Journal of Econometrics 201 (2017) 384–399

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Using principal component analysis to estimate a high dimensional factor model with high-frequency data☆

Yacine Aït-Sahalia [a,*], Dacheng Xiu [b]

[a] Department of Economics, Princeton University and NBER, 26 Prospect Avenue, Princeton, NJ 08540, USA
[b] Booth School of Business, University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Available online 19 August 2017

JEL classification:
C13
C14
C55
C58
G01

Keywords:
High-dimensional data
High-frequency data
Latent factor model
Principal components
Portfolio optimization

ABSTRACT

This paper constructs an estimator for the number of common factors in a setting where both the sampling frequency and the number of variables increase. Empirically, we document that the covariance matrix of a large portfolio of US equities is well represented by a low rank common structure with sparse residual matrix. When employed for out-of-sample portfolio allocation, the proposed estimator largely outperforms the sample covariance estimator.

1. Introduction

This paper proposes an estimator, using high frequency data, for the number of common factors in a large-dimensional dataset. The estimator relies on principal component analysis (PCA) and novel joint asymptotics where both the sampling frequency and the dimension of the covariance matrix increase. One by-product of the estimation method is a well-behaved estimator of the increasingly large covariance matrix itself, including a split between its systematic and idiosyncratic matrix components.

Principal component analysis (PCA) and factor models represent two of the main methods at our disposal to estimate large covariance matrices. If nonparametric PCA determines that a common structure is present, then a parametric or semiparametric factor model becomes a natural choice to represent the data. Prominent examples of this approach include the arbitrage pricing theory (APT) of Ross (1976) and the intertemporal capital asset pricing model (ICAPM) of Merton (1973), which provide an economic rationale for the presence of a factor structure in asset returns. Chamberlain and Rothschild (1983) extend the APT strict factor model to an approximate factor model, in which the residual covariances are not necessarily diagonal, hence allowing for comovement that is unrelated to the systematic risk factors. Based on this model, Connor and Korajczyk (1993), Bai and Ng (2002), Amengual and Watson (2007), Onatski (2010) and Kapetanios (2010) propose statistical methodologies to determine the number of factors, while Bai (2003) provides tools to conduct statistical inference on the common factors and their loadings. Connor and Korajczyk (1988) use PCA to test the APT.

In parallel, much effort has been devoted to searching for observable empirical proxies for the latent factors. The three-factor model by Fama and French (1993) and its many extensions are widely used examples, with factors constructed using portfolios returns often formed by sorting firm characteristics. Chen et al. (1986) propose macroeconomic variables as factors, including inflation, output growth gap, interest rate, risk premia, and term premia. Estimators of the covariance matrix based on observable factors are proposed by Fan et al. (2008) in the case of a strict factor model and Fan et al. (2011) in the case of an approximate factor model. A factor model can serve as the reference point for shrinkage estimation (see Ledoit and Wolf (2012) and Ledoit and

☆ We are benefited from the very helpful comments of the Editor and two anonymous referees, as well as extensive discussions with Jianqing Fan, Alex Furger, Chris Hansen, Jean Jacod, Yuan Liao, Nour Meddahi, Markus Pelger, and Weichen Wang, as well as seminar and conference participants at CEMFI, Duke University, the 6th French Econometrics Conference in Honor of Christian Gouriéroux, the 8th Annual SoFiE Conference, the 2015 IMS-China International Conference on Statistics and Probability, and the 11th World Congress of the Econometric Society. We are also grateful to Chaoxing Dai for excellent research assistance.

* Corresponding author.
E-mail addresses: yacine@princeton.edu (Y. Aït-Sahalia), dacheng.xiu@chicagobooth.edu (D. Xiu).

http://dx.doi.org/10.1016/j.jeconom.2017.08.015
0304-4076/

Wolf (2004)). Alternative methods rely on various forms of thresholding (Bickel and Levina, 2008a,b; Cai and Liu, 2011; Fryzlewicz, 2013; Zhou et al., 2014) whereas the estimator in Fan et al. (2013) is designed for latent factor models.

The above factor models are static, as opposed to the dynamic factor models introduced in Gouriéroux and Jasiak (2001) to represent stochastic means and volatilities, extreme risks, liquidity and moral hazard in insurance analysis. Dynamic factor models are developed in Forni et al. (2000), Forni and Lippi (2001), Forni et al. (2004) and Doz et al. (2011), in which the lagged values of the unobserved factors may also affect the observed dependent variables; see Croux et al. (2004) for a discussion. Forni et al. (2009) adapt structural vector autoregression analysis to dynamic factor models.

Both static and dynamic factor models in the literature have typically been cast in discrete time. By contrast, this paper provides methods to estimate continuous-time factor models, where the observed variables are continuous Itô semimartingales. The literature dealing with continuous-time factor models has mainly focused on models with observable explanatory variables in a low dimensional setting. For example, Mykland and Zhang (2006) develop tools to conduct analysis of variance as well as univariate regression, while Todorov and Bollerslev (2010) add a jump component in the univariate regression setting and Aït-Sahalia et al. (2014) extend the factor model further to allow for multivariate regressors and time-varying coefficients.

When the factors are latent, however, PCA becomes the main tool at our disposal. Aït-Sahalia and Xiu (2015) extend PCA from its discrete-time low frequency roots to the setting of general continuous-time models sampled at high frequency. The present paper complements it by using PCA to construct estimators for the number of common factors, and exploiting the factor structure to build estimators of the covariance matrix in an increasing dimension setting, without requiring that a set of observable common factors be pre-specified. The analysis is based on a general continuous-time semiparametric approximate factor model, which allows for stochastic variation in volatilities as well as correlations. Independently, Pelger (2015a, b) propose an alternative estimator for the number of factors and factor loadings, with a distributional theory that is entry-wise, whereas the present paper concentrates on the matrix-wise asymptotic properties of the covariance matrix and its inverse.

This paper shares some theoretical insights with the existing literature of approximate factor models in discrete time, in terms of the strategy for estimating the number of factors. However, there are several distinctions, which require a different treatment in our setting. For instance, the identification restrictions we impose differ from those given by e.g., Bai (2003), Doz et al. (2011) and Fan et al. (2013), due to the prevalent presence of heteroscedasticity in high frequency data. Also, the discrete-time literature on determining the number of factors relies on random matrix theory for i.i.d. data (see, e.g., Bai and Ng, 2002; Onatski, 2010; Ahn and Horenstein, 2013; Trapani, 2017), which is not available for semimartingales.

The methods in this paper, including the focus on the inverse of the covariance matrix, can be useful in the context of portfolio optimization when the investable universe consists of a large number of assets. For example, in the Markowitz model of mean–variance optimization, an unconstrained covariance matrix with d assets necessitates the estimation of $d(d+1)/2$ elements, which quickly becomes unmanageable as d grows, and even if feasible would often result in optimal asset allocation weights that have undesirable properties, such as extreme long and short positions. Various approaches have been proposed in the literature to deal with this problem. The first approach consists in imposing some further structure on the covariance matrix to reduce the number of parameters to be estimated, typically in the form of a factor model along the lines discussed above, although Green and Hollifield (1992) argue that the dominance of a single factor in equity returns can lead empirically to extreme portfolio weights. The second approach consists in imposing constraints on the portfolio weights (Jagannathan and Ma, 2003; Pesaran and Zaffaroni, 2008; DeMiguel et al., 2009a; El Karoui, 2010; Fan et al., 2012; Gandy and Veraart, 2013) or penalties Brodie et al. (2009). The third set of approaches are Bayesian and consist in shrinkage of the covariance estimates (Ledoit and Wolf, 2003), assuming a prior distribution for expected returns and covariances and reformulating the Markowitz problem as a stochastic optimization one (Lai et al., 2011), or simulating to select among competing models of predictable returns and maximize expected utility (Jacquier and Polson, 2010). A fourth approach consists in modeling directly the portfolio weights in the spirit of Aït-Sahalia and Brandt (2001) as a function of the asset's characteristics (Brandt et al., 2009). A fifth and final approach consists in abandoning mean–variance optimization altogether and replacing it with a simple equally-weighted portfolio, which may in fact outperform the Markowitz solution in practice (DeMiguel et al., 2009b).

An alternative approach to estimating covariance matrices using high-frequency data is fully nonparametric, i.e., without assuming any underlying factor structure, strict or approximate, latent or not. Two issues have attracted much attention in this part of the literature, namely the potential presence of market microstructure noise in high frequency observations and the potential asynchronicity of the observations: see Aït-Sahalia and Jacod (2014) for an introduction. Various methods are available, including Hayashi and Yoshida (2005), Aït-Sahalia et al. (2010), Christensen et al. (2010), Barndorff-Nielsen et al. (2011), Zhang (2011), Shephard and Xiu (2012) and Bibinger et al. (2014). However, when the dimension of the asset universe increases to a few hundreds, the number of synchronized observations is bound to drop, which requires severe downsampling and hence much longer time series to be maintained. Dealing with an increased dimensionality without a factor structure typically requires the additional assumption that the population covariance matrix itself is sparse (see, e.g., Tao et al., 2011, 2013b, a). Fan et al. (2016) assume a factor model but with factors that are observable.

The rest of the paper is organized as follows. Section 2 sets up the model and assumptions. Section 3 describes the proposed estimators and their properties. We show that both the factor-driven and the residual components of the sample covariance matrix are identifiable, as the cross-sectional dimension increases. The proposed PCA-based estimator is consistent, invertible and well-conditioned. Additionally, based on the eigenvalues of the sample covariance matrix, we provide a new estimator for the number of latent factors. Section 4 provides Monte Carlo simulation evidence.

Section 5 implements the estimator on a large portfolio of stocks. We find a clear block-diagonal pattern in the residual correlations of equity returns, after sorting the stocks by their firms' global industrial classification standard (GICS) codes. This suggests that the covariance matrix can be approximated by a low-rank component representing exposure to some common factors, plus a sparse component, which reflects their sector/industry specific exposure. Empirically, we find that the factors uncovered by PCA explain a larger fraction of the total variation of asset returns than that explained by observable portfolio factors such as the market portfolio, the Fama–French portfolios, as well as the industry-specific ETF portfolios. Also, the residual covariance matrix based on PCA is sparser than that based on observable factors, with both exhibiting a clear block-diagonal pattern. Finally, we find that the PCA-based estimator outperforms the sample covariance estimator in out-of-sample portfolio allocation. Section 6 concludes. Mathematical proofs are in the appendix.

2. Factor model setup

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ be a filtered probability space. Let $\mathcal{M}_{d\times r}$ be the Euclidean space of $d \times r$ matrices. Throughout the paper, we use $\lambda_j(A)$, $\lambda_{\min}(A)$, and $\lambda_{\max}(A)$ to denote the jth, the minimum, and the maximum eigenvalues of a matrix A. In addition, we use $\|A\|_1$, $\|A\|_\infty$, $\|A\|$, and $\|A\|_{\mathrm{F}}$ to denote the $\mathbb{L}_1$ norm, the $\mathbb{L}_\infty$ norm, the operator norm (or $\mathbb{L}_2$ norm), and the Frobenius norm of a matrix A, that is, $\max_j\sum_i|A_{ij}|$, $\max_i\sum_j|A_{ij}|$, $\sqrt{\lambda_{\max}(A^\intercal A)}$, and $\sqrt{\mathrm{Tr}(A^\intercal A)}$, respectively. When A is a vector, both $\|A\|$ and $\|A\|_{\mathrm{F}}$ are equal to its Euclidean norm. We also use $\|A\|_{\mathrm{MAX}} = \max_{i,j}|A_{ij}|$ to denote the $\mathbb{L}_\infty$ norm of A on the vector space. We use e_i to denote a d-dimensional column vector whose ith entry is 1 and 0 elsewhere. K is a generic constant that may change from line to line.

We observe a large intraday panel of asset log-prices, Y on a time interval $[0, T]$ at instants $0, \Delta_n, 2\Delta_n, \ldots, n\Delta_n$, where Δ_n is the sampling frequency and $n = [T/\Delta_n]$. We assume that Y follows a continuous-time factor model,

$$Y_t = \beta X_t + Z_t, \tag{1}$$

where Y_t is a d-dimensional vector process, X_t is a r -dimensional unobservable common factor process, Z_t is the idiosyncratic component, and β is a constant factor loading matrix of size $d \times r$. The constant β assumption, although restrictive, is far from unusual in the literature. In fact, Reiß et al. (forthcoming) find evidence supportive of this assumption using high-frequency data.

The asymptotic framework we employ is one where the time horizon T is fixed (at 1 month in the empirical analysis), the number of factors r is unknown but finite, whereas the cross-sectional dimension d increases to ∞ as the sampling interval Δ_n goes to 0.

To complete the specification, we make additional assumptions on the respective dynamics of the factors and the idiosyncratic components.

Assumption 1. Assume that the common factor X and idiosyncratic component Z are continuous Itô semimartingales, that is,

$$X_t = \int_0^t h_s\,ds + \int_0^t \eta_s dW_s, \quad Z_t = \int_0^t f_s ds + \int_0^t \gamma_s dB_s. \tag{2}$$

We denote the spot covariance of X_t as $e_t = \eta_t\eta_t^\intercal$, and that of Z_t as $g_t = \gamma_t\gamma_t^\intercal$. W_t and B_t are independent Brownian motions. In addition, h_t and f_t are progressively measurable, the process η_t, γ_t are càdlàg, and e_t, e_{t-}, g_t, and g_{t-} are positive-definite. Finally, for all $1 \le i, j \le r$, $1 \le k, l \le d$, there exist a constant K and a locally bounded process H_t, such that $|\beta_{kj}| \le K$, and that $|h_{i,s}|$, $|\eta_{ij,s}|$, $|\gamma_{kl,s}|$, $|e_{ij,s}|$, $|f_{kl,s}|$, and $|g_{kl,s}|$ are all bounded by H_s for all ω and $0 \le s \le t$.

The existence of uniform bounds on all processes is necessary to the development of the large dimensional asymptotic results. This is a fairly standard assumption in the factor model literature, e.g., Bai (2003). Apart from the fact that jumps are excluded, Assumption 1 is fairly general, allowing almost arbitrary forms of heteroscedasticity in both X and Z. While jumps are undoubtedly important to explain asset return dynamics, their inclusion in this context would significantly complicate the model, as jumps may be present in some of the common factors, as well as in the idiosyncratic components (not necessarily simultaneously), and in their respective characteristics $(h_s, \eta_s, f_s, \gamma_s)$. We leave a treatment of jumps to future work.

We also impose the usual exogeneity assumption. Different from those discrete-time regressions or factor models, this assumption imposes path-wise restrictions, which is natural in a continuous-time model.

Assumption 2. For any $1 \le j \le r$, $1 \le k \le d$, and $0 \le t \le T$, $[Z_{k,t}, X_{j,t}] = 0$, where $[\cdot,\cdot]$ denotes the quadratic covariation.

Combined with (1), Assumptions 1 and 2 imply a factor structure on the spot covariance matrix of Y, denoted as c_t:

$$c_t = \beta e_t \beta^\intercal + g_t, \quad 0 \le t \le T. \tag{3}$$

This leads to a key equality:

$$\Sigma = \beta \mathrm{E}\beta^\intercal + \Gamma, \tag{4}$$

where for notational simplicity we omit the dependence of Σ, E, and Γ on the fixed T,

$$\Sigma = \frac{1}{T}\int_0^T c_t dt, \quad \Gamma = \frac{1}{T}\int_0^T g_t dt, \quad \text{and} \quad \mathrm{E} = \frac{1}{T}\int_0^T e_t dt. \tag{5}$$

To complete the model, we need an additional assumption on the residual covariance matrix Γ. We define

$$m_d = \max_{1\le i\le d}\sum_{1\le j\le d} 1_{\{\Gamma_{ij}\neq 0\}} \tag{6}$$

and impose a sparsity assumption on Γ, i.e., Γ cannot have too many non-zero elements.

