S_N = \max_{i \leq N} \sum_{j=1}^{N} \mathbb{1}\left\{\sigma_{u,ij} \neq 0\right\},$$

where $\mathbb{1}\left\{\cdot\right\}$ defines an indicator function that is equal to one if the argument in braces is true. Hence, L_N is the number of non-zero elements in the factor loadings matrix Λ and S_N denotes the maximum number of non-zero elements in each row of Σ_u, following Bickel and Levina (2008a).

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.

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

Proof. The unpenalized log-likelihood

$$\mathcal{L}(\Lambda, \Sigma_u) = \frac{1}{N} \log \left|\Lambda\Lambda' + \Sigma_u\right| + \frac{1}{N} \text{tr}\left(S_x \left(\Lambda\Lambda' + \Sigma_u\right)^{-1}\right), \tag{34}$$

can be decomposed in a similar fashion as in *Lemma A.2.* in Bai and Liao (2016).

The first term in equation (34) can be written as:

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

Hence, we have

$$\frac{1}{N}\log\left|\Lambda\Lambda'+\Sigma_u\right| = \frac{1}{N}\log|\Sigma_u| + \mathcal{O}\left(\frac{\log N^{\beta}}{N}\right) \tag{35}$$

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

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

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:

$$\begin{aligned}\frac{1}{N}\text{tr}\left(S_x\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right) = \frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda_0\right) + \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)\\ - A_1 + A_2 + A_3 - A_4,\end{aligned} \tag{36}$$

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

Subsequently, we look at the terms $A_1 - A_4$, respectively.
Since $\pi_{\max}(\Sigma_u)$ and $\pi_{\min}^{-1}(\Lambda'\Lambda)$ are bounded from above uniformly in Ψ_δ, we can derive the following expressions similarly as in Bai and Liao (2016):

$$\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}\pi_{\max}\left[\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\right] \leq \sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}\frac{\pi_{\max}(\Sigma_u)}{\pi_{\min}(\Lambda'\Lambda)} = \mathcal{O}\left(N^{-\beta}\right) \tag{37}$$

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

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 (37) and (38), by the following:

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

Hence,

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

In the following, we define $s_i(A)$ as the ith singular value of a $(m \times n)$ matrix A. Using Lemma A.1., Fact 9.14.3 and Fact 9.14.23 in Bernstein (2009) and the fact that

$$\pi_{\max}\left[\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right] \leq \pi_{\max}\left[\left(\Lambda\Lambda'\right)^{-1}\right] = \mathcal{O}\left(N^{-\beta}\right),$$

we have:

$$\begin{aligned}
\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}|A_2| &\leq \frac{1}{N}\sum_{i=1}^{N} s_i\left(\frac{1}{T}\sum_{t=1}^{T}\Lambda_0 f_t u_t'\right) s_i\left(\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right) \\
&\leq \frac{1}{2N}\sum_{i=1}^{r} s_i\left(\Lambda_0'\Lambda_0+\frac{1}{T}\sum_{t=1}^{T} f_t u_t' u_t f_t'\right) s_i\left(\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right) \\
&\leq \frac{r}{2N} s_{\max}\left(\Lambda_0'\Lambda_0+\frac{1}{T}\sum_{t=1}^{T} f_t u_t' u_t f_t'\right) s_{\max}\left(\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right) \\
&\leq \frac{r}{2N}\left[s_{\max}\left(\Lambda_0'\Lambda_0\right)+s_{\max}\left(\frac{1}{T}\sum_{t=1}^{T} f_t u_t' u_t f_t'\right)\right] s_{\max}\left(\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right) \\
&= \frac{r}{2N}\left[\pi_{\max}^{1/2}\left(\Lambda_0'\Lambda_0\Lambda_0'\Lambda_0\right)+\pi_{\max}^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T} f_t u_t' u_t f_t' f_t u_t' u_t f_t'\right)\right] \\
&\quad\cdot \pi_{\max}^{1/2}\left(\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right) \\
&= \frac{r}{2N}\left(\left\|\Lambda_0'\Lambda_0\right\|+\left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t' u_t f_t'\right\|\right)\left\|\left(\Lambda\Lambda'+\Sigma_u\right)^{-1}\right\| \\
&\leq \frac{r}{2N}\left(\mathcal{O}(1)+\mathcal{O}\left(N^{-\beta}\right)\left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\right\|^2\right) \\
&\leq \frac{r}{2N}\left(\mathcal{O}(1)+\mathcal{O}\left(N^{-\beta}\right) N\cdot r\left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\right\|_{\max}^2\right) = \mathcal{O}_p\left(\frac{1}{N}+\frac{1}{N^\beta}\frac{\log N}{T}\right)
\end{aligned}$$

Similarly, we have that $\sup_{(\Lambda,\Sigma_u)\in\Psi_\delta}|A_3| = \mathcal{O}_p\left(\frac{1}{N}+\frac{1}{N^\beta}\frac{\log N}{T}\right)$.

