Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

The Five Trolls Under the Bridge: Principal Component Analysis With Asynchronous and Noisy High Frequency Data

Dachuan Chen , Per A. Mykland & Lan Zhang

To cite this article: Dachuan Chen , Per A. Mykland & Lan Zhang (2020) The Five Trolls Under the Bridge: Principal Component Analysis With Asynchronous and Noisy High Frequency Data, Journal of the American Statistical Association, 115:532, 1960-1977, DOI: 10.1080/01621459.2019.1672555

To link to this article: https://doi.org/10.1080/01621459.2019.1672555

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 1960–1977: Theory and Methods
https://doi.org/10.1080/01621459.2019.1672555

The Five Trolls Under the Bridge: Principal Component Analysis With Asynchronous and Noisy High Frequency Data

Dachuan Chen[a], Per A. Mykland[b], and Lan Zhang[c]

[a]Department of Information and Decision Sciences, The University of Illinois at Chicago, Chicago, IL; [b]Department of Statistics, The University of Chicago, Chicago, IL; [c]Department of Finance, The University of Illinois at Chicago, Chicago, IL

ABSTRACT

We develop a principal component analysis (PCA) for high frequency data. As in Northern fairy tales, there are trolls waiting for the explorer. The first three trolls are market microstructure noise, asynchronous sampling times, and edge effects in estimators. To get around these, a robust estimator of the spot covariance matrix is developed based on the smoothed two-scale realized variance (S-TSRV). The fourth troll is how to pass from estimated time-varying covariance matrix to PCA. Under finite dimensionality, we develop this methodology through the estimation of realized spectral functions. Rates of convergence and central limit theory, as well as an estimator of standard error, are established. The fifth troll is high dimension on top of high frequency, where we also develop PCA. With the help of a new identity concerning the spot principal orthogonal complement, the high-dimensional rates of convergence have been studied after eliminating several strong assumptions in classical PCA. As an application, we show that our first principal component (PC) closely matches but potentially outperforms the S&P 100 market index. From a statistical standpoint, the close match between the first PC and the market index also corroborates this PCA procedure and the underlying S-TSRV matrix, in the sense of Karl Popper. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2018
Accepted August 2019

KEYWORDS
Asynchronous sampling times; Factor model; High dimensionality; High frequency; Market microstructure noise; Principal component analysis

1. Introduction

In his will, Warren Buffett recommends that his wife should invest her inheritance in an index fund (Buffett 2014, p. 20). Many investors share this preference.

We shall argue that they may be better off by investing in a statistically estimated principal component (PC) instead. The economic arguments for these two approaches are closely related (Section 1.2), and we corroborate this with our empirical analysis in Sections 7.1–7.3. The main barrier to PC investing has so far been the quality of the statistical estimates, both in terms of method, and in terms of data size. With the ever increasing frequency of trading and liquidity of markets, the data are now available. This article is about getting the statistical method right.

This is an article about statistics, about principal component analysis (PCA) for data that are large in two different ways. The dimension is large, and the frequency of the data is also very high. In our empirical example, the dimension is 70–100, and the amount of data in each dimension is up to several observations per second, for 11 years (2007–2017). In the asymptotic theory, the dimension may stay fixed or go to infinity, and the sampling frequency in all cases becomes infinite.

The high frequency permits the precise construction of time-varying eigenvalues and PCs. We use a nonparametric Itô process model (which also permits leverage effect, see Section 2.1 for a precise description). As a result, scientific problems can be investigated with much less statistical uncertainty. Also, if eigenvalues and PCs form part of a measurement or an algorithm, high frequency estimates permit rapid updating under unstable conditions. This methodology can be applied wherever high frequency data can be found, such as in neuroscience, geoscience, climate recordings, wind measurements, turbulence, finance, economics, and on the Internet. The approach extends to factor analysis (see Sections 1.2, 1.3, and 5).

This is also an article about finance, which is our empirical application. Our findings are interesting in their own right. The high precision and the rapid updating means that investment allocations are less likely to be stale. We shall see in Sections 7.1–7.3 that this is indeed the case.

The article can therefore be read for its finance, or it can be read for its statistics, with finance as an incidental choice of application.

The challenge posed by high-frequency PCA is that it requires a most careful construction to give meaningful answers. One cannot use common shortcuts, such as ignoring noise or asynchronicity, or throwing out data to make the dataset nicer, or replacing spot by integrated covariances. We find in Section 6 (Figure 2) and Appendix G in the supplementary materials that eigenvalues and PCs may come out very wrong by making such shortcuts.