Assumption 3. When $d \to \infty$, the degree of sparsity of Γ, m_d, grows at a rate which satisfies

$$d^{-a} m_d \to 0 \tag{7}$$

where a is some positive constant.

At low frequency, Bickel and Levina (2008a) establish the asymptotic theory for a thresholded sample covariance matrix estimator using this notion of sparsity for the covariance matrix. The degree of sparsity determines the convergence rate of their estimator. In a setting with low-frequency time series data, Fan et al. (2011, 2013) suggest imposing the sparsity assumption on the residual covariance matrix. As we will see, a low-rank plus sparsity structure turns out to be a good match for asset returns data at high frequency.

3. Estimators: Factor structure and number of factors

3.1. Identification and approximation

There is fundamental indeterminacy in a latent factor model. For instance, one can rotate the factors and their loadings simultaneously without changing the covariance matrix Σ. The canonical form of a classical factor model, e.g., Anderson (1958), imposes the identification restrictions that the covariance matrix E is the identity matrix and that $\beta^\intercal\beta$ is diagonal. The identification restriction $E = \mathrm{I}_r$ is often adopted by the literature of approximate factor models as well, e.g., Doz et al. (2011) or Fan et al. (2013). However, this is not appropriate in our setting, since the factor covariance matrix E depends on the sample path and hence is non-deterministic.

The goal in this paper is to propose a new covariance matrix estimator, taking advantage of the assumed low-rank plus sparsity structure. We do not, however, try to identify the factors or their loadings, which can be pinned down by imposing sufficiently many identification restrictions by adapting to the continuous-time setting the approach of, e.g., Bai and Ng (2013). Since we only need to separate $\beta\mathrm{E}\beta^\intercal$ and Γ from Σ, we can avoid some strict and, for this purpose unnecessary, restrictions.

Chamberlain and Rothschild (1983) study the identification problem of a general approximate factor model in discrete time. One of their key identification assumptions is that the eigenvalues of Γ are bounded, whereas the eigenvalues of $\beta\mathrm{E}\beta^\intercal$ diverge

because the factors are assumed pervasive. It turns out that for the purpose of covariance matrix estimation, we can relax the boundedness assumption on the eigenvalues of Γ.[1] In fact, the sparsity condition imposed on Γ, implies that its largest eigenvalue diverges but at a slower rate compared to the eigenvalues of $\beta \mathrm{E} \beta^\intercal$.

These considerations motivate the pervasiveness assumption below.

Assumption 4. E is a positive-definite covariance matrix, with distinct eigenvalues bounded away from 0. Moreover, $\left\| d^{-1} \beta^\intercal \beta - \mathrm{I}_r \right\| \to 0$, as $d \to \infty$.

This leads to our result on the identification of number of factors and the approximation of $\beta \mathrm{E} \beta^\intercal$ using eigenvalues and eigenvectors of Σ.

Theorem 1. *Suppose Assumptions 1, 2, 3 with $a = 1/2$, and 4 hold. Also, assume that $\|\mathrm{E}\|_{\mathrm{MAX}} \leq K$, $\|\Gamma\|_{\mathrm{MAX}} \leq K$ almost surely. Then r can be identified as $d \to \infty$. That is, if d is sufficiently large, $\bar{r} = r$, where $\bar{r} = \mathrm{argmin}_{1 \leq j \leq d} (d^{-1} \lambda_j + j d^{-1/2} m_d) - 1$, and $\{\lambda_j, 1 \leq j \leq d\}$ are the eigenvalues of Σ. Moreover, $\beta \mathrm{E} \beta^\intercal$ and Γ can be approximated by the eigenvalues and eigenvectors of Σ using*

$$\left\| \sum_{j=1}^{\bar{r}} \lambda_j \xi_j \xi_j^\intercal - \beta \mathrm{E} \beta^\intercal \right\|_{\mathrm{MAX}} \leq K d^{-1/2} m_d, \quad \text{and}$$

$$\left\| \sum_{j=\bar{r}+1}^{d} \lambda_j \xi_j \xi_j^\intercal - \Gamma \right\|_{\mathrm{MAX}} \leq K d^{-1/2} m_d,$$

where $\{\xi_j, 1 \leq j \leq d\}$ are the corresponding eigenvectors of Σ.

The key identification condition is $d^{-1/2} m_d = o(1)$, which creates a sufficiently wide gap between two groups of eigenvalues, so that we can identify the number of factors as well as approximate the two components of Σ. To identify the number of factors only, $d^{-1/2} m_d$ can be replaced by other penalty functions that dominate $d^{-1} m_d$, so that $d^{-1/2} m_d = o(1)$ can be relaxed, as shown in Theorem 2 below. Note that the identification and approximation are only possible when d is sufficiently large – the so called "blessing of dimensionality." This is in contrast with the result for a classical strict factor model, where the identification is achieved by matching the number of equations with the number of unknown parameters.

This model falls into the class of models with "spiked eigenvalues" in the literature, e.g., Doz et al. (2011) or Fan et al. (2013), except that the gap between the magnitudes of the spiked eigenvalues and the remaining ones is smaller in our situation. Moreover, our model is distinct from others in the class of spiked eigenvalue models discussed by Paul (2007) and Johnstone and Lu (2009), in which all eigenvalues are of the same order, and are bounded as the dimension grows. This explains the difference between our result and theirs – the eigenvalues and eigenvectors of the sample covariance matrix can be consistently recovered in our setting even when d grows faster than n does, as shown below. The next section provides a simple nonparametric covariance matrix estimator with easy-to-interpret tuning parameters, such as the number of digits of the GICS code and the number of latent factors. We also provide a new estimator to determine the number of factors.

3.2. High-frequency estimation of the covariance matrix

Let $\Delta_i^n Y = Y_{i\Delta_n} - Y_{(i-1)\Delta_n}$ denote the observed log-returns at sampling frequency Δ_n. The estimator begins with principal component decomposition of the covariance matrix estimator, using results from Aït-Sahalia and Xiu (2015).[2] Let the sample covariance matrix estimator be

$$\widehat{\Sigma} = \frac{1}{T} \sum_{i=1}^{n} (\Delta_i^n Y)(\Delta_i^n Y)^\intercal \tag{8}$$

and let $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \cdots > \widehat{\lambda}_d$ denote the simple eigenvalues of $\widehat{\Sigma}$, and $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_d$ the corresponding eigenvectors.

With $\widehat{r}$, an estimator of r discussed below, we can in principle separate Γ from Σ:

$$\widehat{\Gamma} = \sum_{j=\widehat{r}+1}^{d} \widehat{\lambda}_j \widehat{\xi}_j \widehat{\xi}_j^\intercal.$$

Since Γ is assumed sparse, we can enforce sparsity through, e.g., soft-, hard-, or adaptive thresholding; see, e.g., Rothman et al. (2009) for a discussion of thresholding techniques. But this would inevitably introduce tuning parameters that might be difficult to select and interpret. Moreover, it is difficult to ensure that after thresholding, the resulted covariance matrix estimator remains positive semi-definite in finite samples.

We adopt a different approach motivated from the economic intuition that firms within similar industries, e.g., Pepsico and Coca Cola, or Target and Walmart, are expected to have higher correlations beyond what can be explained by their loadings on common and systematic factors. This intuition motivates a block-diagonal structure on the residual covariance matrix Γ, once stocks are sorted by their industrial classification. This strategy leads to a simpler, positive semi-definite by construction, and economically-motivated estimator. It requires the following assumption.

Assumption 5. Γ is a block diagonal matrix, and the set of its non-zero entries, denoted by S, is known prior to the estimation.

The block-diagonal assumption is compatible with the sparsity Assumption 3. In fact, m_d in (6) is the size of the largest block. There is empirical support for the block-diagonal assumption on Γ: for instance, Fan et al. (2016) find such a pattern of Γ in their regression setting, after sorting the stocks by the GICS code and stripping off the part explained by observable factors. Fig. 1 illustrates the structure of the covariance matrix.

Our covariance matrix estimator $\widehat{\Sigma}^S$ is then given by

$$\widehat{\Sigma}^S = \sum_{j=1}^{\widehat{r}} \widehat{\lambda}_j \widehat{\xi}_j \widehat{\xi}_j^\intercal + \widehat{\Gamma}^S, \tag{9}$$

where by imposing the block-diagonal structure,

$$\widehat{\Gamma}^S = (\widehat{\Gamma}_{ij} 1_{(i,j) \in S}). \tag{10}$$

This covariance matrix estimator is similar in construction to the POET estimator by Fan et al. (2013) for discrete time series, except that we block-diagonalize Γ instead of using soft- or hard-thresholding.

Equivalently, we can also motivate our estimator from least-squares estimation analogously to Stock and Watson (2002), Bai and Ng (2013) and Fan et al. (2013) in a discrete-time low frequency setting. Our estimator can be re-written as

$$\widehat{\Sigma}^S = T^{-1} F G G^\intercal F^\intercal + \widehat{\Gamma}^S, \quad \widehat{\Gamma} = T^{-1} (\mathcal{Y} - FG)(\mathcal{Y} - FG)^\intercal, \quad \text{and}$$
$$\widehat{\Gamma}^S = (\widehat{\Gamma}_{ij} 1_{(i,j) \in S}), \tag{11}$$

[1] This unboundedness issue has also been studied by Onatski (2010) in a different setting.

[2] Without the benefit of a factor model (1), PCA should be employed on the spot covariance matrices instead of the integrated covariance matrix.

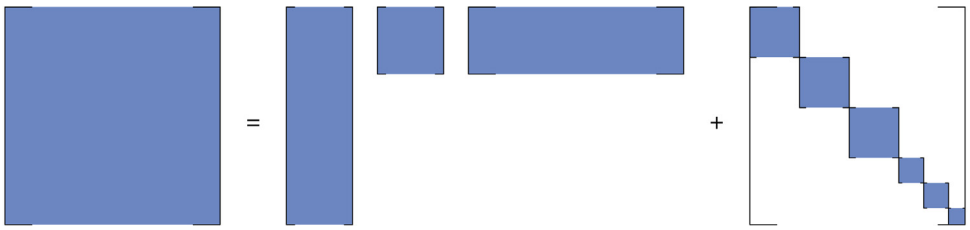

Fig. 1. The structure of the covariance matrix. Note: This figure illustrates the structure of the covariance matrix we impose in (4):
$$\Sigma = \beta \mathrm{E} \beta^\intercal + \Gamma,$$
where Γ, sorted by GICS codes, is block diagonal, and Σ, E, and Γ depend on realizations of the sample paths.

where $\mathcal{Y} = \left(\Delta_1^n Y, \Delta_2^n Y, \ldots, \Delta_n^n Y\right)$ is a $d \times n$ matrix, $G = (g_1, g_2, \ldots, g_n)$ is $\widehat{r} \times n$, $F = (f_1, f_2, \ldots, f_d)^\intercal$ is $d \times \widehat{r}$, and F and G solve the least-squares problem:

$$\begin{aligned}(F, G) &= \arg \min_{f_k, g_i \in \mathbb{R}^{\widehat{r}}} \sum_{i=1}^{n} \sum_{k=1}^{d} \left(\Delta_i^n Y_k - f_k^\intercal g_i\right)^2 \\ &= \arg \min_{F \in \mathcal{M}_{d \times \widehat{r}}, G \in \mathcal{M}_{\widehat{r} \times n}} \|\mathcal{Y} - FG\|_{\mathrm{F}}^2 \end{aligned} \tag{12}$$

subject to the constraints

$$d^{-1} F^\intercal F = \mathrm{I}_{\widehat{r}}, \quad GG^\intercal \text{ is an } \widehat{r} \times \widehat{r} \text{ diagonal matrix.} \tag{13}$$

The least-squares estimator is employed by Bai and Ng (2002), Bai (2003) and Fan et al. (2013). Bai and Ng (2002) suggest that PCA can be applied to either the $d \times d$ matrix $\mathcal{Y}\mathcal{Y}^\intercal$ or the $n \times n$ matrix $\mathcal{Y}^\intercal\mathcal{Y}$, depending on the relative magnitude of d and n. We apply PCA to the $d \times d$ matrix $\mathcal{Y}\mathcal{Y}^\intercal$ regardless, because in our high frequency continuous-time setting, the spot covariance matrices e_t and c_t are stochastically time-varying, so that the $n \times n$ matrix is conceptually more difficult to analyze. It is straightforward to verify that $F = d^{1/2}\left(\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{\widehat{r}}\right)$ and $G = d^{-1} F^\intercal \mathcal{Y}$ are the solutions to this optimization problem, and the estimator given by (11) is then the same as that given by (9) and (10).

3.3. High-frequency estimation of the number of factors

To determine the number of factors, we propose the following estimator using a penalty function g:

$$\widehat{r} = \arg \min_{1 \le j \le r_{\max}} \left(d^{-1} \lambda_j(\widehat{\Sigma}) + j \times g(n, d)\right) - 1, \tag{14}$$

where $r_{\max}$ is an upper bound of $r + 1$. In theory, the choice of $r_{\max}$ does not play a role. It is only used to avoid reaching an economically nonsensical choice of r in finite samples. The penalty function $g(n, d)$ satisfies two criteria. Firstly, the penalty cannot dominate the signal, i.e., the value of $d^{-1}\lambda_j(\Sigma)$, when $1 \le j \le r$. Since $d^{-1}\lambda_r(\Sigma)$ is $O_p(1)$ as d increases, the penalty should shrink to 0. Secondly, the penalty should dominate the estimation error as well as $d^{-1}\lambda_{r+1}(\Sigma)$ when $r + 1 \le j \le d$ to avoid overshooting.

This estimator is similar in spirit to that introduced by Bai and Ng (2002) in the classical low frequency setting. They suggest to estimate r by minimizing the penalized objective function:

$$\widehat{r} = \arg \min_{1 \le j \le r_{\max}} (d \times T)^{-1} \|\mathcal{Y} - F(j)G(j)\|_{\mathrm{F}}^2 + \text{penalty}, \tag{15}$$

where the dependence of F and G on j is highlighted. It turns out, perhaps not surprisingly, that

$$(d \times T)^{-1} \|\mathcal{Y} - F(j)G(j)\|_{\mathrm{F}}^2 = d^{-1} \sum_{i=j+1}^{d} \lambda_i(\widehat{\Sigma}), \tag{16}$$

which is closely related to our proposed objective function. It is, however, easier to use our proposal as it does not involve estimating the sum of many eigenvalues. The proof is also simpler.

Alternative methods to determine the number of factors include Hallin and Liška (2007), Amengual and Watson (2007), Alessi et al. (2010), Kapetanios (2010), and Onatski (2010). Ahn and Horenstein (2013) propose an estimator by maximizing the ratios of adjacent eigenvalues. Their approach is convenient in that it does not involve any penalty function. The consistency of their estimator relies on the random matrix theory established by, e.g., Bai and Yin (1993), so as to establish a sharp convergence rate for the eigenvalue ratio of the sample covariance matrix. Such a theory is not available for continuous-time semimartingales to the best of our knowledge. So we propose an alternative estimator, for which we can establish the desired consistency in the continuous-time context without using random matrix theory.

3.4. Consistency of the estimators

Recall that our asymptotics are based on a dual increasing frequency and dimensionality, while the number of factors is finite. That is, $\Delta_n \to 0$, $d \to \infty$, and r is fixed (but unknown). We first establish the consistency of $\widehat{r}$.