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

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

From equations (37) and (38), we see that the second term on the right hand side is uniformly

of a smaller order than the first term. The first term of A_4 is bounded by:

$$\begin{aligned}
A_4 &= \text{tr}\left[\left(\Sigma_u^{-1}S_u\Sigma_u^{-1}\right)^{1/2}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\left(\Sigma_u^{-1}S_u\Sigma_u^{-1}\right)^{1/2}\right] \\
&\leq \text{tr}\left[\Sigma_u^{-1}S_u\Sigma_u^{-1}\right]\pi_{\max}\left(\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\right) \\
&\leq \text{tr}\left[\left(S_u\Sigma_u^{-1}\right)^{1/2}\Sigma_u^{-1}\left(S_u\Sigma_u^{-1}\right)^{1/2}\right]\mathcal{O}(1) \\
&\leq \text{tr}\left(S_u\Sigma_u^{-1}\right)\mathcal{O}(1)
\end{aligned}$$

Hence, we can bound the unpenalized log-likelihood function by:

$$\begin{aligned}
\mathcal{L}(\Lambda,\Sigma_u) &= \frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda_0\right)+\frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)+\frac{1}{N}\log|\Sigma_u| \\
&\quad -\frac{1}{N}\text{tr}\left(\Lambda_0'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0\right)+\mathcal{O}_p\left(\frac{\log N^\beta}{N}+\frac{1}{N^\beta}\frac{\log N}{T}\right) \\
&= \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)+\frac{1}{N}\log|\Sigma_u|+\frac{1}{N}\text{tr}\left[(\Lambda-\Lambda_0)'\Sigma_u^{-1}(\Lambda-\Lambda_0)\right] \\
&\quad -\frac{1}{N}\text{tr}\left[(\Lambda-\Lambda_0)'\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}(\Lambda-\Lambda_0)\right]+\mathcal{O}_p\left(\frac{\log N^\beta}{N}+\frac{1}{N^\beta}\frac{\log N}{T}\right) \\
&\leq \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)+\frac{1}{N}\log|\Sigma_u|+\frac{1}{N}\text{tr}\left[(\Lambda-\Lambda_0)'(\Lambda-\Lambda_0)\right]\pi_{\max}\left(\Sigma_u^{-1}\right) \\
&\quad -\frac{1}{N}\text{tr}\left[(\Lambda-\Lambda_0)'(\Lambda-\Lambda_0)\right]\pi_{\min}\left(\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right)+\mathcal{O}_p\left(\frac{\log N^\beta}{N}+\frac{1}{N^\beta}\frac{\log N}{T}\right) \\
&\leq \frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)+\frac{1}{N}\log|\Sigma_u|+\mathcal{O}_p\left(\frac{L_N}{N}\right)\max_{i\leq N}\|\lambda_i-\lambda_{i0}\|^2+\mathcal{O}\left(\frac{\log N^\beta}{N}+\frac{1}{N^\beta}\frac{\log N}{T}\right) \\
&= \frac{1}{N}\log|\Sigma_u|+\frac{1}{N}\text{tr}\left(S_u\Sigma_u^{-1}\right)+Q_2(\Lambda)+\mathcal{O}_p\left(\frac{\log N^\beta}{N}+\frac{1}{N^\beta}\frac{\log N}{T}\right)
\end{aligned}$$

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

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

□

Lemma A.3. *For* $d_T = \frac{\log N^\beta}{N} + \frac{1}{N^\beta}\frac{\log N}{T}$

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

Proof. If we consider equation (33) at the true parameter values, we get

$$\begin{aligned}
\mathcal{L}_p(\Lambda_0, \Sigma_{u0}) = & \frac{1}{N}\log\left|\Lambda_0\Lambda_0' + \Sigma_{u0}\right| + \frac{1}{N}\text{tr}\left(S_x\left(\Lambda_0\Lambda_0' + \Sigma_{u0}\right)^{-1}\right) \\
& - Q_2(\Lambda_0, \Sigma_{u0}) - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\text{tr}\left(S_u\Sigma_{u0}^{-1}\right) \\
& + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}| \\
= & \, Q_3(\Lambda_0, \Sigma_{u0})
\end{aligned} \tag{39}$$

Hence, by (32) and (39), we have

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

Therefore, by Lemma A.2. we have

$$Q_1(\hat{\Lambda}, \hat{\Sigma}_u) + Q_2(\hat{\Lambda}) \leq d_T, \tag{40}$$

□

Lemma A.4.