A special feature of our dataset is that it provides a particularly stern test for any PCA procedure, call it the *index test*, as follows. Economic theory provides reasons to think that we know à priori what the first PC should look like: it should be

CONTACT Per A. Mykland mykland@pascal.uchicago.edu Department of Statistics, The University of Chicago, Chicago, IL 60637, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

very close to the corresponding value weighted stock index, see the discussion in Section 1.2.

Our article meets this challenge and provides a carefully constructed high-frequency PCA. We outline in Section 1.3 what is technically new in this article. As validation that our method is indeed highly accurate, we shall see in our application that it enables us to draw highly precise and also long-term conclusions about the relationship between PCs and currently known financial factors (Section 7). In particular, it passes the index test very well, to our knowledge better than any other known PCA procedure, see Figures 5 and 6 and our comments in Section 7.2. This match to the index also suggests that our procedure uses a particularly well behaved covariance estimator in the form of the smoothed two-scale realized variance (S-TSRV, Section 1.3). In the sense of Popper (1959), this match is the positive outcome of the test of a theoretical prediction. Since the test is passed, it corroborates the accuracy of our PCA and S-TSRV methods.

The accuracy of our PCA may provide a firmer footing on which to "export" the index concept to markets (such as commodities) where there is less theoretical basis for how to weigh index components. Indices currently do exist in these cases, of course, but with less foundation than is the case for equities. Indices have substantial social value.

We stand "on the shoulders of Giants," and we start by reviewing the background for this problem (Sections 1.1–1.3).

1.1. PCA and Factor Analysis (in Statistics and Econometrics)

PCA is a form of unsupervised learning (see, e.g., Hastie, Tibshirani, and Friedman 2009). PCA was pioneered by Pearson (1901) and Hotelling (1933), and further developed in a large statistical literature (see, e.g., Anderson 1958, 1963; Mardia, Kent, and Bibby 1979 for the classical theory).

PCA is frequently also appropriate for factor analysis: estimate the first few PCs, and these are then also estimators of the main factors. This important insight originated in econometrics (Chamberlain and Rothschild 1983; Connor and Korajczyk 1986; Stock and Watson 1998, 2002; see also the survey in Chapter 6 of Campbell, Lo, and MacKinlay (1997)), and is a much simpler approach than the usual treatment of factor analysis that can (at the time of writing) be found in most current books on multivariate statistics. It is notable that this approximation relies on dimension going to infinity with the number of observations.

The approach has since been generalized to time dependent systems, notably by Bai and Ng (2002), Fan, Liao, and Mincheva (2013), Aït-Sahalia and Xiu (2017), Kong (2017), Pelger (2019a), and other papers by the same authors. This is an important thread in this article, and we return to this below in Section 1.3.

For the present, we emphasize that this construction relies on an assumption that a finite number of common factors dominate the system (they are "pervasive," in contemporary parlance (Section 5)). This not only makes the PCA and the factor analysis a good proxy for each other. It also means that the PCA and the factor analysis avoid any nasty statistical inconsistencies. We note that the situation where inconsistencies do occur has meanwhile also been a fruitful topic of research, in the form of random matrix theory (including Johnstone 2001; Tao 2012).

1.2. PCA and Factor Analysis (in Finance and Economics)

It is widely agreed that financial markets can be described by a small number of factors. This goes back to the so-called capital asset pricing model (Markowitz 1952, 1959; Sharpe 1964; Lintner 1965; Black 1972), which predicts that a single factor drives asset prices. It was later realized that prices may be driven by multiple factors. Particularly well known (empirical) factors are those developed by Fama and French (1992, 2017) and Carhart (1997). Meanwhile, theoretical multifactor (and approximate multifactor) models were developed starting with Ross (1976) and Chamberlain and Rothschild (1983). There is a vast literature in this area. For literature reviews, see, for example, Campbell, Lo, and MacKinlay (1997) and Cochrane (2005).

The literature on factor models is a main motivation for investing in index funds. Especially for the one factor model, economic theory predicts that this factor becomes the value of the entire market (see, e.g., Cochrane 2005, chap. 9). It is arguably a collective form of unsupervised learning. The literature cited in Section 1.1, however, predicts that the same factor can be found by PCA. To quote (Chamberlain and Rothschild 1983, p. 1285): "Thus, principal component analysis […] is an appropriate technique for finding an approximate factor structure." For multifactor models, similar considerations apply. The question then arises: should one find the factors, as in Fama and French (1992) and their successors, or should one invest based on the several main principal components? We shall look more closely at this question in Sections 7.1–7.3.