Theorem 2. *Suppose Assumptions 1, 2, 3 with $a = 1$, and 4 hold. Suppose that $\Delta_n \log d \to 0$, $g(n, d) \to 0$, and $g(n, d)\left((\Delta_n \log d)^{1/2} + d^{-1} m_d\right)^{-1} \to \infty$, we have $\mathbb{P}(\widehat{r} = r) \to 1$.*

A choice of the penalty function could be

$$g(n, d) = \mu \left((n^{-1} \log d)^{1/2} + d^{-1} m_d\right)^{\kappa}, \tag{17}$$

where μ and κ are some constants and $0 < \kappa < 1$. While it might be difficult/arbitrary to choose these tuning parameters in practice, the covariance matrix estimates are not overly sensitive to the numbers of factors. Also, the scree plot output from PCA offers guidance as to the value of r and can be used as a check on the resulting estimator. Practically speaking, r is no different from a "tuning parameter." And it is much easier to interpret r than μ and κ above. In the later portfolio allocation study, we choose a range of values of r to compare the covariance matrix estimator with that using observable factors. As long as r is larger than 3 but not as large as, say, 20, the results do not change much and the interpretation remains the same. A rather small value of r, all the way to $r = 0$, results in model misspecification, whereas a rather large r leads to overfitting.

It is worth mentioning that to identify and estimate the number of factors consistently, the weaker assumption $a = 1$ in Assumption 3 is imposed, compared to the stronger assumption $a = 1/2$ required in Theorem 1, which we need to identify $\beta \mathrm{E} \beta^\intercal$ and Γ.

The next theorem establishes the desired consistency of the covariance matrix estimator.

Theorem 3. *Suppose Assumptions 1, 2 , 3 with $a = 1/2$, 4 and 5 hold. Suppose that $\Delta_n \log d \to 0$. Suppose further that $\widehat{r} \to r$ with probability approaching 1, then we have*

$$\left\|\widehat{\Gamma}^S - \Gamma\right\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right).$$

Moreover, we have

$$\left\|\widehat{\Sigma}^S - \Sigma\right\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right).$$

Compared to the rate of convergence of the regression based estimator in Fan et al. (2016) where factors are observable, i.e., $O_p((\Delta_n \log d)^{1/2})$, the convergence rate of the PCA-based estimator depends on a new term $d^{-1/2} m_d$, due to the presence of unobservable factors, as can be seen from Theorem 1. We consider the consistency under the entry-wise norm instead of the operator norm, partially because the eigenvalues of Σ themselves grow at the rate of $O(d)$, so that their estimation errors do not shrink to 0, when the dimension d increases exponentially, relative to the sampling frequency Δ_n.

In terms of the portfolio allocation, the precision matrix perhaps plays a more important role than the covariance matrix. For instance, the minimum variance portfolio is determined by the inverse of the Σ instead of Σ itself. The estimator we propose is not only positive-definite, but is also well-conditioned. This is because the minimum eigenvalue of the estimator is bounded from below with probability approaching 1. The next theorem describes the asymptotic behavior of the precision matrix estimator under the operator norm.

Theorem 4. *Suppose Assumptions 1, 2, 3 with $a = 1/2$, 4, and 5 hold. Suppose that $\Delta_n \log d \to 0$. Suppose further that $\widehat{r} \to r$ with probability approaching 1, then we have*

$$\left\|\widehat{\Gamma}^S - \Gamma\right\| = O_p\left(m_d(\Delta_n \log d)^{1/2} + d^{-1/2} m_d^2\right).$$

If in addition, $\lambda_{\min}(\Gamma)$ is bounded away from 0 almost surely, $d^{-1/2} m_d^2 = o(1)$ and $m_d(\Delta_n \log d)^{1/2} = o(1)$, then $\lambda_{\min}(\widehat{\Sigma}^S)$ is bounded away from 0 with probability approaching 1, and

$$\left\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\right\| = O\left(m_d^3\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right)\right).$$

The convergence rate of the regression based estimator in Fan et al. (2016) with observable factors is $O_p(m_d(\Delta_n \log d)^{1/2})$. In their paper, the eigenvalues of Γ are bounded from above, whereas we relax this assumption in this paper, which explains the extra powers of m_d here. As above, $d^{-1/2} m_d$ reflects the loss due to ignorance of the latent factors.

As a by-product, we can also establish the consistency of factors and loadings up to some matrix transformation.

Theorem 5. *Suppose Assumptions 1, 2, 3 with $a = 1/2$, and 4 hold. Suppose that $\Delta_n \log d \to 0$. Suppose further that $\widehat{r} \to r$ with probability approaching 1, then there exists a $r \times r$ matrix H, such that with probability approaching 1, H is invertible, $\|HH^\intercal - \mathrm{I}_r\| = \|H^\intercal H - \mathrm{I}_r\| = o_p(1)$, and*

$$\|F - \beta H\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right),$$
$$\left\|G - H^{-1}\mathcal{X}\right\| = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right),$$

where F and G are defined in (12), and $\mathcal{X} = \left(\Delta_1^n X, \Delta_2^n X, \ldots, \Delta_n^n X\right)$ is a $r \times n$ matrix.

The presence of the H matrix is due to the indeterminacy of a factor model. Bai and Ng (2013) impose further assumptions so as to identify the factors. For instance, one set of identification assumptions may be that the first few observed asset returns are essentially noisy observations of the factors themselves. For the purpose of covariance matrix estimation, such assumptions are not needed. It is also worth pointing out that for the estimation of factors, Assumption 5 is not needed.

4. Monte Carlo simulations

In order to concentrate on the effect of an increasing dimensionality, without additional complications, we have established the theoretical asymptotic results in an idealized setting without market microstructure noise. This setting is realistic and relevant only for returns sampled away from the highest frequencies. In this section, we examine the effect of subsampling on the performance of our estimators, making them robust to the presence of both asynchronous observations and microstructure noise.

We sample paths from a continuous-time r-factor model of d assets specified as follows:

$$\begin{aligned} dY_{i,t} &= \sum_{j=1}^{r} \beta_{i,j} dX_{j,t} + dZ_{i,t}, \quad dX_{j,t} = b_j dt + \sigma_{j,t} dW_{j,t}, \\ dZ_{i,t} &= \gamma_i^\intercal dB_{i,t}, \end{aligned} \tag{18}$$

where W_j is a standard Brownian motion and B_i is a d-dimensional Brownian motion, for $i = 1, 2, \ldots, d$, and $j = 1, 2, \ldots, r$. They are mutually independent. X_j is the jth unobservable factor. One of the Xs, say the first, is the market factor, so that its associated βs are positive. The covariance matrix of Z is a block-diagonal matrix, denoted by Γ, that is, $\Gamma_{il} = \gamma_i^\intercal \gamma_l$. We allow for time-varying $\sigma_{j,t}$ which evolves according to the following system of equations:

$$d\sigma_{j,t}^2 = \kappa_j(\theta_j - \sigma_{j,t}^2)dt + \eta_j \sigma_{j,t} d\widetilde{W}_{j,t}, \quad j = 1, 2, \ldots, r, \tag{19}$$

where $\widetilde{W}_j$ is a standard Brownian motion with $\mathbb{E}[dW_{j,t} d\widetilde{W}_{j,t}] = \rho_j dt$. We choose $d = 500$ and $r = 3$. In addition, $\kappa = (3, 4, 5)$, $\theta = (0.05, 0.04, 0.03)$, $\eta = (0.3, 0.4, 0.3)$, $\rho = (-0.60, -0.40, -0.25)$, and $b = (0.05, 0.03, 0.02)$. In the cross-section, we sample $\beta_1 \sim \mathcal{U}[0.25, 1.75]$, and sample $\beta_2, \beta_3 \sim \mathcal{N}(0, 0.5^2)$. The variances on the diagonal of Γ are uniformly generated from $[0.05, 0.20]$, with constant within-block correlations sampled from $\mathcal{U}[0.10, 0.50]$ for each block. To generate blocks, we fix the largest block size at MAX, and randomly generate the sizes of the remaining blocks from a Uniform distribution [10, MAX], such that the total sizes of all blocks is equal to d. The number of blocks is thereby random. The cross-sectional βs, and the covariance matrix Γ, including its block structure, its diagonal variances, and its within-block correlations are randomly generated once and then fixed for all Monte Carlo repetitions. Their variations do not change the simulation results. We fix MAX at 15, 25, and 35, respectively, and there are 41, 30, and 23 blocks, accordingly.

To mimic the effect of microstructure noise and asynchronicity, we add a Gaussian noise with mean zero and variance 0.001^2 to the simulated log prices. The data are then censored using Poisson sampling, where the number of observations for each asset is drawn from a truncated log-normal distribution. The log-normal distribution $\log \mathcal{N}(\mu, \sigma^2)$ has parameters $\mu = 2$, 500 and $\sigma = 0.8$. The lower and upper truncation boundaries are 500 and 23 400, respectively, for data generated at 1-second frequency. We estimate the covariance matrix based on data subsampled at various frequencies, from every 5 s to 2 observations per day, using the previous-tick approach from a $T = 21$-day interval, with each day having 6.5 trading hours. We sample 100 paths.

Table 1 provides the averages of $\|\widehat{\Sigma}^S - \Sigma\|_{\text{MAX}}$ and $\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\|$ in various scenarios. We apply the PCA approach suggested in this paper, and the regression estimator of Fan et al. (2016), which assumes X to be observable, to the idealized dataset without any noise or asynchronicity. The results are shown in Columns PCA* and REG*. Columns PCA and REG contain the estimation results using the polluted data where noise and censoring have been applied. In the last column, we report the estimated number of factors with the polluted data. We use as tuning parameters $\kappa = 0.5$, $r_{\max} = 20$, and $\mu = 0.02 \times \lambda_{\min(d,n)/2}(\widehat{\Sigma})$. The use of the

Table 1
Simulation results.

MAX	Freq.	$\|\widehat{\Sigma}^S - \Sigma\|_{MAX}$				$\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\|$				# Factors
		REG*	PCA*	REG	PCA	REG*	PCA*	REG	PCA	
	5	0.004	0.010	2.063	2.063	0.51	1.42	31.94	31.94	1
	15	0.007	0.011	0.785	0.785	0.88	1.55	31.60	31.59	1
	30	0.009	0.012	0.404	0.404	1.23	1.72	31.05	31.02	2
	60	0.014	0.015	0.229	0.217	1.81	2.03	29.94	29.87	3
15	300	0.031	0.032	0.137	0.112	4.30	4.20	24.58	24.17	3
	900	0.054	0.054	0.078	0.070	8.33	8.07	18.04	17.31	3
	1800	0.078	0.079	0.084	0.084	13.44	13.18	13.52	12.95	3
	3900	0.112	0.113	0.116	0.116	26.22	26.05	21.11	21.19	3
	4680	0.124	0.124	0.124	0.125	31.01	30.94	26.69	26.80	3
	11700	0.195	0.196	0.196	0.196	116.23	117.10	113.27	112.45	1
	5	0.004	0.012	2.063	2.063	0.63	2.38	35.66	35.66	1
	15	0.007	0.013	0.785	0.785	1.10	2.51	35.32	35.30	1
	30	0.010	0.014	0.403	0.403	1.57	2.66	34.76	34.73	1
	60	0.014	0.017	0.247	0.222	2.21	2.89	33.63	33.55	3
25	300	0.032	0.033	0.148	0.118	5.46	5.32	27.76	27.20	3
	900	0.055	0.055	0.080	0.071	10.90	10.65	21.15	20.09	3
	1800	0.078	0.078	0.084	0.083	18.51	18.20	16.34	15.46	3
	3900	0.116	0.116	0.116	0.117	37.71	37.07	31.60	31.60	3
	4680	0.128	0.128	0.130	0.131	47.72	47.28	40.93	41.41	3
	11700	0.199	0.200	0.200	0.201	315.69	314.05	312.97	307.50	1
	5	0.004	0.010	2.064	2.064	0.59	1.71	28.46	28.46	1
	15	0.007	0.011	0.785	0.785	1.04	1.80	28.15	28.11	1
	30	0.010	0.013	0.404	0.404	1.50	1.97	27.65	27.59	1
	60	0.014	0.016	0.227	0.217	2.16	2.28	26.74	26.62	3
35	300	0.031	0.032	0.139	0.111	5.50	5.42	21.34	20.94	3
	900	0.053	0.054	0.079	0.070	11.49	11.41	15.12	14.50	3
	1800	0.076	0.076	0.082	0.081	20.66	20.48	14.87	14.82	3
	3900	0.115	0.116	0.117	0.117	49.08	48.72	43.54	43.61	3
	4680	0.124	0.124	0.127	0.127	65.04	65.08	59.80	59.96	3
	11700	0.193	0.194	0.195	0.196	2523.60	2422.00	2359.70	2327.80	1

Note: In this table, we report the values of $\|\widehat{\Sigma}^S - \Sigma\|_{MAX}$ and $\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\|$ for each subsampling frequency ranging from one observation every 5 s to 2 observations per day within a 21-day fixed period, with the size of the largest block being 15, 25, and 35 respectively. The first column displays the size of the largest block. The second column displays the sampling frequencies in seconds. Columns REG* and PCA* report the results of regression and the PCA methods respectively, using synchronous observations without microstructure noise. Columns REG and the PCA are based on the polluted data. Columns REG*, REG, and PCA* all assume 3 factors. The results in the PCA column are obtained by estimating the number of factors first. The last column reports the median estimates of the number of factors using the polluted data.

median eigenvalue $\lambda_{\min(d,n)/2}(\widehat{\Sigma})$ helps adjust the level of average eigenvalues for better accuracy.

We find the following. First, the values of $\|\widehat{\Sigma} - \Sigma\|_{MAX}$ in Columns REG and PCA are almost identical. This is due to the fact that the largest entry-wise errors are likely achieved along the diagonals, and that the estimates on the diagonal are identical to the sample covariance estimates, regardless of whether the factors are observable or not. As to the precision matrix under the operator norm, i.e., $\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\|$, the differences between the two estimators are noticeable despite being very small. While the PCA approach uses less information by construction, it can perform as well as the REG approach. That said, the benefit of using observable factors is apparent from the comparison between Columns REG* and PCA* when the sampling frequency is high, as the results based on the PCA* are worse. This also agrees with what our theory suggests: when the sampling frequency is high, the $d^{-1/2}m_d$ term dominates; whereas when the frequency is low, the $(\Delta_n \log d)^{1/2}$ term is more important. Second, microstructure effect does negatively affect the estimates when the data is sampled every few seconds or more frequently. Subsampling does mitigate the effect of microstructure noise, but it also raises another concern with a relatively increasing dimensionality – the ratio of the cross-sectional dimension against the number of observations. The sweet spot in that trade-off appears to be in the range between 15 and 30 min given an overall length of $T = 21$ days. Third, as the size of the largest blocks m_d increases, the performance of the estimators deteriorates, as expected from the theory. Finally, the number of factors is estimated fairly precisely for most frequencies. Not surprisingly, the estimates are off at both ends of the sampling frequency (due to insufficient amount of data in one case, and microstructure noise in the other).

5. Empirical results

5.1. Data

We collect intraday observations of the S&P 500 index constituents from January 2004 to December 2012 from the TAQ database. We follow the usual procedures, see, e.g., Aït-Sahalia and Jacod (2014), to clean the data and subsample returns of each asset every 15 min. The overnight returns are excluded to avoid dividend issuances and stock splits.

The S&P 500 constituents have obviously been changing over this long period. As a result, there are in total 736 stocks in our dataset, with 498–502 of them present on any given day. We calculate the covariance matrix for all index constituents that have transactions every day both for this month and the next. We do not require stocks to have all 15-minute returns available, as we use the previous tick method to interpolate the missing observations. As a result, each month we have over 491 names, and the covariance matrix for these names is positive-definite. Since we remove the stocks de-listed during the next period, there is potential for some slight survivorship bias. However, all the strategies we compare are exposed to the same survivorship bias, hence this potential bias should not affect the comparisons below. Also, survivorship bias in this setup only matters for a maximum of one month ahead, because the analysis is repeated each month. This is potentially an important advantage of using high frequency data compared to the long time series needed at low frequency.