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

Proof. By equation (40) and the definition of $Q_1(\hat{\Lambda}, \hat{\Sigma}_u)$ and $Q_2(\hat{\Lambda})$, we get

$$B_1 + B_2 \leq d_T, \tag{41}$$

where B_1 and B_2 are defined as

$$B_1 = \frac{1}{N}\log\left|\hat{\Sigma}_u\right| + \frac{1}{N}\text{tr}\left(S_u\hat{\Sigma}_u^{-1}\right) - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\text{tr}\left(S_u\Sigma_{u0}^{-1}\right)$$
$$B_2 = \max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|^2 + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}\right| - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}|$$

By equation (41), we can see that

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

and

$$\frac{1}{N}\log\left|\hat{\Phi}_u\right| + \frac{1}{N}\text{tr}\left(S_u\hat{\Phi}_u^{-1}\right) - \frac{1}{N}\log|\Phi_{u0}| - \frac{1}{N}\text{tr}\left(S_u\Phi_{u0}^{-1}\right) \leq d_T, \tag{42}$$

where $\Phi_u = \text{diag}\left(\Sigma_u\right)$ and denotes a covariance matrix that contains only the elements of the main diagonal of Σ_u. Using the same argument as in the proof of *Lemma B.1.* in Bai and Liao (2016), we get

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

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

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

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

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

□

Lemma A.5.

$$\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\| = \mathcal{O}_p\left(\mu + \sqrt{\frac{Nd_T}{L_N}}\right)$$

Proof. If we consider equation (41) and Lemma A.4., we have

$$\begin{aligned}\mathcal{O}\left(\frac{L_N}{N}\right)\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|^2 + \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik}\right| - |\lambda_{ik0}| &\leq d_T \\ \mathcal{O}\left(\frac{L_N}{N}\right)\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|^2 - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}|\lambda_{ik0}| - \left|\hat{\lambda}_{ik}\right| &\leq d_T \\ \mathcal{O}\left(\frac{L_N}{N}\right)\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|^2 - \frac{1}{N}\mu\sum_{k=1}^{r}\sum_{i=1}^{N}\left|\hat{\lambda}_{ik} - \lambda_{ik0}\right| &\leq d_T \\ \mathcal{O}\left(\frac{L_N}{N}\right)\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|^2 - \mathcal{O}\left(\frac{L_N}{N}\right)\mu\max_{i\leq N}\sum_{k=1}^{r}\left|\hat{\lambda}_{ik} - \lambda_{ik0}\right| &\leq d_T \\ \mathcal{O}\left(\frac{L_N}{N}\right)\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|^2 - \mathcal{O}\left(\frac{L_N}{N}\right)\mu\sqrt{r}\sqrt{\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|^2} &\leq d_T\end{aligned}$$

Solving for $\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\|$ yields

$$\begin{aligned}\max_{i\leq N}\left\|\hat{\lambda}_i - \lambda_{i0}\right\| &\leq \mu + \sqrt{\mu^2 + \mathcal{O}\left(\frac{Nd_T}{L_N}\right)} \\ &\leq \mu + \mathcal{O}\left(\sqrt{\frac{Nd_T}{L_N}}\right)\end{aligned}$$

□

Lemma A.6.

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_t - f_t\right\|^2 = o_p(1)$$

Proof. By the definition of the factor estimator in equation (5) we have:

$$\hat{f}_t - f_t = -\left(\hat{\Lambda}'\hat{\Phi}_u^{-1}\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Phi}_u^{-1}\left(\hat{\Lambda}-\Lambda\right)f_t + \left(\hat{\Lambda}'\hat{\Phi}_u^{-1}\hat{\Lambda}\right)^{-1}\hat{\Lambda}'\hat{\Phi}_u^{-1}u_t \tag{43}$$

As $L_N = \mathcal{O}\left(N^{\beta}\right)$, the first term on the right-hand side can be bounded by:

$$\begin{aligned}
&\mathcal{O}_p\left(N^{-\beta}\right)\sqrt{\sum_{i=1}^{N}\left\|\left(\hat{\Lambda}'\hat{\Phi}_u^{-1}\right)_i\left(\hat{\lambda}_i-\lambda_i\right)\right\|^2}\,\|f_t\| \\
&\leq \mathcal{O}_p\left(N^{-\beta}\right)\sqrt{\mathcal{O}_p\left(\sum_{i=1}^{N}\left\|\hat{\lambda}_i-\lambda_i\right\|^2\right)} \\
&\leq \mathcal{O}_p\left(N^{-\beta}\right)\sqrt{\mathcal{O}_p\left(L_N\max_{i\leq N}\left\|\hat{\lambda}_i-\lambda_i\right\|^2\right)} \\
&= \mathcal{O}_p\left(\frac{\sqrt{L_N}}{N^{\beta}}\right)o_p(1) = o_p(1)
\end{aligned} \tag{44}$$