The one factor case is the basis of the "index test" of a PCA procedure: the first PC should be close to the stock index. In the multifactor case, this would approximately remain the case in the commonly assumed scenario where the index is the main factor driving asset returns.

1.3. Time-Varying and High-Frequency PCA and Factor Analysis

We build on three pillars. In a seminal paper, Aït-Sahalia and Xiu (2019) developed high frequency PCA with the elegant use of spectral functions. In an equally pioneering article, Fan, Liao, and Mincheva (2013) developed the principal orthogonal complement thresholding estimator (POET) method to parlay time discrete PCA into a factor analysis along the lines of Section 1.1, but, critically, using sparsity to obtain the separation of the factor and residual part. A third pillar is the S-TSRV as developed in Mykland, Zhang, and Chen (2019).

Important other papers on high frequency PCA and factor analysis include, in particular, Aït-Sahalia and Xiu (2017), Kong (2017), and Pelger (2019a, 2019b), but we shall not build on these directly. A main advantage of the high frequency approach is that one avoids stationarity assumptions, which may be unrealistic in economic or financial data (Sections 3 and 4).

The main difficulty with the existing literature on high frequency PCA is that it does not permit the data to be noisy or

asynchronous (except Dai, Lu, and Xiu 2019). The effect of noise can be devastating (Zhang, Mykland, and Aït-Sahalia 2005) on variances and covariances, and we shall see that this is also the case for PCA. Noise leads to over-estimation of eigenvalues, and the PCs do not come out correctly (Section 6.3, in particular Figure 2, and Appendix G in the supplementary materials, both in this article). Asynchronous times can also cause severe problems, especially when one tries to sweep them under the carpet with pre-averaging (Mykland, Zhang, and Chen 2019).

In the current article, we solve this problem by constructing a PCA for noisy high frequency data under irregular trading (observation) times. This is done by estimating instantaneous eigenvalues and eigenvectors based on an instantaneous version of the S-TSRV. To set standard errors, an observed asymptotic variance estimator (Mykland and Zhang 2017) emerges naturally under the same conditions (Sections 3 and 4.)

We then proceed to design (in Section 5) a new estimation methodology for high-dimensional spot covariance and precision matrices through high frequency PCA, which can be regard as the realized version of POET from Fan, Liao, and Mincheva (2013). The new methodology allows for time-varying volatility and for time-varying factor loadings. We assume (i) conditional sparsity structure of the spot covariance matrix and (ii) the pervasiveness of the common factors. The estimation starts with the constrained least quadratic variation (CLQV) optimization subject to canonical conditions. It is shown that the CLQV optimization is an asymptotic version of the constrained least squares (CLS) optimization from Fan, Liao, and Mincheva (2013). The equivalence between CLQV and asymptotic CLS yields a useful identity about the spot principal orthogonal complement, which completely frees us from the higher order assumptions on common factor and idiosyncratic component in classical PCA (Section 5.2.1). The asymptotics of the new methodology only relies on very basic assumptions about the spot factor loadings and the spot idiosyncratic covariance matrix, in analogy with Assumptions 2(b) and 4(a) in Fan, Liao, and Mincheva (2013). Following the general approach of Bai and Ng (2002), a data-driven approach is proposed to consistently estimate the number of common factors. As the building block of new methodology, the spot principal orthogonal complement is obtained through the CLQV optimization for the spot covariance matrix, of which the convergence rate under elementwise max norm is shown to be $\left(\Delta T_n \log d\right)^{1/2} + d^{-1/2}$, where $\Delta T_n = \left[(K-J)\,\Delta\tau_n^+\right]^{1/2}$ and $\log d = o\left(\Delta T_n\right)$ as $n, d \to \infty$. Finally, the estimator is obtained by thresholding the spot principal orthogonal complement, of which the inversion matrix is a consistent estimator for the spot precision matrix under classical conditions.