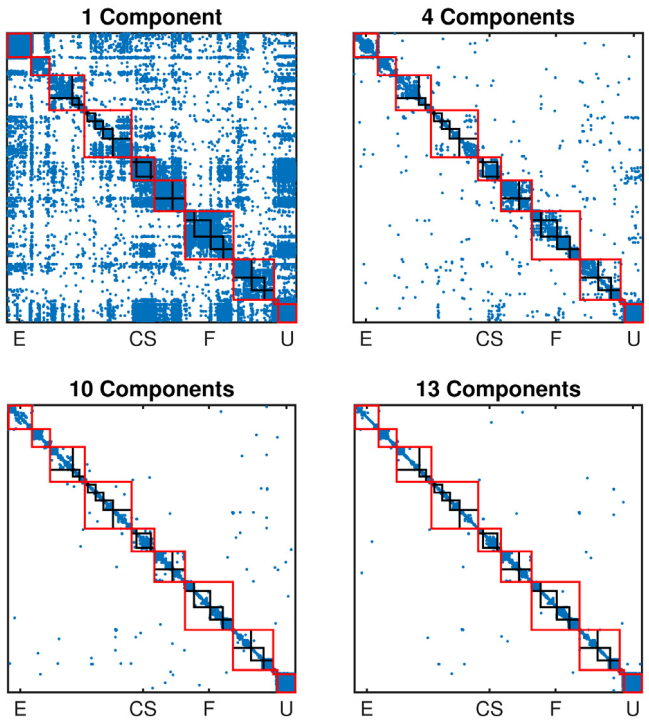

Fig. 2. The sparsity pattern of the residual covariance matrices. Note: The figure displays the significant entries of the residual covariance matrices, relative to 1, 4, 10, and 13 latent factors. The red (resp. black) squares highlight those stocks that belong to the same sector (resp. industry group). We highlight 4 sectors on the *x*-axis, including E (Energy), CS (Consumer Staples), F (Financials), and U (Utilities).

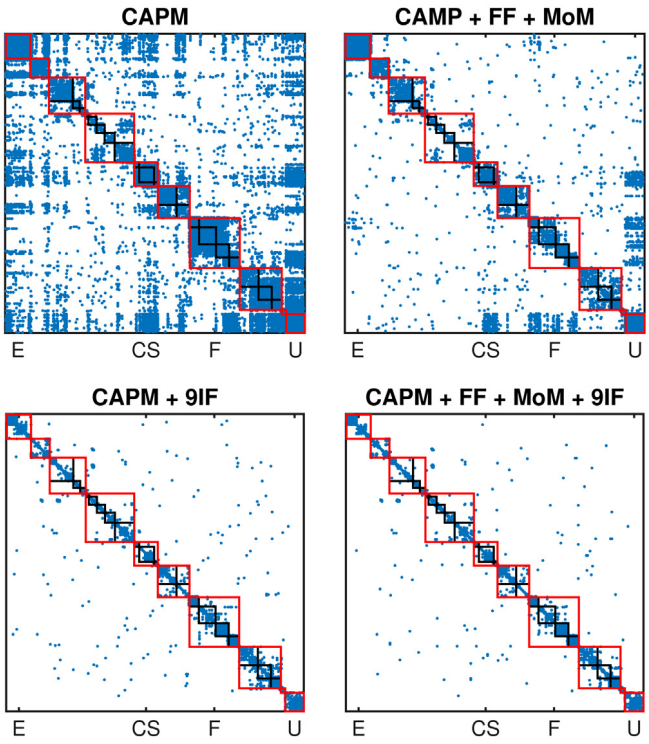

Fig. 3. The sparsity pattern of the residual covariance matrices. Note: The figure displays the significant entries of the residual covariance matrices, relative to 1, 4, 10, and 13 observable factors. The red (resp. black) squares highlight those stocks that belong to the same sector (resp. industry group). CAPM denotes one factor case using the market portfolio, FF refers to the two additional Fama–French factors, MoM denotes the momentum factor, whereas the 9IF refers to the 9 industrial ETF factors. We highlight 4 sectors on the *x*-axis, including E (Energy), CS (Consumer Staples), F (Financials), and U (Utilities).

In addition, we collect the Global Industrial Classification Standard (GICS) codes from the Compustat database. These 8-digit codes are assigned to each company in the S&P 500. The code is split into 4 groups of 2 digits. Digits 1–2 describe the company's sector; digits 3–4 describe the industry group; digits 5–6 describe the industry; digits 7–8 describe the sub-industry. The GICS codes are used to sort stocks and form blocks of the residual covariance matrices. The GICS codes also change over time. The time series median of the largest block size is 77 for sector-based classification, 38 for industry group, 24 for industry, and 14 for sub-industry categories.

For comparison purpose, we also make use of observable factors constructed from high-frequency returns, including the market portfolio, the small-minus-big market capitalization (SMB) portfolio, and high-minus-low price–earnings ratio (HML) portfolio in the Fama–French 3 factor model, as well as the daily-rebalanced momentum portfolio formed by sorting stock returns between the past 250 days and 21 days. We construct these factors by adapting the Fama–French procedure to a high frequency setting (see Aït-Sahalia et al., 2014). We also collect from TAQ the 9 industry SDPR ETFs (Energy (XLE), Materials (XLB), Industrials (XLI), Consumer Discretionary (XLY), Consumer Staples (XLP), Health Care (XLV), Financial (XLF), Information Technology (XLK), and Utilities (XLU)).

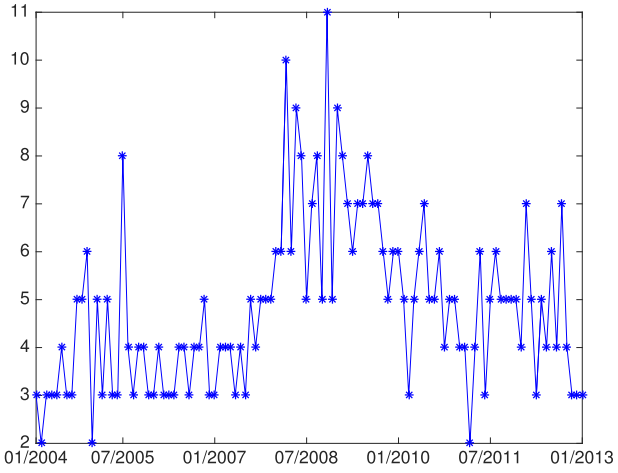

Fig. 4. Estimates of the number of factors. Note: This figure plots the time series of the estimated number of factors using PCA. The tuning parameters in the penalty function are $\mu = 0.02 \times \lambda_{\min(d,n)/2}(\widehat{\Sigma})$, $\kappa = 0.5$, and $r_{\max} = 20$.

5.2. The number of factors

Prior to estimating the number of factors, we verify empirically the sparsity and block-diagonal pattern of the residual covariance matrix using various combinations of factors. In Figs. 2 and 3, we indicate the economically significant entries of the residual covariance estimates for the year 2012, after removing the part driven by 1, 4, 10, and 13 PCA-based factors, respectively. The criterion we employ to indicate economic significance is that the correlation is at least 0.15 for at least 1/3 of the year. These two thresholds as well as the choice of the year 2012 are arbitrary, but varying these numbers or the subsample do not change the pattern and the message of the plots. We also compare these plots with those based on observable factors. The benchmark one-factor model we use is the CAPM. For the 4-factor model, we use the 3 Fama–French portfolios plus the momentum portfolio. The 10-factor model is based on the market portfolio and 9 industrial ETFs. The 13-factor model uses all above observable factors.

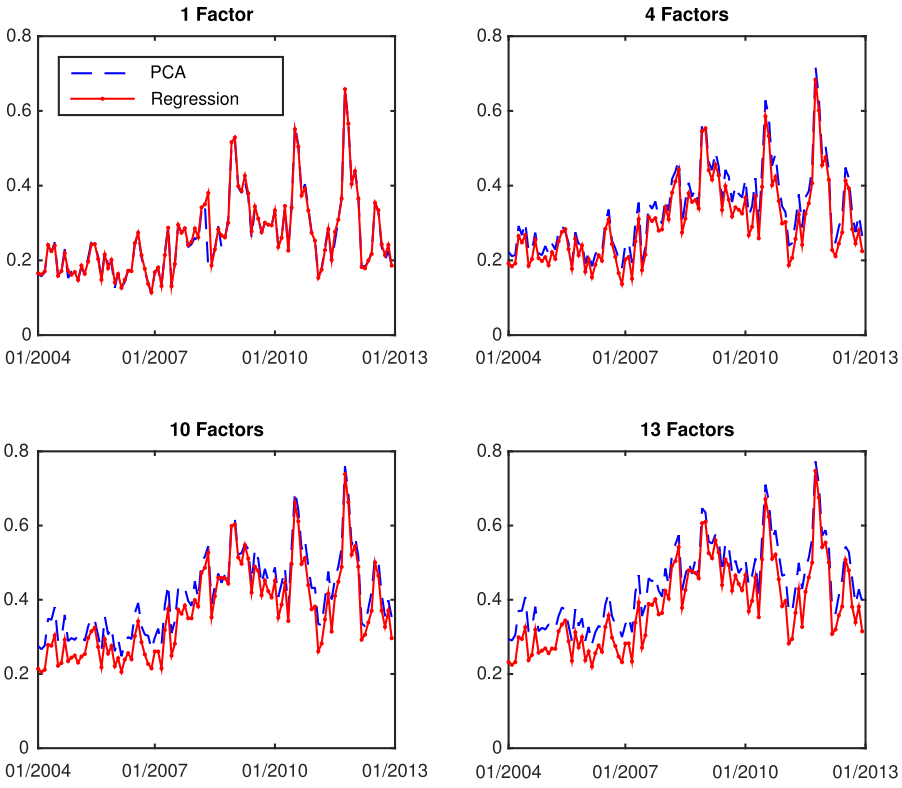

Fig. 5. In-sample R^2 comparison. Note: This figure plots the time series of the cross-sectional medians of R^2s based on the latent factors identified from the PCA, as well as those based on the observable factors. The number of factors refers to the number of latent components from the PCA approach and the number of portfolios used in the regression approach.

We find that the PCA approach provides sharp results in terms of identifying the latent factors. The residual covariance matrix exhibits a clear block-diagonal pattern after removing as few as 4 latent factors. The residual correlations are likely due to idiosyncrasies within sectors or industrial groups. This pattern empirically documents the low-rank plus sparsity structure we imposed in the theoretical analysis. Instead of thresholding all off-diagonal entries as suggested by the strict factor model, we maintain within-sector or within-industry correlations, and produce more accurate estimates. As documented in Fan et al. (2016), a similar pattern holds with observable factors, but more such factors are necessary to obtain the same degree of the sparsity obtained here by the PCA approach.

We then use the estimator $\widehat{r}$ to determine the number of common factors each month. The time series plot is shown in Fig. 4. The times series is relatively stable, identifying 3 to 5 factors for most of the sample subperiods. The result agrees with the pattern in the residual sparsity plot, and is consistent with the scree plot shown in Aït-Sahalia and Xiu (2015) for S&P 100 constituents.

5.3. In-sample R^2 comparison

We now compare the variation explained by an increasing number of latent factors with the variation explained by the same number of observable factors. We calculate the in-sample R^2 respectively for each stock and for each month, and plot the time series of their cross-sectional medians in Fig. 5. Not surprisingly, the first latent factor agrees with the market portfolio return and explains as much variation as the market portfolio does. When additional factors are included, both the latent factors and the observable factors can explain more variation, with the former explaining slightly more. Both methods end up in large agreement in terms of explained variation, suggesting that the observable factors identified in the literature are fairly effective at capturing the latent common factors.

One interesting finding is that the R^2s based on high frequency data are significantly higher than those reported in the literature with daily data, see e.g., Herskovic et al. (2016). This may reflect the increased signal-to-noise ratio from intraday data sampled at an appropriate frequency.

5.4. Out-of-sample portfolio allocation

We then examine the effectiveness of the covariance estimates in terms of portfolio allocation. We consider the following constrained portfolio allocation problem:

$$\min_{\omega} \omega^{\intercal} \widehat{\Sigma}^{S} \omega, \quad \text{subject to } \omega^{\intercal} 1 = 1, \|\omega\|_1 \leq \gamma, \tag{20}$$

where $\|\omega\|_1 \leq \gamma$ imposes an exposure constraint. When $\gamma = 1$, short-sales are ruled out, i.e., all portfolio weights are non-negative (since $\sum_{i=1}^{d} \omega_i = 1$, $\sum_{i=1}^{d} |\omega_i| \leq 1$ imposes that $\omega_i \geq 0$ for all $i = 1, \ldots, d$). When γ is small, the optimal portfolio is sparse, i.e., many weights are zero. When the γ constraint is not binding, the optimal portfolio coincides with the global minimum variance portfolio.

For each month from February 2004 to December 2012, we build the optimal portfolio based on the covariance estimated during the past month.[3] This amounts to assuming that $\widehat{\Sigma}_t^S \approx \mathrm{E}_t(\Sigma_{t+1})$, which is a common empirical strategy in practice. We compare the out-of-sample performance of the portfolio allocation problem (20) with a range of exposure constraints. The results are shown in Fig. 6.

[3] We estimate the covariance matrix for stocks that are constituents of the index during the past month and the month ahead. Across all months in our sample, we have over 491 stocks available.

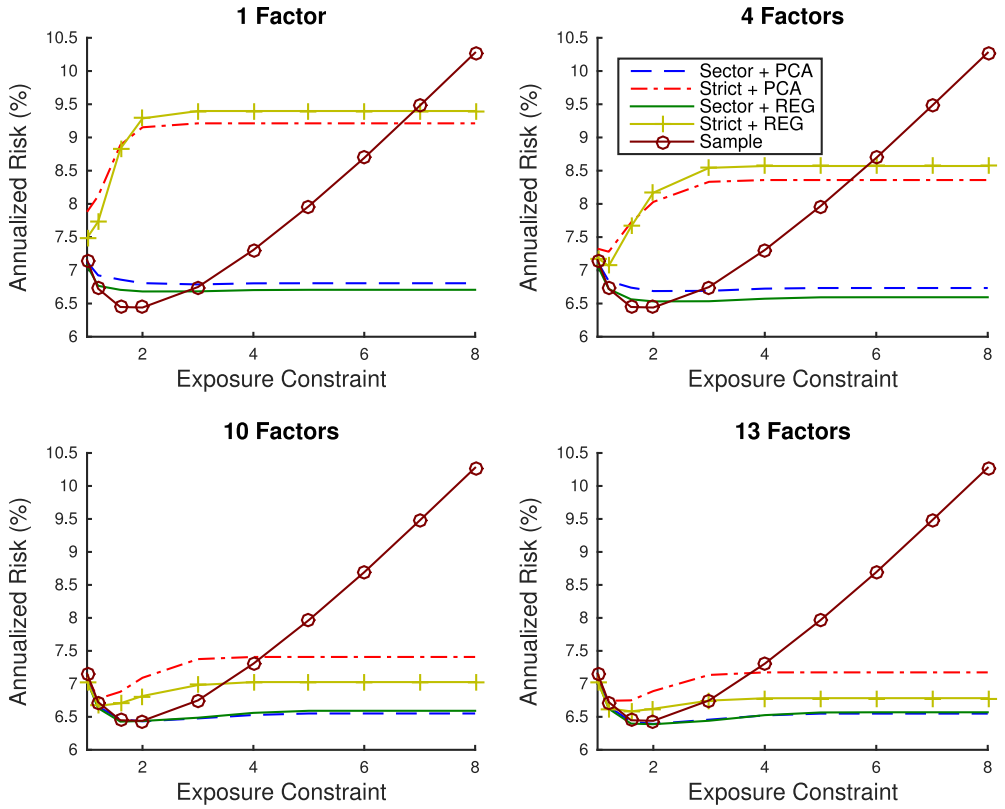

Fig. 6. Out-of-Sample Risk of the Portfolio. Note: This figure compares the time series average of the out-of-sample monthly volatility from 2004 and 2012. The x-axis is the exposure constraint γ in the optimization problem (20). The results are based on 5 covariance matrix estimators, including the sample covariance matrix (Sample), the PCA approach with sector-grouped block-diagonal residual covariance (Sector + PCA), PCA with diagonal residual covariance (Strict + PCA), and their regression counterparts (Sector + REG, Strict + REG). The number of factors refers to the number of principal components for the PCA approach and the number of portfolios factors for the regression approach.