Now, we are going to bound the second term on the right-hand side of (43). For this we first analyse the term $\hat{\Lambda}'\hat{\Phi}_u^{-1}u_t$.

$$\begin{aligned}
&\mathcal{O}_p\left(N^{-\beta}\right)\left\|\left(\hat{\Lambda}'\hat{\Phi}_u^{-1}-\Lambda_0'\Phi_{u0}^{-1}\right)u_t\right\|_F \leq \\
&\quad\mathcal{O}_p\left(N^{-\beta}\right)\left\|\left(\hat{\Lambda}-\Lambda_0\right)'\hat{\Phi}_u^{-1}u_t\right\|_F + \mathcal{O}_p\left(N^{-\beta}\right)\left\|\Lambda_0'\left(\hat{\Phi}_u^{-1}-\Phi_{u0}^{-1}\right)u_t\right\|_F
\end{aligned}$$

Using Lemma A.5., the first term can be bounded by:

$$\begin{aligned}
&\mathcal{O}_p\left(N^{-\beta}\right)\sqrt{\sum_{i=1}^{N}\left\|\left(\hat{\lambda}_i-\lambda_{i0}\right)\left(\hat{\Phi}_u^{-1}u_t\right)_i\right\|^2} \\
&\quad\mathcal{O}_p\left(N^{-\beta}\right)\sqrt{L_N\max_{i\leq N}\left\|\hat{\lambda}_i-\lambda_{i0}\right\|^2\mathcal{O}_p(1)} \\
&= \mathcal{O}_p\left(\frac{\sqrt{L_N}}{N^{\beta}}\right)o_p(1) = o_p(1)
\end{aligned} \tag{45}$$

The second term can be bounded using Lemma A.4. according to:

$$\begin{aligned}\mathcal{O}_p\left(N^{-\beta}\right)\left\|\Lambda_0'\left(\hat{\Phi}_u^{-1}-\Phi_{u0}^{-1}\right)u_t\right\|_F &= \mathcal{O}_p\left(N^{-\beta}\right)\sqrt{\sum_{i=1}^{N}\left\|\left(\Lambda_0'\Phi_u^{-1}\right)_i\left(\phi_{iu0}-\hat{\phi}_{iu}\right)\left(\hat{\Phi}_u^{-1}u_t\right)_i\right\|^2}\\ &\leq \mathcal{O}_p\left(N^{-\beta}\right)\sqrt{\sum_{i=1}^{N}\left\|\hat{\phi}_{iu}-(\phi_{iu0})\right\|^2\left\|\left(\Lambda_0'\Phi_u^{-1}\right)_i\right\|^2\left\|\left(\hat{\Phi}_u^{-1}u_t\right)_i\right\|^2}\\ &= \mathcal{O}_p\left(\frac{\log N}{N^{\beta}}\left\|\hat{\Phi}_u-\Phi_{u0}\right\|_F\right)=o_p(1) \qquad (46)\end{aligned}$$

Hence, using (44), (45) and (46) yields:

$$\left\|\hat{f}_t-f_t\right\|=\mathcal{O}_p\left(N^{-\beta}\right)\sum_{i=1}^{N}\left\|\left(\Lambda_0'\Phi_{u0}^{-1}\right)_i u_{it}\right\|+o_p(1)=\mathcal{O}_p\left(N^{-\beta/2}\right)+o_p(1)=o_p(1)$$

□

Lemma A.7.

$$\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}|\hat{u}_{it}-u_{it}|^2=\mathcal{O}_p\left(\mu^2+\frac{Nd_T}{L_N}\right)$$

Proof. Since $\hat{u}_{it}-u_{it}=\left(\hat{\lambda}_i-\lambda_i\right)\hat{f}_t'+\lambda_i\left(\hat{f}_t-f_t\right)'$, we have by Lemma A.5. and Lemma A.6.:

$$\begin{aligned}\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}|\hat{u}_{it}-u_{it}|^2 &\leq 2\max_{i\leq N}\left\|\hat{\lambda}_i-\lambda_i\right\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_t\right\|^2+2\max_{i\leq N}\|\lambda_i\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_t-f_t\right\|^2\\ &\leq \mathcal{O}_p\left(\max_{i\leq N}\left\|\hat{\lambda}_i-\lambda_i\right\|^2\right)+\mathcal{O}_p\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_t-f_t\right\|^2\right)\\ &= \mathcal{O}_p\left(\mu^2+\frac{Nd_T}{L_N}\right)\end{aligned}$$

□

Lemma A.8.

$$\max_{i,j\leq N}\left|\hat{\sigma}_{ij}-\sigma_{ij}\right|=\mathcal{O}_p\left(\sqrt{\mu^2+\frac{Nd_T}{L_N}}\right),$$

where $d_T=\frac{\log N^{\beta}}{N}+\frac{1}{N^{\beta}}\frac{\log N}{T}$.