In recent years, high frequency data have been connected to the high-dimensional factor model while eliminating the stationarity conditions in classical PCA. In particular, important extensions include allowing time-varying volatilities in the log price processes (Aït-Sahalia and Xiu 2017), or allowing jumps in log price processes (Pelger 2019a, 2019b), or allowing noisy and (mildly) asynchronous observations (i.e., Dai, Lu, and Xiu 2019). The existing literature on high frequency data analysis conduct PCA on either the integrated covariance matrix $\int_0^{\mathcal{T}} c_t dt$, or the averaged covariance matrix $\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}} c_s ds$, where $(c_t)_{0\le t\le\mathcal{T}}$ denotes the process of spot covariance matrix and the time horizon $\mathcal{T}$ is fixed. However, based on the Weyl's theorem, the difference $\left|\bar{\lambda}^{(j)} - \lambda_t^{(j)}\right|$ can be large, that is, of order $O_p(d)$ provided $\sup_t \max_{r,s}\left|c_t^{(r,s)}\right| < \infty$, for any $1 \le j \le d$ and $0 \le t \le \mathcal{T}$ when $\mathcal{T}$ is fixed, where d is the cross-sectional dimension, and $\bar{\lambda}^{(j)}$ and $\lambda_t^{(j)}$ are the jth eigenvalues of $\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}} c_s ds$ and c_t, respectively. Also, the cited papers either do not take account of microstructure, or they use pre-averaging without taking account of the potentially misleading effects of irregular times (see Mykland, Zhang, and Chen 2019, sec. 2). These are reasons why the instantaneous behavior of the latent structures cannot be easily detected by existing techniques.

1.4. Organization and Notation

This article is organized as follows. Section 2 sets up the model, and provides a more precise decomposition of the S-TSRV estimator. Section 3 provides the estimator for the spot covariance matrix. Section 4 proposes the estimators for the realized spectral functions and develops the asymptotic theory under finite dimensionality assumption. Section 5 shows the connection between high frequency PCA and high dimensional factor models, by estimating the high dimensional spot covariance and precision matrices using the realized POET. Section 6 and Appendix G in the supplementary materials report the Monte Carlo evidence. Section 7 focuses on empirical work. All mathematical proofs are collected in Appendices A–F in the supplementary materials.

We draw attention to the following notation, which is used throughout this article. For a matrix $\mathbf{A}$, we denote its (i,j)th element by $\mathbf{A}^{(i,j)}$, its ith row by $(\mathbf{A})_{i,\bullet}$, and its jth column by $(\mathbf{A})_{\bullet,j}$. We denote the largest and smallest eigenvalue of matrix $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively. We denote by $\|\mathbf{A}\|, \|\mathbf{A}\|_1, \|\mathbf{A}\|_{\mathrm{F}}, \|\mathbf{A}\|_{\max}$ the spectral norm, L_1-norm, Frobenius norm, and elementwise max norm of matrix $\mathbf{A}$, defined as $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}^\intercal\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_j \Sigma_i \left|\mathbf{A}^{(i,j)}\right|$, $\|\mathbf{A}\|_{\mathrm{F}} = \mathrm{tr}^{1/2}(\mathbf{A}^\intercal\mathbf{A})$, $\|\mathbf{A}\|_{\max} = \max_{i,j}\left|\mathbf{A}^{(i,j)}\right|$. If $\mathbf{A}$ is a vector, then $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_{\mathrm{F}}$ are equal to its Euclidean norm. For two sequences, we write $x_n \asymp y_n$ if $x_n = O_p(y_n)$ and $y_n = O_p(x_n)$.

2. Basic Setup

2.1. The Model

Assume that the process $(X_t)_{0\le t\le\mathcal{T}} = \left(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(d)}\right)_{0\le t\le\mathcal{T}}$ is a d-dimensional continuous semimartingale (Itô processes) in the sense that

$$dX_t = \mu_t dt + \sigma_t dW_t,$$

where W_t is Brownian motion; μ_t and σ_t are Itô processes which can be mutually dependent with W. This is comparable to Definition 1 in Mykland and Zhang (2006), as well as Conditions 1 and 2 in Mykland, Zhang, and Chen (2019).

We define the spot covariance process as follows:

$$c_t = \left(\sigma\sigma^\intercal\right)_t, \tag{1}$$

which belongs to the set of positive-semidefinite matrices for any $0 \le t \le \mathcal{T}$. If X_t is continuous, then its quadratic variation $[X,X]_t = \int_0^t c_s ds$.