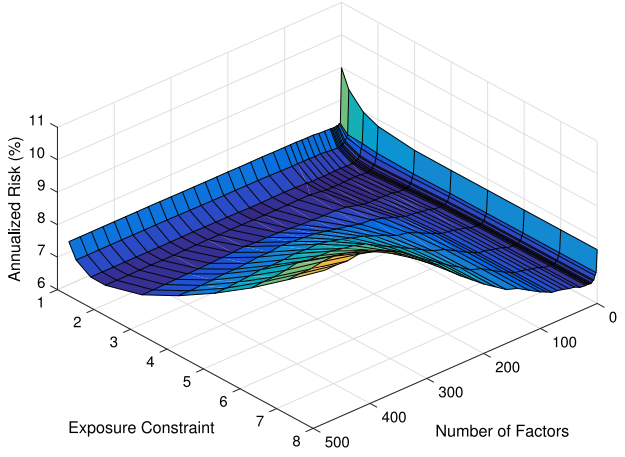

Fig. 7. Out-of-sample risk of the portfolio as a function of the exposure constraint and number of factors. Note: This figure reports the time series average of the out-of-sample monthly volatility (z-axis) from 2004 and 2012. The x-axis is the exposure constraint γ in the optimization problem (20), whereas the y-axis is the number of factors, i.e., the number of principal components used by the PCA approach with sector-grouped block-diagonal residual covariance (Sector + PCA).

We find that for the purpose of portfolio allocation, PCA performs out of sample as well as the regression method does. The performance of PCA further improves when combined with the sector-based block-diagonal structure of the residual covariance matrix. The allocation based on the sample covariance matrix only performs reasonably well when the exposure constraint is very tight. As the constraint relaxes, more stocks are selected into the portfolio, and the in-sample risk of the portfolio decreases. However, the risk of the portfolio based on the sample covariance matrix increases out-of-sample, suggesting that the covariance matrix estimates are ill-conditioned and that the allocation becomes noisy and unstable. Both PCA and the regression approach produce stable out-of-sample risk, as the exposure constraint relaxes. For comparison, we also build up an equal-weight portfolio, which is independent of the exposure constraints and the numbers of factors. Its annualized risk is 17.9%.

Fig. 7 further illustrates how the out-of-sample portfolio risk using the PCA approach with the sector-based block-diagonal structure of the residual covariance matrix varies with different number of factors for a variety of exposure constraints. When the number of factors is 0, i.e., the estimator is a block-diagonal thresholded sample covariance matrix, the out-of-sample risk explodes due to the obvious model misspecification (no factor structure). The risk drops rapidly, as soon as a few factors are added. Nonetheless, when tens of factors are included, the risk surges again due to overfitting. The estimator with 500 factors corresponds to the sample covariance matrix estimator (without any truncation), which performs well only when a binding exposure constraint is imposed.

6. Conclusion

We propose a PCA-based estimator of the large covariance matrix from a continuous-time model using high frequency returns. The approach is semiparametric, and relies on a latent factor structure following dynamics represented by arbitrary Itô semimartingales with continuous paths. This includes for instance general forms of stochastic volatility. The estimator is positive-definite by construction and well-conditioned. We also provide an

estimator of the number of latent factors and show consistency of these estimators under dual increasing frequency and dimension asymptotics. Empirically, we document a latent low-rank and sparsity structure in the covariances of the asset returns. A comparison with observable factors shows that the Fama–French factors, the momentum factor, and the industrial portfolios together, approximate the span of the latent factors quite well.

Acknowledgments

Xiu gratefully acknowledges financial support from the Fama–Miller Center for Research in Finance and the IBM Faculty Scholar Fund at the University of Chicago Booth School of Business.

Appendix. Mathematical proofs

A.1. Proof of Theorem 1

Proof of Theorem 1. First, we write $\mathrm{B} = \beta\sqrt{\mathrm{E}}U = (\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_r)$ with $\|b_j\|$s sorted in a descending order, where U is an orthogonal matrix such that $U^\intercal\sqrt{\mathrm{E}}\beta^\intercal\beta\sqrt{\mathrm{E}}U$ is a diagonal matrix. Note that $\left\{\|\mathrm{b}_j\|^2, 1 \le j \le r\right\}$ are the non-zero eigenvalues of $\mathrm{BB}^\intercal$. Therefore by Weyl's inequalities, we have

$$|\lambda_j(\Sigma) - \|\mathrm{b}_j\|^2| \le \|\Gamma\|, 1 \le j \le r; \quad \text{and}$$
$$|\lambda_j(\Sigma)| \le \|\Gamma\|, r+1 \le j \le d.$$

On the other hand, the non-zero eigenvalues of $\mathrm{BB}^\intercal$ are the eigenvalues of $\mathrm{B}^\intercal\mathrm{B}$, and the eigenvalues of $\mathrm{E} = \sqrt{\mathrm{E}}UU^\intercal\sqrt{\mathrm{E}}$ are the eigenvalues of $U^\intercal\mathrm{E}U$. By Weyl's inequalities and Assumption 4, we have, for $1 \le j \le r$,

$$\begin{aligned}\left|d^{-1}\lambda_j(\mathrm{B}^\intercal\mathrm{B}) - \lambda_j(\mathrm{E})\right| &= \left|d^{-1}\lambda_j\left(U^\intercal\sqrt{\mathrm{E}}\beta^\intercal\beta\sqrt{\mathrm{E}}U\right) - \lambda_j(U^\intercal\mathrm{E}U)\right| \\ &\le \|\mathrm{E}\|\,\|U\|^2\left\|d^{-1}\beta^\intercal\beta - \mathrm{I}_r\right\| = o(1).\end{aligned}$$

Therefore, $\|\mathrm{b}_j\|^2 = O(d)$, and $K'd \le \lambda_j(\Sigma) \le Kd$, for $1 \le j \le r$. Since $\|\Gamma\| \le \|\Gamma\|_1 \le Km_d$ and $\lambda_j(\Sigma) \ge \lambda_j(\Gamma)$ for $1 \le j \le d$, it follows that $K' \le \lambda_j(\Sigma) \le Km_d$, for $r+1 \le j \le d$. This implies that $d^{-1}\lambda_j(\Sigma) \ge d^{-1}\lambda_r(\Sigma) \ge K'$, for $1 \le j \le r$; $d^{-1}\lambda_j(\Sigma) \le d^{-1}m_d$, for $r+1 \le j \le d$. Since $d^{-1/2}m_d = o(1)$, it follows that $d^{-1}m_d < d^{-1/2}m_d < K'$. Therefore, we have, as $d \to \infty$:

$$\bar{r} = \arg\min_{1\le j\le d}\left(d^{-1}\lambda_j(\Sigma) + jd^{-1/2}m_d\right) - 1 \to r.$$

Next, by the Sin theta theorem in Davis and Kahan (1970), we have

$$\left\|\xi_j - \frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\| \le \frac{K\|\Gamma\|}{\min\left(\left|\lambda_{j-1}(\Sigma) - \|\mathrm{b}_j\|^2\right|, \left|\lambda_{j+1}(\Sigma) - \|\mathrm{b}_j\|^2\right|\right)}.$$

By the triangle inequality, we have

$$\begin{aligned}\left|\lambda_{j-1}(\Sigma) - \|\mathrm{b}_j\|^2\right| &\ge \left|\|\mathrm{b}_{j-1}\|^2 - \|\mathrm{b}_j\|^2\right| - \left|\lambda_{j-1}(\Sigma) - \|\mathrm{b}_{j-1}\|^2\right| \\ &\ge \left|\|\mathrm{b}_{j-1}\|^2 - \|\mathrm{b}_j\|^2\right| - \|\Gamma\| > Kd,\end{aligned}$$

because for any $1 \le j \le r$, the proof above shows that $\|\mathrm{b}_{j-1}\|^2 - \|\mathrm{b}_j\|^2 = d\left(\lambda_{j-1}(\mathrm{E}) - \lambda_j(\mathrm{E})\right) + o(1)$. Similarly, $\left|\lambda_{j+1}(\Sigma) - \|\mathrm{b}_j\|^2\right| > Kd$, when $j \le r-1$. When $j = r$, we have $\|\mathrm{b}_r\|^2 - \lambda_{j+1}(\Sigma) \ge \|\mathrm{b}_r\|^2 - \|\Gamma\| > Kd$. Therefore, it implies that

$$\left\|\xi_j - \frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\| = O\left(d^{-1}m_d\right), \quad 1 \le j \le r.$$

This, along with the triangle inequality, $\|\mathrm{B}\|_{\mathrm{MAX}} \le \|\beta\|_{\mathrm{MAX}}\left\|\mathrm{E}^{1/2}U\right\|_1 \le K$, and $\|\cdot\|_{\mathrm{MAX}} \le \|\cdot\|$, implies that for $1 \le j \le r$,

$$\|\xi_j\|_{\mathrm{MAX}} \le \left\|\frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}} + O\left(d^{-1}m_d\right) \le O(d^{-1/2}) + O\left(d^{-1}m_d\right).$$

Since $\bar{r} = r$, for d sufficiently large, by triangle inequalities and that $\|\cdot\|_{\mathrm{MAX}} \le \|\cdot\|$ again, we have

$$\begin{aligned}\left\|\sum_{j=1}^r \lambda_j\xi_j\xi_j^\intercal - \mathrm{BB}^\intercal\right\|_{\mathrm{MAX}} &\le \sum_{j=1}^r\|b_j\|^2\left\|\frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}}\left\|\xi_j - \frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}} \\ &\quad + \sum_{j=1}^r|\lambda_j - \|b_j\|^2|\,\|\xi_j\xi_j^\intercal\|_{\mathrm{MAX}} \\ &\quad + \sum_{j=1}^r\|b_j\|^2\|\xi_j\|_{\mathrm{MAX}}\left\|\xi_j - \frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}} \\ &\le Kd^{-1/2}m_d.\end{aligned}$$

Hence, since $\Sigma = \sum_{j=1}^d\lambda_j\xi_j\xi_j^\intercal$, it follows that

$$\left\|\sum_{j=r+1}^d\lambda_j\xi_j\xi_j^\intercal - \Gamma\right\|_{\mathrm{MAX}} \le Kd^{-1/2}m_d,$$

which concludes the proof. □

A.2. Proof of Theorem 2

Throughout the proofs of Theorems 2 to 5, we will impose the assumption that $\|\beta\|_{\mathrm{MAX}}$, $\|\Gamma\|_{\mathrm{MAX}}$, $\|\mathrm{E}\|_{\mathrm{MAX}}$, $\|\mathcal{X}\|_{\mathrm{MAX}}$, $\|\mathcal{Z}\|_{\mathrm{MAX}}$, are bounded by K uniformly across time and dimensions. This is due to Assumption 1, the fact that X and Z are continuous, and the localization argument in Section 4.4.1 of Jacod and Protter (2012).

We need one lemma on the concentration inequalities for continuous Itô semimartingales.

Lemma 1. *Suppose Assumptions 1 and 2 hold, then we have*

$$\text{(i)}\ \max_{1\le l,k\le d}\left|\sum_{i=1}^{[T/\Delta_n]}(\Delta_i^nZ_l)(\Delta_i^nZ_k) - \int_0^T g_{s,lk}ds\right| = O_p\left((\Delta_n\log d)^{1/2}\right), \tag{A.1}$$

$$\text{(ii)}\ \max_{1\le j\le r,1\le l\le d}\left|\sum_{i=1}^{[T/\Delta_n]}(\Delta_i^nX_j)(\Delta_i^nZ_l)\right| = O_p\left((\Delta_n\log d)^{1/2}\right), \tag{A.2}$$

$$\begin{aligned}\text{(iii)}\ &\max_{1\le j\le r,1\le l\le r}\left|\sum_{i=1}^{[T/\Delta_n]}(\Delta_i^nX_j)(\Delta_i^nX_l) - \int_0^T e_{s,jl}ds\right| \\ &= O_p\left((\Delta_n\log d)^{1/2}\right).\end{aligned} \tag{A.3}$$

Proof of Lemma 1. The proof of this lemma follows by (i), (iii), (iv) of Lemma 2 in Fan et al. (2016). □

Proof of Theorem 2. We first recall some notation introduced in the main text. Let $n = [T/\Delta_n]$. Suppose that $\mathcal{Y} = \left(\Delta_1^nY, \Delta_2^nY, \ldots, \Delta_n^nY\right)$ is a $d \times n$ matrix, where $\Delta_i^nY = Y_{i\Delta_n} - Y_{(i-1)\Delta_n}$. Similarly, $\mathcal{X}$ and $\mathcal{Z}$ are $r\times n$ and $d\times n$ matrices, respectively. Therefore, we have $\mathcal{Y} = \beta\mathcal{X} + \mathcal{Z}$ and $\widehat{\Sigma} = T^{-1}\mathcal{Y}\mathcal{Y}^\intercal$. Let $f(j) = d^{-1}\lambda_j(\widehat{\Sigma}) + j\times g(n,d)$. Suppose $\mathrm{R} = \{j|1 \le j \le k_{\max}, j \ne r\}$.

Note that using $\|\beta\| \le r^{1/2}d^{1/2}\|\beta\|_{\mathrm{MAX}} = O(d^{1/2})$ and $\|\Gamma\|_\infty \le Km_d$ we have

$$\begin{aligned}\|\mathcal{Y}\mathcal{Y}^\intercal - \beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal\| &\le \|\mathcal{Z}\mathcal{X}^\intercal\beta^\intercal\| + \|\beta\mathcal{X}\mathcal{Z}^\intercal\| + \|\mathcal{Z}\mathcal{Z}^\intercal - \Gamma\| + \|\Gamma\| \\ &\le Kd^{1/2}\|\beta\|\,\|\mathcal{Z}\mathcal{X}^\intercal\|_{\mathrm{MAX}} + d\,\|\mathcal{Z}\mathcal{Z}^\intercal - \Gamma\|_{\mathrm{MAX}}\end{aligned}$$

$$+ \|\Gamma\|_\infty$$
$$= O_p\left(d(\Delta_n \log d)^{1/2} + m_d\right).$$

where we use the following bounds, implied by Lemma 1:

$$\|\mathcal{Z}\mathcal{Z}^\intercal - \Gamma\|_{\text{MAX}} = \max_{1\le k,l\le d}\left(\left|\sum_{i=1}^n (\Delta_i^n Z_l)(\Delta_i^n Z_k) - \int_0^T g_{s,lk}ds\right|\right)$$
$$= O_p((\Delta_n \log d)^{1/2}), \text{ and}$$
$$\|\mathcal{Z}\mathcal{X}^\intercal\|_{\text{MAX}} = O_p((\Delta_n \log d)^{1/2}).$$

Therefore, by Weyl's inequality we have for $1 \le j \le r$,

$$|\lambda_j(\widehat{\Sigma}) - \lambda_j(T^{-1}\beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal)| = O_p\left(d(\Delta_n \log d)^{1/2} + m_d\right).$$

On the other hand, the non-zero eigenvalues of $T^{-1}\beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal$ are identical to the eigenvalues of $T^{-1}\sqrt{\mathcal{X}\mathcal{X}^\intercal}\beta^\intercal\beta\sqrt{\mathcal{X}\mathcal{X}^\intercal}$. By Weyl's inequality again, we have for $1 \le j \le r$,

$$\left|d^{-1}\lambda_j\left(T^{-1}\sqrt{\mathcal{X}\mathcal{X}^\intercal}\beta^\intercal\beta\sqrt{\mathcal{X}\mathcal{X}^\intercal}\right) - \lambda_j(T^{-1}\mathcal{X}\mathcal{X}^\intercal)\right|$$
$$\le T^{-1}\|\mathcal{X}\mathcal{X}^\intercal\|\left\|d^{-1}\beta^\intercal\beta - \mathrm{I}_r\right\| = o_p(1),$$

where we use

$$\|\mathcal{X}\| = \sqrt{\lambda_{\max}(\mathcal{X}\mathcal{X}^\intercal)} \le r^{1/2}\max_{1\le l,j\le r}\left|\sum_{i=1}^n (\Delta_i^n X_l)(\Delta_i^n X_j)\right|^{1/2}$$
$$= O_p(1). \tag{A.4}$$