Proof. Based on *Lemma A.3.(iii)* by Fan, Liao, and Mincheva (2011) we have:

$$\max_{i,j\leq N}\left|\hat{\sigma}_{ij}-\sigma_{ij}\right| \leq \max_{i,j\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-\sigma_{ij}\right| + \max_{i,j\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}-u_{it}u_{jt}\right|, \tag{47}$$

where the authors show that the first term on the right-hand side is $\mathcal{O}_p\left(\sqrt{\frac{\log N}{T}}\right)$. Now we are going to analyse the second term on the right-hand side of equation (47). In Lemma A.7. we have shown that $\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left|\hat{u}_{it}-u_{it}\right|^2 = o_p(1)$. Hence, the result follows from Lemma A.3.(ii) by Fan, Liao, and Mincheva (2011).

□

A.2 Rate of convergence for the idiosyncratic error covariance matrix estimator

In what follows, we are going to determine the convergence rate of the idiosyncratic error covariance matrix estimator based on soft-thresholding.

Lemma A.9.

$$\left\|\hat{\Sigma}_u^{\tau}-\Sigma_u\right\| = \mathcal{O}_p\left(S_N\sqrt{\mu^2+\frac{Nd_T}{L_N}}\right)$$

Proof. The result follows from Lemma A.8. and Theorem A.1. of Fan, Liao, and Mincheva (2013).

□

A.3 Convergence Rates for the Covariance Matrix Estimator

Proof: Theorem 3.2 (Convergence Rates for the Covariance Matrix Estimator)

$$\Sigma = \Lambda_0 \Lambda_0' + \Sigma_{u0} \tag{48}$$

$$\hat{\Sigma}_{\text{SAF}} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Sigma}_u^{\tau}, \tag{49}$$

where $\hat{\Sigma}_u^{\tau}$ corresponds to the POET estimator of Fan, Liao, and Mincheva (2013). Similar as in Fan, Liao, and Mincheva (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.10.

$$\frac{1}{N} \left\| \hat{\Sigma}_{SAF} - \Sigma \right\|_{\Sigma}^2 = \mathcal{O}_p \left(\frac{L_N^2}{N^2} \left[\mu^4 + \left(\frac{N}{L_N} d_T \right)^2 \right] + \left[\frac{N^{\beta} L_N}{N^2} + \frac{S_N^2}{N} \right] \left[\mu^2 + \frac{N}{L_N} d_T \right] \right)$$

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

$$\left\| \hat{\Sigma}_{\text{SAF}} - \Sigma \right\|_{\Sigma}^2 \leq \left\| \hat{\Lambda}\hat{\Lambda}' - \Lambda_0 \Lambda_0' \right\|_{\Sigma}^2 + \left\| \hat{\Sigma}_u^{\tau} - \Sigma_{u0} \right\|_{\Sigma}^2 \tag{50}$$

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

$$\begin{aligned} CC' &= \hat{\Lambda}\hat{\Lambda}' - \hat{\Lambda}\Lambda_0' - \Lambda_0\hat{\Lambda}' + \Lambda_0\Lambda_0' \\ \Lambda_0 C' &= \Lambda_0\hat{\Lambda}' - \Lambda_0\Lambda_0' \\ C\Lambda_0' &= \hat{\Lambda}\Lambda_0' - \Lambda_0\Lambda_0' \end{aligned}$$

Using the previous definitions, we can rewrite the first term in (50) in the following form

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

Hence, equation (50) can be expressed as:

$$\left\|\hat{\Sigma}_{\text{SAF}} - \Sigma\right\|_{\Sigma}^2 \leq \left\|CC'\right\|_{\Sigma}^2 + \left\|C\Lambda_0'\right\|_{\Sigma}^2 + \left\|\Lambda_0 C'\right\|_{\Sigma}^2 + \left\|\hat{\Sigma}_u^{\tau} - \Sigma_u\right\|_{\Sigma}^2 \tag{51}$$

Now we analyse each term in (51) separately:

$$\begin{aligned}
\left\|\Lambda_0 C'\right\|_{\Sigma}^2 &= N^{-1}\text{tr}\left(\Sigma^{-1/2}\Lambda_0 C'\Sigma^{-1/2}\Sigma^{-1/2}C\Lambda_0'\Sigma^{-1/2}\right) \\
&= N^{-1}\text{tr}\left(\Lambda_0'\Sigma^{-1}\Lambda_0 C'\Sigma^{-1}C\right) \\
&\leq N^{-1}\left\|\Lambda_0'\Sigma^{-1}\Lambda_0\right\|\left\|\Sigma^{-1}\right\|\left\|C\right\|_F^2 = \mathcal{O}_p\left(\frac{N^{\beta}}{N}\left\|C\right\|_F^2\right)
\end{aligned}$$