For the financial application, $\{X_t\}$ is not observed and can be considered as latent efficient prices (in logarithmic form). We assume that the observed process (observed log stock prices) $Y = \left(Y^{(1)}, Y^{(2)}, \ldots, Y^{(d)}\right)$ is contaminated by the market microstructure noise ϵ as follows:

$$Y^{(r)}_{t^{(r)}_j} = X^{(r)}_{t^{(r)}_j} + \epsilon^{(r)}_{t^{(r)}_j}, \quad \text{for } r = 1, 2, \ldots, d.$$

For each process $\{ Y^{(r)}_t\}$, it is observed not continuously, but on the grid $\mathcal{G}^{(r)} = \left\{0 = t^{(r)}_0 < t^{(r)}_1 < \cdots < t^{(r)}_{n^{(r)}} = \mathcal{T}\right\}$. In this article, the assumptions about the sampling times $t^{(r)}_j$ and microstructure noise $\epsilon^{(r)}$ follow from Conditions 1–4 in Mykland, Zhang, and Chen (2019).

We also make the following assumption about the covariation between spot volatility processes as follows.

Assumption 1 (Assumption on covariation of spot volatility processes). Assume that for all pairs of (r_1, s_1) and (r_2, s_2), $\left\langle c^{(r_1,s_1)}, c^{(r_2,s_2)}\right\rangle_t$ are continuously differentiable and $\left\langle c^{(r_1,s_1)}, c^{(r_2,s_2)}\right\rangle'_t$ are Itô processes in the sense of Definition 1 in Mykland and Zhang (2006). Also assume that $\sup_{0\le t\le\mathcal{T}} \|c_t\|_{\max} < \infty$.

Recall that eigenvalues are analytic functions of the corresponding covariance matrix so long as they have multiplicity one (e.g., Tsing, Fan, and Verriest 1994, Proposition 4.1, p. 168). In this case, therefore, the eigenvalues are also Itô processes, and they satisfy the statements of Assumption 1.

2.2. The Smoothed-TSRV

To estimate the integrated covariance matrix $\langle X,X\rangle_t$, we construct the S-TSRV estimator $\widehat{\langle X,X\rangle}_t$ on a synchronous grid

$$\left\{0 = \tau_{n,0} < \tau_{n,1} < \cdots < \tau_{n,N} = \mathcal{T}\right\}. \tag{2}$$

Denote $M^{(r)}_{n,i} = \#\left\{j : \tau_{n,i-1} < t^{(r)}_j \le \tau_{n,i}\right\}$. We can set $\Delta\tau^+_n = \max_i \Delta\tau_{n,i}$ and $M^-_n = \min_{i,r} M^{(r)}_{n,i}$. For the structure of blocks, we assume Condition 3 in Mykland, Zhang, and Chen (2019).

We also make two more assumptions in this article for the simplicity of discussion.

Assumption 2 (Assumption on averaged noise). We suppose that there is stationarity enough to assure $\mathrm{cov}\left(\bar\epsilon^{(s_1)}_i, \bar\epsilon^{(s_2)}_i\right) = \left(M^-_n\right)^{-1} \varsigma^{(s_1,s_2)}$ and $\sup_i \mathrm{var}\left(\bar\epsilon^{(s_1)}_i \bar\epsilon^{(s_2)}_i\right) = O_p\left(\left(M^-_n\right)^{-2}\right)$.

Assumption 3 (Assumption on block structure). Assume that $\Delta\tau^+_n \asymp M^-_n/n$, in which case the number of blocks $N = N_n$ is of exact order $O\left(n/M^-_n\right)$.

For $0 \le t \le \mathcal{T}$ and a pair (J,K), set

$$K\widetilde{\left[\bar Y^{(r)}, \bar Y^{(s)}\right]}^{(K)}_t = \left(\frac{1}{2}\sum_{i=1}^{b-K} + \sum_{i=b-K+1}^{N^*(t)-b} + \frac{1}{2}\sum_{i=N^*(t)-b+1}^{N^*(t)-K}\right) \times \left(\bar Y^{(r)}_{i+K} - \bar Y^{(r)}_i\right)\left(\bar Y^{(s)}_{i+K} - \bar Y^{(s)}_i\right),$$

where

$$N^*(t) = \max\left\{1 \le i \le N : \tau_{n,i} \le t\right\} \text{ and } b = K + J, \tag{3}$$

and for $1 \le i \le N$ and $1 \le r \le d$, the pre-averaged price is defined as

$$\bar Y^{(r)}_i = \frac{1}{M^{(r)}_{n,i}} \sum_{\tau_{n,i-1} < t^{(r)}_j \le \tau_{n,i}} Y^{(r)}_{t^{(r)}_j}. \tag{4}$$

We define $J\widetilde{\left[\bar Y^{(r)}, \bar Y^{(s)}\right]}^{(J)}$ similarly by switching J and K.