Also, for $1 \le j \le r$, by Weyl's inequality and Lemma 1, we have

$$|\lambda_j(T^{-1}\mathcal{X}\mathcal{X}^\intercal) - \lambda_j(\mathrm{E})| \le \left\|T^{-1}\mathcal{X}\mathcal{X}^\intercal - \mathrm{E}\right\| = O_p\left((\Delta_n \log d)^{1/2}\right).$$

Combining the above inequalities, we have for $1 \le j \le r$,

$$|d^{-1}\lambda_j(\widehat{\Sigma}) - \lambda_j(\mathrm{E})| \le O_p\left((\Delta_n \log d)^{1/2} + d^{-1}m_d\right) + o_p(1).$$

Therefore, for $1 \le j < r$, we have

$$\lambda_{j+1}(\mathrm{E}) - o_p(1) < d^{-1}\lambda_{j+1}(\widehat{\Sigma}) < \lambda_{j+1}(\mathrm{E}) + o_p(1)$$
$$< \lambda_j(\mathrm{E}) - o_p(1) < d^{-1}\lambda_j(\widehat{\Sigma}). \tag{A.5}$$

Next, note that

$$\mathcal{Y}\mathcal{Y}^\intercal = \tilde{\beta}\mathcal{X}\mathcal{X}^\intercal\tilde{\beta}^\intercal + \mathcal{Z}\left(\mathrm{I}_n - \mathcal{X}^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}\mathcal{X}\right)\mathcal{Z}^\intercal$$

where $\tilde{\beta} = \beta + \mathcal{Z}\mathcal{X}^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}$. Since $\text{rank}(\tilde{\beta}\mathcal{X}\mathcal{X}^\intercal\tilde{\beta}^\intercal) = r$, and by (4.3.2a) of Theorem 4.3.1 and (4.3.14) of Corollary 4.3.12 in Horn and Johnson (2013), we have for $r + 1 \le j \le d$,

$$\lambda_j(\mathcal{Y}\mathcal{Y}^\intercal) \le \lambda_{j-r}\left(\mathcal{Z}\left(\mathrm{I}_n - \mathcal{X}^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}\mathcal{X}\right)\mathcal{Z}^\intercal\right) + \lambda_{r+1}(\tilde{\beta}\mathcal{X}\mathcal{X}^\intercal\tilde{\beta}^\intercal)$$
$$\le \lambda_{j-r}(\mathcal{Z}\mathcal{Z}^\intercal) \le \lambda_1(\mathcal{Z}\mathcal{Z}^\intercal).$$

Since by Lemma 1 we have

$$\lambda_1(\mathcal{Z}\mathcal{Z}^\intercal) = \|\mathcal{Z}\mathcal{Z}^\intercal\| \le \|\mathcal{Z}\mathcal{Z}^\intercal\|_\infty$$
$$\le \max_{1\le j,l\le d}\left\{d|(\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)_{jl}| + m_d|\Gamma_{jl}|\right\}$$
$$= O_p(d(\Delta_n \log d)^{1/2} + m_d), \tag{A.6}$$

it thus implies that for $r + 1 \le j \le d$, there exists some $K > 0$, such that

$$d^{-1}\lambda_j(\widehat{\Sigma}) \le K(\Delta_n \log d)^{1/2} + Kd^{-1}m_d.$$

In sum, for $1 \le j \le r$,

$$f(j) - f(r+1) = d^{-1}\left(\lambda_j(\widehat{\Sigma}) - \lambda_{r+1}(\widehat{\Sigma})\right) + (j - r - 1)g(n,d)$$
$$> \lambda_j(\mathrm{E}) + o_p(1) > K,$$

for some $K > 0$. Since $g(n,d)\left((\Delta_n \log d)^{1/2} + d^{-1}m_d\right)^{-1} \to \infty$, it follows that for $r + 1 < j \le d$,

$$\mathbb{P}(f(j) < f(r+1))$$
$$= \mathbb{P}\left((j - r - 1)g(n,d) < d^{-1}\left(\lambda_{r+1}(\widehat{\Sigma}) - \lambda_j(\widehat{\Sigma})\right)\right) \to 0.$$

This establishes the desired result. □

A.3. Proof of Theorem 3

First, we can assume $\widehat{r} = r$. Since it holds with probability approaching 1 as established by Theorem 2, a simple conditioning argument, see, e.g., footnote 5 of Bai (2003), is sufficient to show this is without loss of rigor. Recall that

$$\Lambda = \text{Diag}\left(\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_r\right), \quad F = d^{1/2}\left(\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_r\right), \quad \text{and}$$
$$G = d^{-1}F^\intercal\mathcal{Y}.$$

We write

$$H = T^{-1}\mathcal{X}\mathcal{X}^\intercal\beta^\intercal F\Lambda^{-1}.$$

It is easy to verify that

$$\widehat{\Sigma}F = F\Lambda, \quad GG^\intercal = Td^{-1} \times \Lambda, \quad F^\intercal F = d \times \mathrm{I}_r, \quad \text{and}$$
$$\widehat{\Gamma} = T^{-1}(\mathcal{Y} - FG)(\mathcal{Y} - FG)^\intercal = T^{-1}\mathcal{Y}\mathcal{Y}^\intercal - d^{-1}F\Lambda F^\intercal.$$

We now need a few more lemmas. The proofs of these lemmas rely on similar arguments to those developed in Doz et al. (2011) and Fan et al. (2013).

Lemma 2. *Under Assumptions 1–4, $d^{-1/2}m_d = o(1)$, and $\Delta_n \log d = o(1)$, we have*

$$\text{(i)}\ \|F - \beta H\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right). \tag{A.7}$$
$$\text{(ii)}\ \left\|H^{-1}\right\| = O_p(1). \tag{A.8}$$
$$\text{(iii)}\ \left\|G - H^{-1}\mathcal{X}\right\| = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right). \tag{A.9}$$

Proof of Lemma 2. (i) By simple calculations, we have

$$F - \beta H = T^{-1}\left(\mathcal{Y}\mathcal{Y}^\intercal - \beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal\right)F\Lambda^{-1}$$
$$= T^{-1}\left(\beta\mathcal{X}\mathcal{Z}^\intercal F\Lambda^{-1} + \mathcal{Z}\mathcal{X}^\intercal\beta^\intercal F\Lambda^{-1}\right.$$
$$\left. + (\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)F\Lambda^{-1} + \Gamma F\Lambda^{-1}\right). \tag{A.10}$$

We bound these terms separately. First, we have

$$\left\|(\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)F\Lambda^{-1}\right\|_{\text{MAX}} \le \|\mathcal{Z}\mathcal{Z}^\intercal - \Gamma\|_{\text{MAX}}\|F\|_1\left\|\Lambda^{-1}\right\|_{\text{MAX}}.$$

Moreover, $\|F\|_1 \le d^{1/2}\|F\|_{\text{F}} = d$, and by (A.5), $\left\|\Lambda^{-1}\right\|_{\text{MAX}} = O_p(d^{-1})$, which implies that

$$\left\|(\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)F\Lambda^{-1}\right\|_{\text{MAX}} = O_p((\Delta_n \log d)^{1/2}).$$

In addition, since $\|\Gamma\|_\infty \le Km_d$ and $\|F\|_{\text{MAX}} \le \|F\|_{\text{F}} = d^{1/2}$, it follows that

$$\left\|\Gamma F\Lambda^{-1}\right\|_{\text{MAX}} \le \|\Gamma\|_\infty\|F\|_{\text{MAX}}\left\|\Lambda^{-1}\right\|_{\text{MAX}} = O_p(d^{-1/2}m_d).$$

Also, we have

$$\left\|\beta\mathcal{X}\mathcal{Z}^\intercal F\Lambda^{-1}\right\|_{\text{MAX}} \le \|\beta\|_{\text{MAX}}\|\mathcal{X}\mathcal{Z}^\intercal\|_1\|F\|_1\left\|\Lambda^{-1}\right\|_{\text{MAX}}$$
$$= O_p((\Delta_n \log d)^{1/2}).$$

where we use the fact that $\|\beta\|_{\text{MAX}} \le K$ and the bound below derived from (A.2):

$$\|\mathcal{X}\mathcal{Z}^\intercal\|_1 = \max_{1\le l\le d}\sum_{j=1}^r\left|\sum_{i=1}^n (\Delta_i^n X_j)(\Delta_i^n Z_l)\right|$$
$$\le r\max_{1\le l\le d,\,1\le j\le r}\left|\sum_{i=1}^n (\Delta_i^n X_j)(\Delta_i^n Z_l)\right| = O_p((\Delta_n \log d)^{1/2}).$$

The remainder term can be bounded similarly.

(ii) Since $\|\beta\| = O(d^{1/2})$ and $\left\|T^{-1}\mathcal{X}\mathcal{X}^\intercal\right\| = O_p(1)$, we have

$$\|H\| = \left\|T^{-1}\mathcal{X}\mathcal{X}^\intercal\beta^\intercal F\Lambda^{-1}\right\| \le \left\|T^{-1}\mathcal{X}\mathcal{X}^\intercal\right\|\|\beta\|\|F\|\left\|\Lambda^{-1}\right\| = O_p(1).$$

By triangle inequalities, and that $\|F-\beta H\| \le (rd)^{1/2}\|F-\beta H\|_{\max}$, we have

$$\begin{aligned}\|H^\top H - \mathrm{I}_r\| &\le \left\|H^\top H - d^{-1}H^\top\beta^\top\beta H\right\| + d^{-1}\|H^\top\beta^\top\beta H - d\mathrm{I}_r\| \\ &\le \|H\|^2\left\|\mathrm{I}_r - d^{-1}\beta^\top\beta\right\| + d^{-1}\|H^\top\beta^\top\beta H - F^\top F\| \\ &\le \|H\|^2\left\|\mathrm{I}_r - d^{-1}\beta^\top\beta\right\| + d^{-1}\|F-\beta H\|\,\|\beta H\| \\ &\quad + d^{-1}\|F-\beta H\|\,\|F\| \\ &= o_p(1).\end{aligned}$$

By Weyl's inequality again, we have $\lambda_{\min}(H^\top H) > 1/2$ with probability approaching 1. Therefore, H is invertible, and $\left\|H^{-1}\right\| = O_p(1)$.

(iii) We use the following decomposition:

$$\begin{aligned}G - H^{-1}\mathcal{X} &= d^{-1}F^\top(\beta H - F)H^{-1}\mathcal{X} + d^{-1}(F^\top - H^\top\beta^\top)\mathcal{Z} \\ &\quad + d^{-1}H^\top\beta^\top\mathcal{Z}.\end{aligned}$$

Note that by (A.4), we have $\|\mathcal{X}\| = O_p(1)$. Moreover, since $\|F\| \le \|F\|_{\mathrm{F}}$ and $\|F-\beta H\| \le r^{1/2}d^{1/2}\|F-\beta H\|_{\max}$, we have

$$\begin{aligned}\left\|d^{-1}F^\top(\beta H - F)H^{-1}\mathcal{X}\right\| &\le d^{-1}\|F\|\,\|F-\beta H\|\left\|H^{-1}\right\|\|\mathcal{X}\| \\ &= O_p\left((\Delta_n\log d)^{1/2} + d^{-1/2}m_d\right).\end{aligned}$$

Similarly, by (A.6) we have

$$\|\mathcal{Z}\| = O_p(d^{1/2}(\Delta_n\log d)^{1/4} + m_d^{1/2}),$$

which leads to

$$\begin{aligned}\left\|d^{-1}(F^\top - H^\top\beta^\top)\mathcal{Z}\right\| = O_p\Big(&\left((\Delta_n\log d)^{1/4} + d^{-1/2}m_d^{1/2}\right) \\ &\times\ \left((\Delta_n\log d)^{1/2} + d^{-1/2}m_d\right)\Big).\end{aligned}$$

Moreover, we can apply Lemma 1 to $\beta^\top\mathcal{Z}$, which is an $r\times n$ matrix, so we have

$$\begin{aligned}\|\beta^\top\mathcal{Z}\| &= \sqrt{\|\beta^\top\mathcal{Z}\mathcal{Z}^\top\beta\|} \le \sqrt{\|\beta^\top\mathcal{Z}\mathcal{Z}^\top\beta\|_\infty} \\ &\le \sqrt{\|\beta^\top\mathcal{Z}\mathcal{Z}^\top\beta - \beta^\top\Gamma\beta\|_\infty + \|\Gamma\|_\infty\|\beta\|_\infty\|\beta\|_1} \\ &\le K(\Delta_n\log d)^{1/4} + Km_d^{1/2}d^{1/2},\end{aligned}$$

where we use $\|\beta\|_\infty \le r\|\beta\|_{\max}$ and $\|\beta\|_1 \le d\|\beta\|_{\max}$. This leads to

$$\left\|d^{-1}H^\top\beta^\top\mathcal{Z}\right\| = O_p\left(d^{-1}(\Delta_n\log d)^{1/4} + d^{-1/2}m_d^{1/2}\right).$$

This concludes the proof. □

Lemma 3. *Under Assumptions 1–4, $d^{-1/2}m_d = o(1)$, and $\Delta_n\log d = o(1)$, we have*

$$\begin{aligned}\left\|\widehat{\Gamma}^S - \Gamma\right\|_{\max} &\le \left\|\widehat{\Gamma} - \Gamma\right\|_{\max} \\ &= O_p\left((\Delta_n\log d)^{1/2} + d^{-1/2}m_d\right).\end{aligned} \tag{A.11}$$

Proof of Lemma 3. We write $G = (g_1, g_2, \ldots, g_n)$, $F = (f_1, f_2, \ldots, f_d)^\top$, $\beta = (\beta_1, \beta_2, \ldots, \beta_d)^\top$, and $\widehat{\Delta_i^n Z_k} = \Delta_i^n Y_k - f_k^\top g_i$. Hence, $\widehat{\Gamma}_{lk} = T^{-1}\sum_{i=1}^n(\widehat{\Delta_i^n Z_l})(\widehat{\Delta_i^n Z_k})$.