Similarly, we get $\left\|C\Lambda_0'\right\|_{\Sigma}^2 = \mathcal{O}_p\left(\frac{N^{\beta}}{N}\left\|C\right\|_F^2\right)$. Further, $\left\|CC'\right\|_{\Sigma}^2 = \frac{1}{N}\left\|C\right\|_F^4$.
Hence, by Lemma A.9. we get:

$$\begin{aligned}
\left\|\hat{\Sigma}_{\text{SAF}} - \Sigma\right\|_{\Sigma}^2 &= \mathcal{O}_p\left(\frac{1}{N}\left\|C\right\|_F^4 + \frac{N^{\beta}}{N}\left\|C\right\|_F^2\right) + \mathcal{O}_p\left(\left\|\hat{\Sigma}_u^{\tau} - \Sigma_u\right\|_{\Sigma}^2\right) \\
&= \mathcal{O}_p\left(\frac{L_N^2}{N}\left[\mu^4 + \left(\frac{N}{L_N}d_T\right)^2\right] + \frac{N^{\beta}L_N}{N}\left[\mu^2 + \frac{N}{L_N}d_T\right]\right) + \mathcal{O}_p\left(S_N^2\left[\mu^2 + \frac{N}{L_N}d_T\right]\right) \\
&= \mathcal{O}_p\left(\frac{L_N^2}{N}\left[\mu^4 + \left(\frac{N}{L_N}d_T\right)^2\right] + \left[\frac{N^{\beta}L_N}{N} + S_N^2\right]\left[\mu^2 + \frac{N}{L_N}d_T\right]\right)
\end{aligned}$$

□

Under the **Frobenius norm** we have:

Lemma A.11.

$$\frac{1}{N}\left\|\hat{\Sigma}_{SAF} - \Sigma\right\|_F^2 = \mathcal{O}_p\left(\frac{L_N^2}{N}\left[\mu^2 + \frac{N}{L_N}d_T\right]^2 + \left[\frac{N^{\beta}L_N}{N} + S_N^2\right]\left[\mu^2 + \frac{N}{L_N}d_T\right]\right)$$

Proof. A similar argument as in Lemma A.10 leads to:

$$\left\|\hat{\Sigma}_{\text{SAF}} - \Sigma\right\|_F^2 \leq \left\|CC'\right\|_F^2 + \left\|\Lambda_0 C'\right\|_F^2 + \left\|C\Lambda_0'\right\|_F^2 + \left\|\hat{\Sigma}_u^{\tau} - \Sigma_u\right\|_F^2, \tag{52}$$

where the second term can be bounded by

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

Furthermore, the first term in (52) has the same upper bound. Hence, again by using Lemma A.9 we get:

$$\begin{aligned}\left\|\hat{\Sigma}_{\text{SAF}} - \Sigma\right\|_F^2 &\leq \mathcal{O}_p\left(\|C\|_F^4 + N^\beta \|C\|_F^2\right) + \mathcal{O}_p\left(\left\|\hat{\Sigma}_u^\tau - \Sigma_u\right\|_F^2\right) \\ &\leq \mathcal{O}_p\left(L_N^2\left[\mu^2 + \frac{N}{L_N}d_T\right]^2 + N^\beta L_N\left[\mu^2 + \frac{N}{L_N}d_T\right]\right) + \mathcal{O}_p\left(N\left[\mu^2 + \frac{N}{L_N}d_T\right]S_N^2\right) \\ &= \mathcal{O}_p\left(L_N^2\left[\mu^2 + \frac{N}{L_N}d_T\right]^2 + \left[N^\beta L_N + NS_N^2\right]\left[\mu^2 + \frac{N}{L_N}d_T\right]\right)\end{aligned}$$

□

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.12.

(i) $\left\|\hat{G}\right\| = \mathcal{O}_p\left(N^{-\beta}\right)$

(ii) $\left\|\hat{G}^{-1} - G_0^{-1}\right\|_F = \mathcal{O}_p\left(N^\beta\left(N^{-\beta/2}\|C\|_F + \left\|\left(\hat{\Sigma}_u^\tau\right)^{-1} - \Sigma_u^{-1}\right\|_F\right)\right)$

Proof.