The S-TSRV is defined as

$$\widehat{\left\langle X^{(r)}, X^{(s)}\right\rangle}_t = \frac{1}{(1 - b/N)(K-J)} \times \left\{K\widetilde{\left[\bar Y^{(r)}, \bar Y^{(s)}\right]}^{(K)}_t - J\widetilde{\left[\bar Y^{(r)}, \bar Y^{(s)}\right]}^{(J)}_t\right\}.$$

If we assume that $K - J = O_p\left(\left(N/M^-_n\right)^{2/3}\right)$, as well as the other conditions to support the central limit theorem (CLT) in Theorem 5 and formula (39) of Mykland, Zhang, and Chen (2019), we have the following expression:

$$\widehat{\left\langle X^{(r)}, X^{(s)}\right\rangle}_t = \int_0^t c^{(r,s)}_u du + O_p\left(a_n\right), \tag{5}$$

where $c^{(r,s)}_t$ is the (r,s)th element of c_t, that is, defined in (1), and where the sequence $\{a_n\}_{n\ge1}$ is defined as

$$a_n = \left[(K-J)\,\Delta\tau^+_n\right]^{\frac{1}{2}}. \tag{6}$$

Moreover, under Assumptions 2 and 3, and assuming $K - J = O_p\left(\left(N/M^-_n\right)^{2/3}\right)$, the estimation error has a sharper representation as follows:

$$\widehat{\left\langle X^{(r)}, X^{(s)}\right\rangle}_t - \int_0^t c^{(r,s)}_u du = M^{(r,s)}_t + \tilde e^{(r,s)}_t - e^{(r,s)}_0, \tag{7}$$

where the main martingale term can be expressed as

$$M^{(r,s)}_t = M^{X,(r,s)}_t + M^{\epsilon,(r,s)}_t + o_p\left(a_n\right), \tag{8}$$

and

$$M^{X,(r,s)}_t = \sum_{p=1}^{K-J-1}\left(\frac{K-J-p}{K-J}\right)\sum_{i=J+p+1}^{N^*(t)} \Delta X^{(r)}_{\tau_{i-p}} \Delta X^{(s)}_{\tau_i}[2],$$

$$M^{\epsilon,(r,s)}_t = \frac{1}{K-J}\sum_{i=K+1}^{N^*(t)}\left(\bar\epsilon^{(r)}_{i-J} - \bar\epsilon^{(r)}_{i-K}\right)\bar\epsilon^{(s)}_i[2],$$

while the edge effect terms $e_0^{(r,s)}$ and $\tilde{e}_t^{(r,s)}$ are of order $O_p\left(a_n^2\right)$, and can be further expressed as

$$e_0^{(r,s)} = \frac{1}{K-J}\sum_{i=J+1}^{K}\bar{\epsilon}_{i-J}^{(r)}\bar{\epsilon}_i^{(s)}[2] + \sum_{p=1}^{K-J-1}\sum_{i=1}^{K-J-p} \times\left(\frac{K-J-p-i}{K-J}\right)\Delta X_{\tau_{J+i}}^{(r)}\Delta X_{\tau_{J+i+p}}^{(s)}[2] + \sum_{i=1}^{K-J}\left(\frac{K-J-i}{K-J}\right)\Delta X_{\tau_{J+i}}^{(r)}\Delta X_{\tau_{J+i}}^{(s)} + o_p\left(a_n^2\right), \quad (9)$$

and

$$\tilde{e}_t^{(r,s)} = -\frac{1}{K-J}\sum_{i=J}^{K-1}\bar{\epsilon}_{N^*(t)-i-J}^{(r)}\bar{\epsilon}_{N^*(t)-i}^{(s)}[2] - \sum_{p=1}^{K-J-1}\sum_{i=0}^{K-J-p} \times\left(\frac{K-J-p-i}{K-J}\right)\Delta X_{\tau_{N^*(t)-i-p}}^{(r)}\Delta X_{\tau_{N^*(t)-i}}^{(s)}[2] - \sum_{i=0}^{K-J}\left(\frac{K-J-i}{K-J}\right)\Delta X_{\tau_{N^*(t)-i}}^{(r)}\Delta X_{\tau_{N^*(t)-i}}^{(s)} + o_p\left(a_n^2\right). \quad (10)$$

Proof. The proof of this expression is gathered in Appendix A in the supplementary materials. □