For $1\le k\le d$ and $1\le i\le n$, we have

$$\begin{aligned}\Delta_i^n Z_k - \widehat{\Delta_i^n Z_k} &= \Delta_i^n Y_k - \beta_k^\top\Delta_i^n X - (\Delta_i^n Y_k - f_k^\top g_i) = f_k^\top g_i - \beta_k^\top\Delta_i^n X \\ &= \beta_k^\top H(g_i - H^{-1}\Delta_i^n X) + (f_k^\top - \beta_k^\top H)(g_i - H^{-1}\Delta_i^n X) \\ &\quad + (f_k^\top - \beta_k^\top H)H^{-1}\Delta_i^n X.\end{aligned}$$

Therefore, using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we have

$$\begin{aligned}&\sum_{i=1}^n\left(\Delta_i^n Z_k - \widehat{\Delta_i^n Z_k}\right)^2 \\ \le&3\sum_{i=1}^n\left(\beta_k^\top H(g_i - H^{-1}\Delta_i^n X)\right)^2 \\ &+ 3\sum_{i=1}^n\left((f_k^\top - \beta_k^\top H)(g_i - H^{-1}\Delta_i^n X)\right)^2 \\ &+ 3\sum_{i=1}^n\left((f_k^\top - \beta_k^\top H)H^{-1}\Delta_i^n X\right)^2.\end{aligned}$$

Using $v^\top Av \le \lambda_{\max}(A)v^\top v$ repeatedly, it follows that

$$\begin{aligned}&\sum_{i=1}^n\left(\beta_k^\top H(g_i - H^{-1}\Delta_i^n X)\right)^2 \\ &= \sum_{i=1}^n\beta_k^\top H(G - H^{-1}\mathcal{X})e_i e_i^\top(G - H^{-1}\mathcal{X})^\top H^\top\beta_k \\ &\le \lambda_{\max}\left(\left(G - H^{-1}\mathcal{X}\right)\left(G - H^{-1}\mathcal{X}\right)^\top\right)\lambda_{\max}(HH^\top)\beta_k^\top\beta_k \\ &\le r\left\|G - H^{-1}\mathcal{X}\right\|^2\|H\|^2\max_{1\le l\le r}|\beta_{kl}|^2\end{aligned}$$

Similarly, we can bound the other terms.

$$\begin{aligned}&\sum_{i=1}^n\left((f_k^\top - \beta_k^\top H)(g_i - H^{-1}\Delta_i^n X)\right)^2 \\ &\quad\le r\left\|G - H^{-1}\mathcal{X}\right\|^2\max_{1\le l\le r}(F_{kl} - (\beta_k^\top H)_l)^2, \\ &\sum_{i=1}^n\left((f_k^\top - \beta_k^\top H)H^{-1}\Delta_i^n X\right)^2 \\ &\quad\le rT\|\mathrm{E}\|\left\|H^{-1}\right\|^2\max_{1\le l\le r}(F_{kl} - (\beta_k^\top H)_l)^2.\end{aligned}$$

As a result, by Lemma 2, we have

$$\begin{aligned}&\max_{1\le k\le d}\sum_{i=1}^n\left(\Delta_i^n Z_k - \widehat{\Delta_i^n Z_k}\right)^2 \\ \le&K\left\|G - H^{-1}\mathcal{X}\right\|^2\|H\|^2\|\beta\|_{\max}^2 + K\left\|G - H^{-1}\mathcal{X}\right\|^2\|F-\beta H\|_{\max}^2 \\ &+ K\|\mathrm{E}\|\left\|H^{-1}\right\|^2\|F-\beta H\|_{\max}^2 \\ \le&O_p\left((\Delta_n\log d) + d^{-1}m_d^2\right)\end{aligned}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}&\max_{1\le l,k\le d}\left|\sum_{i=1}^n(\widehat{\Delta_i^n Z_l})(\widehat{\Delta_i^n Z_k}) - \sum_{i=1}^n(\Delta_i^n Z_l)(\Delta_i^n Z_k)\right| \\ \le&\max_{1\le l,k\le d}\left|\sum_{i=1}^n\left(\widehat{\Delta_i^n Z_l} - \Delta_i^n Z_l\right)\left(\widehat{\Delta_i^n Z_k} - \Delta_i^n Z_k\right)\right| \\ &+ 2\max_{1\le l,k\le d}\left|\sum_{i=1}^n\left(\Delta_i^n Z_l\right)\left(\widehat{\Delta_i^n Z_k} - \Delta_i^n Z_k\right)\right| \\ \le&\max_{1\le l\le d}\sum_{i=1}^n\left(\widehat{\Delta_i^n Z_l} - \Delta_i^n Z_l\right)^2 \\ &+ 2\sqrt{\max_{1\le l\le d}\sum_{i=1}^n(\Delta_i^n Z_l)^2\max_{1\le l\le d}\sum_{i=1}^n\left(\widehat{\Delta_i^n Z_l} - \Delta_i^n Z_l\right)^2} \\ =&O_p\left((\Delta_n\log d)^{1/2} + d^{-1/2}m_d\right),\end{aligned}$$

Finally, by the triangular inequality,

$$\begin{aligned}\max_{1\le l,k\le d,(l,k)\in S}\left|\widehat{\Gamma}_{lk}-\Gamma_{lk}\right| &\le \max_{1\le l,k\le d}\left|\widehat{\Gamma}_{lk}-\Gamma_{lk}\right| \\ &\le \max_{1\le l,k\le d}\left|\sum_{i=1}^{n}(\Delta_i^n Z_l)(\Delta_i^n Z_k)-\int_0^T g_{s,lk}ds\right| \\ &\quad+\max_{1\le l,k\le d}\left|\sum_{i=1}^{n}(\widehat{\Delta_i^n Z_l})(\widehat{\Delta_i^n Z_k})-\sum_{i=1}^{n}(\Delta_i^n Z_l)(\Delta_i^n Z_k)\right|,\end{aligned}$$

which yields the desired result by using (A.1). □

Lemma 4. *Under Assumptions 1–4, $d^{-1/2}m_d = o(1)$, and $\Delta_n \log d = o(1)$, we have*

$$\left\|T^{-1}FGG^{\intercal}F^{\intercal}-\beta\mathrm{E}\beta^{\intercal}\right\|_{\mathrm{MAX}} = O_p\left((\Delta_n\log d)^{1/2}+d^{-1/2}m_d\right).$$

Proof. Using $GG^{\intercal}=Td^{-1}\times\Lambda$, we can write

$$\begin{aligned}&T^{-1}FGG^{\intercal}F^{\intercal}\\ =\;& d^{-1}F\Lambda F^{\intercal}=d^{-1}(F-\beta H+\beta H)\Lambda(F-\beta H+\beta H)^{\intercal}\\ =\;& d^{-1}(F-\beta H)\Lambda(F-\beta H)^{\intercal}+d^{-1}\beta H\Lambda(F-\beta H)^{\intercal}\\ &+d^{-1}(\beta H\Lambda(F-\beta H)^{\intercal})^{\intercal}+d^{-1}\beta H\Lambda H^{\intercal}\beta^{\intercal}.\end{aligned}$$

Moreover, we can derive

$$\begin{aligned}&d^{-1}\beta H\Lambda H^{\intercal}\beta^{\intercal}=T^{-1}\beta HGG^{\intercal}H^{\intercal}\beta^{\intercal}\\ =\;& T^{-1}\beta H(G-H^{-1}\mathcal{X}+H^{-1}\mathcal{X})(G-H^{-1}\mathcal{X}+H^{-1}\mathcal{X})^{\intercal}H^{\intercal}\beta^{\intercal}\\ =\;& T^{-1}\beta H(G-H^{-1}\mathcal{X})(G-H^{-1}\mathcal{X})^{\intercal}H^{\intercal}\beta^{\intercal}\\ &+T^{-1}\beta H(G-H^{-1}\mathcal{X})\mathcal{X}^{\intercal}\beta^{\intercal}\\ &+T^{-1}(\beta H(G-H^{-1}\mathcal{X})\mathcal{X}^{\intercal}\beta^{\intercal})^{\intercal}+T^{-1}\beta\mathcal{X}\mathcal{X}^{\intercal}\beta^{\intercal}.\end{aligned}$$

Therefore, combining the above equalities and applying the triangular inequality, we obtain

$$\begin{aligned}&\left\|T^{-1}FGG^{\intercal}F^{\intercal}-\beta\mathrm{E}\beta^{\intercal}\right\|_{\mathrm{MAX}}\\ &\le d^{-1}\left\|(F-\beta H)\Lambda(F-\beta H)^{\intercal}\right\|_{\mathrm{MAX}}\\ &\quad+2d^{-1}\left\|\beta H\Lambda(F-\beta H)^{\intercal}\right\|_{\mathrm{MAX}}\\ &\quad+T^{-1}\left\|\beta H(G-H^{-1}\mathcal{X})(G-H^{-1}\mathcal{X})^{\intercal}H^{\intercal}\beta^{\intercal}\right\|_{\mathrm{MAX}}\\ &\quad+2T^{-1}\left\|\beta H(G-H^{-1}\mathcal{X})\mathcal{X}^{\intercal}\beta^{\intercal}\right\|_{\mathrm{MAX}}\\ &\quad+\left\|\beta(T^{-1}\mathcal{X}\mathcal{X}^{\intercal}-\mathrm{E})\beta^{\intercal}\right\|_{\mathrm{MAX}}.\end{aligned}$$

Note that by Lemma 2, (A.4), $\|\beta\|_{\mathrm{MAX}}=O_p(1)$, $\|H\|=O_p(1)$, and $\|\Lambda\|_{\mathrm{MAX}}=O_p(1)$,

$$\begin{aligned}&d^{-1}\left\|(F-\beta H)\Lambda(F-\beta H)^{\intercal}\right\|_{\mathrm{MAX}}\\ &\le r^2d^{-1}\|F-\beta H\|^2_{\mathrm{MAX}}\|\Lambda\|_{\mathrm{MAX}}\\ &\le O_p(\Delta_n\log d+d^{-1}m_d^2),\\ &2d^{-1}\left\|\beta H\Lambda(F-\beta H)^{\intercal}\right\|_{\mathrm{MAX}}\\ &\le 2r^2d^{-1}\|\beta\|_{\mathrm{MAX}}\|H\|\|\Lambda\|_{\mathrm{MAX}}\|F-\beta H\|_{\mathrm{MAX}}\\ &\le O_p\left((\Delta_n\log d)^{1/2}+d^{-1/2}m_d\right),\\ &T^{-1}\left\|\beta H(G-H^{-1}\mathcal{X})(G-H^{-1}\mathcal{X})^{\intercal}H^{\intercal}\beta^{\intercal}\right\|_{\mathrm{MAX}}\\ &\le r^4T^{-1}\|\beta\|^2_{\mathrm{MAX}}\|H\|^2\left\|G-H^{-1}\mathcal{X}\right\|^2\\ &\le O_p(\Delta_n\log d+d^{-1}m_d^2),\\ &2T^{-1}\left\|\beta H(G-H^{-1}\mathcal{X})\mathcal{X}^{\intercal}\beta^{\intercal}\right\|_{\mathrm{MAX}}\\ &\le r^3\|\beta\|^2_{\mathrm{MAX}}\|H\|\left\|G-H^{-1}\mathcal{X}\right\|\|\mathcal{X}\|\\ &\le O_p\left((\Delta_n\log d)^{1/2}+d^{-1/2}m_d\right),\\ &\left\|\beta(T^{-1}\mathcal{X}\mathcal{X}^{\intercal}-\mathrm{E})\beta^{\intercal}\right\|_{\mathrm{MAX}}\\ &\le r^2\|\beta\|_{\mathrm{MAX}}\left\|T^{-1}\mathcal{X}\mathcal{X}^{\intercal}-\mathrm{E}\right\|_{\mathrm{MAX}}\\ &\le O_p\left((\Delta_n\log d)^{1/2}\right).\end{aligned}$$

Combining the above inequalities concludes the proof. □

Proof of Theorem 3. Note that

$$\widehat{\Sigma}^S=d^{-1}F\Lambda F^{\intercal}+\widehat{\Gamma}^S=T^{-1}FGG^{\intercal}F^{\intercal}+\widehat{\Gamma}^S.$$

By Lemma 3, we have

$$\left\|\widehat{\Gamma}^S-\Gamma\right\|_{\mathrm{MAX}}=O_p\left((\Delta_n\log d)^{1/2}+d^{-1/2}m_d\right).$$

By the triangle inequality, we have

$$\left\|\widehat{\Sigma}^S-\Sigma\right\|_{\mathrm{MAX}}\le\left\|d^{-1}F\Lambda F^{\intercal}-\beta\mathrm{E}\beta^{\intercal}\right\|_{\mathrm{MAX}}+\left\|\widehat{\Gamma}^S-\Gamma\right\|_{\mathrm{MAX}}$$

Therefore, the desired result follows from Lemmas 3 and 4. □

A.4. Proof of Theorem 4

Lemma 5. *Under Assumptions 1–4, $d^{-1/2}m_d=o(1)$, and $\Delta_n\log d=o(1)$, we have*

$$\left\|\widehat{\Gamma}^S-\Gamma\right\|=O_p\left(m_d(\Delta_n\log d)^{1/2}+d^{-1/2}m_d^2\right). \tag{A.12}$$

Moreover, if in addition, $d^{-1/2}m_d^2=o(1)$ and $m_d(\Delta_n\log d)^{1/2}=o(1)$ hold, then $\lambda_{\min}\left(\widehat{\Gamma}^S\right)$ is bounded away from 0 with probability approaching 1, and

$$\left\|\left(\widehat{\Gamma}^S\right)^{-1}-\Gamma^{-1}\right\|=O_p\left(m_d(\Delta_n\log d)^{1/2}+d^{-1/2}m_d^2\right).$$

Proof of Lemma 5. Note that since $\widehat{\Gamma}^S-\Gamma$ is symmetric,

$$\begin{aligned}\left\|\widehat{\Gamma}^S-\Gamma\right\|\le\left\|\widehat{\Gamma}^S-\Gamma\right\|_{\infty}&=\max_{1\le l\le d}\sum_{k=1}^{d}\left|\widehat{\Gamma}^S_{lk}-\Gamma_{lk}\right|\\ &\le m_d\max_{1\le l\le d,1\le k\le d}\left|\widehat{\Gamma}^S_{lk}-\Gamma_{lk}\right|\end{aligned}$$

By Lemma 3, we have

$$\begin{aligned}\left\|\widehat{\Gamma}^S-\Gamma\right\|&\le m_d\left\|\widehat{\Gamma}^S-\Gamma\right\|_{\mathrm{MAX}}\\ &=O_p\left(m_d(\Delta_n\log d)^{1/2}+d^{-1/2}m_d^2\right).\end{aligned}$$

Moreover, since $\lambda_{\min}(\Gamma)\ge K$ for some constant K and by Weyl's inequality, we have $\lambda_{\min}(\widehat{\Gamma}^S)>K-o_p(1)$. As a result, we have

$$\begin{aligned}\left\|\left(\widehat{\Gamma}^S\right)^{-1}-\Gamma^{-1}\right\|&=\left\|\left(\widehat{\Gamma}^S\right)^{-1}\left(\Gamma-\left(\widehat{\Gamma}^S\right)\right)\Gamma^{-1}\right\|\\ &\le\lambda_{\min}(\widehat{\Gamma}^S)^{-1}\lambda_{\min}(\Gamma)^{-1}\left\|\Gamma-\widehat{\Gamma}^S\right\|\\ &\le O_p\left(m_d(\Delta_n\log d)^{1/2}+d^{-1/2}m_d^2\right).\end{aligned}$$

□

Proof of Theorem 4. First, by Lemma 5 and the fact that $\lambda_{\min}(\widehat{\Sigma}^S)\ge\lambda_{\min}(\widehat{\Gamma}^S)$, we can establish the first two statements.