(i) Lemma A.9. implies $\left\| \left(\hat{\Sigma}_u^\tau \right)^{-1} \right\| = \mathcal{O}_p(1)$. Then, by the definition of $\hat{G}$ we have:

$$\begin{aligned} \left\| \hat{G} \right\| &\leq \left\| \left(\hat{\Lambda}' \left(\hat{\Sigma}_u^\tau \right)^{-1} \hat{\Lambda} \right)^{-1} \right\| \\ &\leq \frac{\pi_{\max} \left(\hat{\Sigma}_u^\tau \right)}{\pi_{\min} \left(\hat{\Lambda}' \hat{\Lambda} \right)} = \mathcal{O}_p \left(N^{-\beta} \right) \end{aligned}$$

(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:

$$\hat{G}^{-1} - G_0^{-1} = C' \left(\hat{\Sigma}_u^\tau \right)^{-1} \hat{\Lambda} + \Lambda_0' \Sigma_{u0}^{-1} C + \Lambda_0' \left(\left(\hat{\Sigma}_u^\tau \right)^{-1} - \Sigma_{u0}^{-1} \right) \hat{\Lambda} \tag{53}$$

If we bound all three terms on the right hand side of equation (53), we get:

$$\begin{aligned} \left\| \hat{G}^{-1} - G_0^{-1} \right\|_F &\leq \|C\|_F \, \mathcal{O}_p \left(N^{\beta/2} \right) + \left\| \left(\hat{\Sigma}_u^\tau \right)^{-1} - \Sigma_{u0}^{-1} \right\|_F \mathcal{O}_p \left(N^{\beta} \right) \\ &= \mathcal{O}_p \left(N^{\beta} \left(N^{-\beta/2} \|C\|_F + \left\| \left(\hat{\Sigma}_u^\tau \right)^{-1} - \Sigma_u^{-1} \right\|_F \right) \right) \end{aligned}$$

□

Lemma A.13.

$$\frac{1}{N} \left\| \hat{\Sigma}_{SAF}^{-1} - \Sigma^{-1} \right\|_F^2 = \mathcal{O}_p \left(\frac{L_N}{N^{\beta+1}} \left[\mu^2 + \frac{N}{L_N} d_T \right] + S_N^2 \left[\mu^2 + \frac{N}{L_N} d_T \right] \right)$$

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

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

where

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

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

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

By Lemma A.12. *(i)* follows that $L_2 \leq \mathcal{O}_p(L_1)$. Similarly, L_3 is also $\mathcal{O}_p(L_1)$. Further,

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

Hence, also by Lemma A.12. *(i)*

$$
L_4 \leq \|C\|_F^2 \, \mathcal{O}_p\left(N^{-\beta}\right) = \mathcal{O}_p\left(\|C\|_F^2 \, N^{-\beta}\right)
$$

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

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

By Lemma A.12. *(ii)* we have,

$$\begin{aligned}
L_6 &\leq \mathcal{O}_p\left(N^{2\beta}\right)\left\|\hat{G}\left(G_0^{-1}-\hat{G}^{-1}\right)G_0\right\|_F^2 \\
&\leq \mathcal{O}_p\left(N^{-2\beta}\right)\left\|G_0^{-1}-\hat{G}^{-1}\right\|_F^2 \\
&= \mathcal{O}_p\left(N^{-2\beta}\right)\mathcal{O}_p\left(N^{2\beta}\left(N^{-\beta}\left\|C\right\|_F^2+\left\|\left(\hat{\Sigma}_u^{\tau}\right)^{-1}-\Sigma_u^{-1}\right\|_F^2\right)\right) \\
&= \mathcal{O}_p\left(N^{-\beta}\left\|C\right\|_F^2+\left\|\left(\hat{\Sigma}_u^{\tau}\right)^{-1}-\Sigma_u^{-1}\right\|_F^2\right)
\end{aligned}$$

Adding up the terms $L_1 - L_6$ gives

$$\frac{1}{N}\left\|\hat{\Sigma}_{\text{SAF}}^{-1}-\Sigma^{-1}\right\|_F^2 = \mathcal{O}_p\left(\frac{L_N}{N^{\beta+1}}\left[\mu^2+\frac{N}{L_N}d_T\right]+S_N^2\left[\mu^2+\frac{N}{L_N}d_T\right]\right)$$