3. Estimator of Spot Covariance

Suppose that $\{\Delta T_n\}_{n\geq 1}$ is a sequence of positive numbers satisfying

$$a_n^{-2}\Delta T_n \to \infty \text{ and } \Delta T_n \to 0 \text{ as } n \to \infty. \quad (11)$$

We define the estimator of spot volatility $c_t^{(r,s)}$ as follows: for $1 \leq r, s \leq d$,

$$\hat{c}_{\Delta T_n,t}^{(r,s)} = \frac{1}{\Delta T_n}\left(\widehat{\left\langle X^{(r)}, X^{(s)}\right\rangle}_{t+\Delta T_n} - \widehat{\left\langle X^{(r)}, X^{(s)}\right\rangle}_t\right). \quad (12)$$

Before stating consistency results, we introduce new quantities as follows:

$$\bar{c}_{\Delta T_n,t}^{(r,s)} = \frac{1}{\Delta T_n}\int_t^{t+\Delta T_n} c_u^{(r,s)}du,$$
$$\bar{\beta}_{\Delta T_n,t}^{(r,s)} = \frac{1}{\Delta T_n}\sum_{i=N^*(t)+1}^{N^*(t+\Delta T_n)}\bar{B}_{t+\Delta T_n,i}^{(r,s)}, \text{ and}$$
$$\tilde{\beta}_{\Delta T_n,t}^{(r,s)} = \frac{1}{\Delta T_n}\sum_{i=N^*(t)+1}^{N^*(t+\Delta T_n)}\tilde{B}_i^{(r,s)}[2], \quad (13)$$

and

$$\varphi_{\Delta T_n,t}^{(r_1,r_2,s_1,s_2)} = \frac{(K-J)\mathcal{T}}{N}\int_t^{t+\Delta T_n} c_u^{(r_1,r_2)}c_u^{(s_1,s_2)}dG_n(u)[2][2] + 2\varsigma^{(r_1,r_2)}\varsigma^{(s_1,s_2)}\frac{N^*(t+\Delta T_n)-N^*(t)}{(K-J)^2\left(M_n^-\right)^2}[2][2], \quad (14)$$

where "[2]" denotes the summation by switching r and s, and "[2][2]" means the summation over four terms where r_1 can change place with s_1 and r_2 can change place with s_2, and

$$\bar{B}_{l,i}^{(r,s)} = \int_{\tau_{i-1}}^{\tau_i}(l-u)\,dc_u^{(r,s)} \text{ for } l \geq \tau_i,$$
$$\tilde{B}_i^{(r,s)} = \left(\sum_{p=1}^{K-J-1}\left(\frac{K-J-p}{K-J}\right)\Delta X_{\tau_{i-p}}^{(r)}\right)\Delta X_{\tau_i}^{(s)} + \frac{1}{(K-J)}\left(\bar{\epsilon}_{i-J}^{(r)} - \bar{\epsilon}_{i-K}^{(r)}\right)\bar{\epsilon}_i^{(s)},$$

and

$$g_i = \frac{N}{(K-J)\mathcal{T}}\sum_{p=1}^{K-J-1}\left(\frac{K-J-p}{K-J}\right)^2\Delta\tau_{i-p} \text{ and}$$
$$G_n(t) = \sum_{\tau_i\leq t} g_i\Delta\tau_i,$$

where $N^*(\cdot)$ is defined in (3).

Lemma 1 (Consistency and optimal convergence rate of the spot volatility estimator). Suppose that ΔT_n is a sequence of positive numbers satisfying (11). Under Assumptions 1–3, for arbitrary $\varepsilon > 0$,

(i) $\left\|\bar{c}_{\Delta T_n,t}^{(r,s)} - c_t^{(r,s)}\right\|_2 = O_p\left(\Delta T_n^{1/2}\right)$ uniformly with respect to t, and consequently,

$$\sup_t\left|\bar{c}_{\Delta T_n,t}^{(r,s)} - c_t^{(r,s)}\right| = O_p\left(\Delta T_n^{1/2-\varepsilon}\right) = o_p(1).$$

More precisely, $\bar{c}_{\Delta T_n,t}^{(r,s)} - c_t^{(r,s)} = \bar{\beta}_{\Delta T_n,t}^{(r,s)} + o_p\left(\Delta T_n^{1/2}\right)$.