To bound $\left\|(\widehat{\Sigma}^S)^{-1}-\Sigma^{-1}\right\|$, by the Sherman–Morrison–Woodbury formula, we have

$$\begin{aligned}&\left(\widehat{\Sigma}^S\right)^{-1}-\left(\widetilde{\Sigma}\right)^{-1}\\ =\;&\left(T^{-1}FGG^{\intercal}F^{\intercal}+\widehat{\Gamma}^S\right)^{-1}-\left(T^{-1}\beta HH^{-1}\mathcal{X}\mathcal{X}^{\intercal}(H^{-1})^{\intercal}H^{\intercal}\beta^{\intercal}+\Gamma\right)^{-1}\\ =\;&\left((\widehat{\Gamma}^S)^{-1}-\Gamma^{-1}\right)-\left((\widehat{\Gamma}^S)^{-1}-\Gamma^{-1}\right)\\ &\times F\left(d\Lambda^{-1}+F^{\intercal}(\widehat{\Gamma}^S)^{-1}F\right)^{-1}F^{\intercal}(\widehat{\Gamma}^S)^{-1}\\ &-\Gamma^{-1}F\left(d\Lambda^{-1}+F^{\intercal}(\widehat{\Gamma}^S)^{-1}F\right)^{-1}F^{\intercal}\left((\widehat{\Gamma}^S)^{-1}-\Gamma^{-1}\right)\\ &+\Gamma^{-1}(\beta H-F)\left(TH^{\intercal}(\mathcal{X}\mathcal{X}^{\intercal})^{-1}H+H^{\intercal}\beta^{\intercal}\Gamma^{-1}\beta H\right)^{-1}H^{\intercal}\beta^{\intercal}\Gamma^{-1}\\ &-\Gamma^{-1}F\left(TH^{\intercal}(\mathcal{X}\mathcal{X}^{\intercal})^{-1}H+H^{\intercal}\beta^{\intercal}\Gamma^{-1}\beta H\right)^{-1}(F^{\intercal}-H^{\intercal}\beta^{\intercal})\Gamma^{-1}\\ &+\Gamma^{-1}F\left(\left(TH^{\intercal}(\mathcal{X}\mathcal{X}^{\intercal})^{-1}H+H^{\intercal}\beta^{\intercal}\Gamma^{-1}\beta H\right)^{-1}\right.\\ &\left.-\left(d\Lambda^{-1}+F^{\intercal}(\widehat{\Gamma}^S)^{-1}F\right)^{-1}\right)F^{\intercal}\Gamma^{-1}\\ =\;&L_1+L_2+L_3+L_4+L_5+L_6.\end{aligned}$$

By Lemma 5, we have

$$\|L_1\| = O_p\left(m_d(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^2\right).$$

For L_2, because $\|F\| = O_p(d^{1/2})$, $\lambda_{\max}\left((\widehat{\Gamma}^S)^{-1}\right) \leq \left(\lambda_{\min}(\widehat{\Gamma}^S)\right)^{-1} \leq K + o_p(1)$,

$$\begin{aligned}\lambda_{\min}\left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right) &\geq \lambda_{\min}\left(F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)\\ &\geq \lambda_{\min}\left(F^\intercal F\right)\lambda_{\min}\left((\widehat{\Gamma}^S)^{-1}\right) \geq m_d^{-1}d,\end{aligned}$$

and by Lemma 5, we have

$$\begin{aligned}\|L_2\| &\leq \left\|\left((\widehat{\Gamma}^S)^{-1} - \Gamma^{-1}\right)\right\| \|F\| \left\|\left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)^{-1}\right\|\\ &\quad\times \left\|F^\intercal(\widehat{\Gamma}^S)^{-1}\right\|\\ &= O_p\left(m_d^2(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^3\right).\end{aligned}$$

The same bound holds for $\|L_3\|$. As for L_4, note that $\|\beta\| = O_p(d^{1/2})$, $\|H\| = O_p(1)$, $\left\|\Gamma^{-1}\right\| \leq (\lambda_{\min}(\Gamma))^{-1} \leq K$, and $\|\beta H - F\| \leq \sqrt{rd}\,\|\beta H - F\|_{\text{MAX}} = O_p(d^{1/2}(\Delta_n \log d)^{1/2} + m_d)$, and that

$$\begin{aligned}&\lambda_{\min}\left(TH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)\\ &\quad\geq \lambda_{\min}\left(H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)\\ &\quad\geq \lambda_{\min}(\Gamma^{-1})\lambda_{\min}(\beta^\intercal\beta)\lambda_{\min}(H^\intercal H)\\ &\quad> Km_d^{-1}d,\end{aligned}$$

hence we have

$$\begin{aligned}\|L_4\| &\leq \left\|\Gamma^{-1}\right\| \|(\beta H - F)\| \left\|\left(TH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)^{-1}\right\|\\ &\quad\times \|H^\intercal\beta^\intercal\| \left\|\Gamma^{-1}\right\|\\ &= O_p(m_d(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^2).\end{aligned}$$

The same bound holds for L_5. Finally, with respect to L_6, we have

$$\begin{aligned}&\left\|\left(TH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)^{-1} - \left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)^{-1}\right\|\\ &\quad\leq Kd^{-2}m_d^2\left\|\left(TH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)\right.\\ &\qquad\left. - \left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)\right\|.\end{aligned}$$

Moreover, since we have

$$\begin{aligned}\left\|TH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H - d\Lambda^{-1}\right\| &= \left\|\Lambda^{-1}F^\intercal(\beta H - F)\right\|\\ &= O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right)\end{aligned}$$

and

$$\begin{aligned}&\left\|H^\intercal\beta^\intercal\Gamma^{-1}\beta H - F^\intercal(\widehat{\Gamma}^S)^{-1}F\right\|\\ \leq\ &\left\|(H^\intercal\beta^\intercal - F^\intercal)\Gamma^{-1}\beta H\right\| + \left\|F^\intercal\Gamma^{-1}(\beta H - F)\right\|\\ &+ \left\|F^\intercal\left(\Gamma^{-1} - (\widehat{\Gamma}^S)^{-1}\right)F\right\|\\ =\ &O_p\left(dm_d(\Delta_n \log d)^{1/2} + d^{1/2}m_d^2\right),\end{aligned}$$

combining these inequalities yields

$$\|L_6\| = O_p\left(m_d^3(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^4\right).$$

On the other hand, using the Sherman–Morrison–Woodbury formula again,

$$\begin{aligned}&\left\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\right\| = \left\|\left(T^{-1}\beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal + \Gamma\right)^{-1} - (\beta\mathrm{E}\beta^\intercal + \Gamma)^{-1}\right\|\\ \leq\ &\left\|\Gamma^{-1}\right\|^2 \|\beta H\|^2 \left\|\left(\left(TH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)^{-1}\right.\right.\\ &\left.\left. - \left(H^\intercal\mathrm{E}^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)^{-1}\right)\right\|\\ \leq\ &Kd\left\|TH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right\|^{-1}\\ &\quad\times \left\|H^\intercal\mathrm{E}^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right\|^{-1}\left\|T(\mathcal{X}\mathcal{X}^\intercal)^{-1} - \mathrm{E}^{-1}\right\|\\ =\ &O_p\left(m_d(\Delta_n \log d)^{1/2}\right).\end{aligned}$$

By the triangle inequality, we obtain

$$\begin{aligned}\left\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\right\| &\leq \left\|(\widehat{\Sigma}^S)^{-1} - \widetilde{\Sigma}^{-1}\right\| + \left\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\right\|\\ &= O_p\left(m_d^3(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^4\right). \quad \square\end{aligned}$$

A.5. Proof of Theorem 5

Proof of Theorem 5. This follows from Lemma 2. □

References

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.

Aït-Sahalia, Y., Brandt, M., 2001. Variable selection for portfolio choice. J. Finance 56, 1297–1351.

Aït-Sahalia, Y., Fan, J., Xiu, D., 2010. High-Frequency covariance estimates with noisy and asynchronous data. J. Amer. Statist. Assoc. 105, 1504–1517.

Aït-Sahalia, Y., Jacod, J., 2014. High Frequency Financial Econometrics. Princeton University Press.

Aït-Sahalia, Y., Kalnina, I., Xiu, D., 2014. The Idiosyncratic Volatility Puzzle: A Reassessment at High Frequency, Tech. Rep. The University of Chicago.

Aït-Sahalia, Y., Xiu, D., 2015. Principal Component Analysis of High Frequency Data. Princeton University and the University of Chicago.

Alessi, L., Barigozzi, M., Capasso, M., 2010. Improved penalization for determining the number of factors in approximate factor models. Statist. Probab. Lett. 80, 1806–1813.

Amengual, D., Watson, M.W., 2007. Consistent estimation of the number of dynamic factors in a large N and T panel. J. Bus. Econom. Statist. 25, 91–96.

Anderson, T.W., 1958. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

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

Bai, J., Ng, S., 2013. Principal components estimation and identification of static factors. J. Econometrics 176 (1), 18–29.

Bai, Z.D., Yin, Y.Q., 1993. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Ann. Probab. 21 (3), 1275–1294.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2011. Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. J. Econometrics 162, 149–169.

Bibinger, M., Hautsch, N., Malec, P., Reiß, M., 2014. Estimating the quadratic covariation matrix from noisy observations: Local method of moments and efficiency. Ann. Statist. 42 (4), 1312–1346.

Bickel, P.J., Levina, E., 2008a. Covariance regularization by thresholding. Ann. Statist. 36 (6), 2577–2604.

Bickel, P.J., Levina, E., 2008b. Regularized estimation of large covariance matrices. Ann. Statist. 36, 199–227.

Brandt, M.W., Santa-Clara, P., Valkanov, R., 2009. Covariance regularization by parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. Rev. Financ. Stud. 22, 3411–3447.

Brodie, J., Daubechies, I., Mol, C.D., Giannone, D., Loris, I., 2009. Sparse and stable Markowitz portfolios. Proc. Natl. Acad. Sci. 106, 12267–12272.

Cai, T., Liu, W., 2011. Adaptive thresholding for sparse covariance matrix estimation. J. Amer. Statist. Assoc. 106, 672–684.

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica 51, 1281–1304.

Chen, N.-F., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. J. Bus. 50 (1).

Christensen, K., Kinnebrock, S., Podolskij, M., 2010. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. J. Econometrics 159, 116–133.

Connor, G., Korajczyk, R., 1988. Risk and return in an equilibrium APT: Application of a new test methodology. J. Financ. Econ. 21, 255–289.

Connor, G., Korajczyk, R., 1993. A test for the number of factors in an approximate factor model. J. Finance 48, 1263–1291.

Croux, C., Renault, E., Werker, B., 2004. Dynamic factor models. J. Econometrics 119, 223–230.

Davis, C., Kahan, W.M., 1970. The rotation of eigenvectors by a perturbation. III. SIAM J. Numer. Anal. 7, 1–46.

DeMiguel, V., Garlappi, L., Nogales, F.J., Uppal, R., 2009a. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. Manage. Sci. 55, 798–812.

DeMiguel, V., Garlappi, L., Nogales, F.J., Uppal, R., 2009b. Optimal versus naive diversifcation: How inefficient is the 1/*N* portfolio strategy? Rev. Financ. Stud. 55, 798–812.
Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J. Econometrics 164, 188–205.
El Karoui, N., 2010. High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation. Ann. Statist. 38, 3487–3566.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.
Fan, J., Fan, Y., Lv, J., 2008. High dimensional covariance matrix estimation using a factor model. J. Econometrics 147, 186–197.
Fan, J., Furger, A., Xiu, D., 2016. Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high frequency data. J. Bus. Econom. Statist. 34 (4), 489–503.
Fan, J., Liao, Y., Mincheva, M., 2011. High-dimensional covariance matrix estimation in approximate factor models. Ann. Statist. 39 (6), 3320–3356.
Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc. Ser. B Stat. Methodol. 75, 603–680.
Fan, J., Zhang, J., Yu, K., 2012. Vast portfolio selection with gross-exposure constraints. J. Amer. Statist. Assoc. 107, 592–606.
Forni, M., Giannone, D., Lippi, M., Reichlin, L., 2009. Opening the black box: Structural factor models with large cross sections. Econometric Theory 25, 1319–1347.
Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic-factor model: Identification and estimation. Rev. Econ. Stat. 82, 540–554.
Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2004. The generalized dynamic factor model: Consistency and rates. J. Econometrics 119 (2), 231–255.
Forni, M., Lippi, M., 2001. The generalized dynamic factor model: Representation theory. Econometric Theory 17, 1113–1141.
Fryzlewicz, P., 2013. High-dimensional volatility matrix estimation via wavelets and thresholding. Biometrika 100, 921–938.
Gandy, A., Veraart, L.A.M., 2013. The effect of estimation in high-dimensional portfolios. Math. Finance 23, 531–559.
Gouriéroux, C., Jasiak, J., 2001. Dynamic factor models. Econometric Rev. 20, 385–424.
Green, R.C., Hollifield, B., 1992. When will mean-variance efficient portfolios be well diversified? J. Finance 47, 1785–1809.
Hallin, M., Liška, R., 2007. Determining the number of factors in the general dynamic factor model. J. Amer. Statist. Assoc. 102 (478), 603–617.
Hayashi, T., Yoshida, N., 2005. On covariance estimation of non-synchronously observed diffusion processes. Bernoulli 11, 359–379.
Herskovic, B., Kelly, B., Lustig, H., Nieuwerburgh, S.V., 2016. The common factor in idiosyncratic volatility: Quantitative asset pricing implications. J. Financ. Econ. 119 (2), 249–283.
Horn, R.A., Johnson, C.R., 2013. Matrix Analysis, Second ed. Cambridge University Press.
Jacod, J., Protter, P., 2012. Discretization of Processes. Springer-Verlag.
Jacquier, E., Polson, N.G., 2010. Simulation-based-estimation in portfolio selection. In: Chen, M.-H., Müller, P., Sun, D., Ye, K., Dey, D. (Eds.), Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger. Springer, pp. 396–410.
Jagannathan, R., Ma, T., 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. J. Finance 58, 1651–1684.
Johnstone, I.M., Lu, A.Y., 2009. On consistency and sparsity for principal components analysis in high dimensions. J. Amer. Statist. Assoc. 104 (486), 682–693.
Kapetanios, G., 2010. A testing procedure for determining the number of factors in approximate factor models. J. Bus. Econom. Statist. 28, 397–409.
Lai, T., Xing, H., Chen, Z., 2011. Mean-variance portfolio optimization when means and covariances are unknown. Ann. Appl. Stat. 5, 798–823.
Ledoit, O., Wolf, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empir. Finance 10, 603–621.
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88, 365–411.
Ledoit, O., Wolf, M., 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. Ann. Statist. 40, 1024–1060.
Merton, R.C., 1973. An intertemporal capital Asset Pricing Model. Econometrica 41, 867–887.
Mykland, P.A., Zhang, L., 2006. ANOVA for diffusions and Itô processes. Ann. Statist. 34, 1931–1963.
Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Rev. Econ. Stat. 92, 1004–1016.
Paul, D., 2007. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statist. Sinica 17, 1617–1642.
Pelger, M., 2015a. Large-Dimensional Factor Modeling Based on High-Frequency Observations, Tech. Rep. Stanford University.
Pelger, M., 2015b. Understanding Systematic Risk: A High-Frequency Approach, Tech. Rep. Stanford University.
Pesaran, M.H., Zaffaroni, P., 2008. Optimal Asset Allocation with Factor Models for Large Portfolios, Tech. Rep. Cambridge University.
Reiß, M., Todorov, V., Tauchen, G.E., 2015. Nonparametric test for a constant beta between itô semi-martingales based on high-frequency data. Stochastic Process. Appl. 125 (8), 2955–2988.
Ross, S.A., 1976. The arbitrage theory of capital asset pricing. J. Econom. Theory 13, 341–360.
Rothman, A.J., Levina, E., Zhu, J., 2009. Generalized thresholding of large covariance matrices. J. Amer. Statist. Assoc. 104, 177–186.
Shephard, N., Xiu, D., 2012. Econometric Analysis of Multivariate Realized QML: Estimation of the Covariation of Equity Prices Under Asynchronous Trading, Tech. Rep. University of Oxford and University of Chicago.
Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97, 1167–1179.
Tao, M., Wang, Y., Chen, X., 2013a. Fast convergence rates in estimating large volatility matrices using high-frequency financial data. Econometric Theory 29 (4), 838–856.
Tao, M., Wang, Y., Yao, Q., Zou, J., 2011. Large volatility matrix inference via combining low-frequency and high-frequency approaches. J. Amer. Statist. Assoc. 106, 1025–1040.
Tao, M., Wang, Y., Zhou, H.H., 2013b. Optimal sparse volatility matrix estimation for high-dimensional Itô processes with measurement errors. Ann. Statist. 41, 1816–1864.
Todorov, V., Bollerslev, T., 2010. Jumps and betas: A new framework for disentangling and estimating systematic risks. J. Econometrics 157, 220–235.
Trapani, L., 2017. A randomised sequential procedure to determine the number of factors. J. Amer. Statist. Assoc., forthcoming.
Zhang, L., 2011. Estimating covariation: Epps effect and microstructure noise. J. Econometrics 160, 33–47.
Zhou, H.H., Cai, T., Ren, Z., 2014. Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation, Tech. Rep. Yale University, pp. 1–54.