□

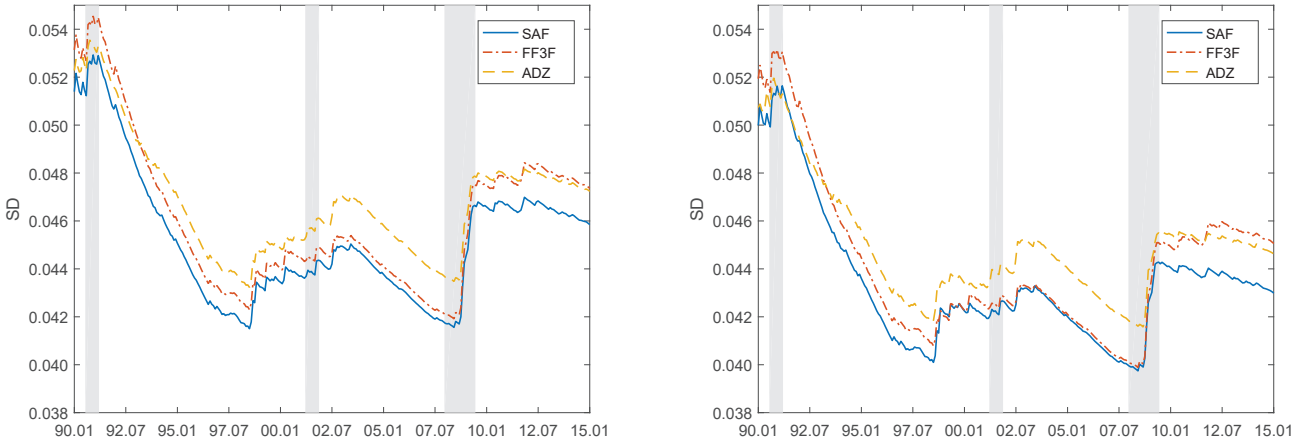
B Tables

Table 4: Summary Statistics for the Estimated Portfolio Weights

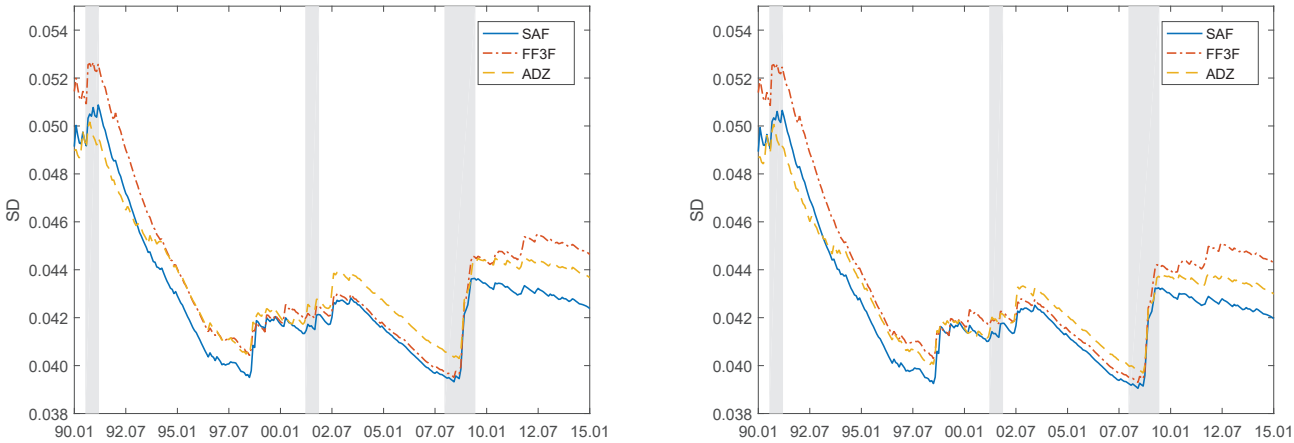
Model	1/N	GMVP	**SAF**	AFM	DFM	SIM	FF3F	LW	KDM	ADZ
					N = 30					
Min	0.0333	-1.0988	-0.1907	-0.5123	-0.4828	0.0295	-0.0669	-0.5176	-0.6166	-0.1776
Max	0.0333	1.0253	0.1996	0.4584	0.3353	0.0448	0.1121	0.3198	0.5534	0.3392
SD	0.0000	0.1366	0.0197	0.0670	0.0543	0.0010	0.0088	0.0648	0.0249	0.0500
MAD	0.0000	0.1075	0.0157	0.0522	0.0421	0.0007	0.0069	0.0506	0.0196	0.0397
					N = 50					
Min	0.0200	-2.7950	-0.1461	-0.4305	-0.4127	0.0176	-0.0701	-0.4237	-1.0228	-0.1492
Max	0.0200	2.7290	0.1634	0.3798	0.2158	0.0270	0.0884	0.2629	1.0182	0.3004
SD	0.0000	0.2255	0.0158	0.0505	0.0394	0.0006	0.0073	0.0513	0.0214	0.0400
MAD	0.0000	0.1777	0.0126	0.0394	0.0307	0.0004	0.0058	0.0399	0.0169	0.0318
					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: Summary statistics for the estimated portfolio weights for our sparse approximate factor model (SAF) are compared to the equally weighted portfolio ($1/N$), the GMVP plug-in estimator (GMVP), the approximate factor model (AFM), the dynamic factor model (DFM), the single factor model by Sharpe (1963) (SIM), the 3 factor model by Fama and French (1993) (FF3F), the estimators by Ledoit and Wolf (2003) (LW), Kourtis, Dotsis, and Markellos (2012) (KDM), Abadir, Distaso, and Žikeš (2014) (ADZ).

C Figures



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



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

Figure 4: SD for different subperiods