(ii) $\left\|\hat{c}_{\Delta T_n,t}^{(r,s)} - \bar{c}_{\Delta T_n,t}^{(r,s)}\right\|_2 = O_p\left(\Delta T_n^{-1/2}a_n\right)$ uniformly with respect to t, and consequently,

$$\sup_t\left|\hat{c}_{\Delta T_n,t}^{(r,s)} - \bar{c}_{\Delta T_n,t}^{(r,s)}\right| = O_p\left(\Delta T_n^{-1}\left(\Delta T_n a_n^2\right)^{1/2-\varepsilon}\right) = o_p(1).$$

More precisely, $\hat{c}_{\Delta T_n,t}^{(r,s)} - \bar{c}_{\Delta T_n,t}^{(r,s)} = \tilde{\beta}_{\Delta T_n,t}^{(r,s)} + O_p\left(\Delta T_n^{-1}(a_n^4)^{1/2-\varepsilon}\right)$.

(iii) If we further assume that $\Delta T_n \asymp a_n$, then the spot volatility estimator reaches the optimal convergence rate $O_p\left(a_n^{1/2}\right)$, that is

$$\sup_t\left|\hat{c}_{\Delta T_n,t}^{(r,s)} - c_t^{(r,s)}\right| = O_p\left(a_n^{1/2-\varepsilon}\right),$$

and more precisely, we have: $\hat{c}_{\Delta T_n,t}^{(r,s)} - c_t^{(r,s)} = \bar{\beta}_{\Delta T_n,t}^{(r,s)} + \tilde{\beta}_{\Delta T_n,t}^{(r,s)} + o_p\left(a_n^{1/2}\right)$.

Proof. The proof of this lemma is collected in Appendix B in the supplementary materials. □

If we further define

$$\beta_{\Delta T_n,t}^{(r,s)} = \hat{c}_{\Delta T_n,t}^{(r,s)} - c_t^{(r,s)}, \quad (15)$$

then we state the second-order behavior of $\beta_{\Delta T_n,t}^{(r,s)}$ in the following lemma.

Lemma 2 (Second-order and higher-order behavior of spot volatility estimator). Suppose that ΔT_n is a sequence of positive numbers satisfying (11). Under Assumptions 1–3:

(i) If we further assume $\inf_n a_n^{-1}\Delta T_n > 0$, then $\beta_{\Delta T_n,t}^{(r_1,s_1)}\beta_{\Delta T_n,t}^{(r_2,s_2)} = O_p(\Delta T_n)$ and for $h \geq 3$, we have $\prod_{l=1}^{h}\beta_{\Delta T_n,t}^{(r_l,s_l)} = O_p\left(\Delta T_n^{h/2}\right)$ uniformly with respect to t.

(ii) If we further assume $a_n^{-1}\Delta T_n \to 0$ as $n \to \infty$, then $\beta_{\Delta T_n,t}^{(r_1,s_1)}\beta_{\Delta T_n,t}^{(r_2,s_2)} = O_p\left(a_n^2\Delta T_n^{-1}\right)$ and for $h \geq 3$, we have $\prod_{l=1}^{h}\beta_{\Delta T_n,t}^{(r_l,s_l)} = O_p\left(\left(a_n\Delta T_n^{-1/2}\right)^h\right)$ uniformly with respect to t.

(iii) If we further assume $a_n^{-1}\Delta T_n \to 0$ as $n \to \infty$, we have

$$\sup_t \left\| E\left(\beta_{\Delta T_n,t}^{(r_1,s_1)}\beta_{\Delta T_n,t}^{(r_2,s_2)}|\mathcal{F}_t\right) - \frac{1}{\Delta T_n^2}\varphi_{\Delta T_n,t}^{(r_1,r_2,s_1,s_2)}\right\|_2 = O_p\left(a_n^4\Delta T_n^{-2}\right) + o_p(a_n), \tag{16}$$

and

$$\sup_t \left\| \beta_{\Delta T_n,t}^{(r_1,s_1)}\beta_{\Delta T_n,t}^{(r_2,s_2)} - \frac{1}{\Delta T_n^2}\varphi_{\Delta T_n,t}^{(r_1,r_2,s_1,s_2)}\right\|_2 = O_p\left(a_n^2\Delta T_n^{-1}\right), \tag{17}$$

where $\varphi_{\Delta T_n,t}^{(r_1,r_2,s_1,s_2)}$ is defined in (14).

Proof. The proof of (i) and (ii) in this lemma is similar to the proof of Lemma 1. The proof of (iii) is collected in Appendix C in the supplementary materials. □

4. High Frequency PCA Under Finite Dimensionality

When the dimension d is finite, PC analysis using high frequency data may conveniently be based on the estimation of integrals $\int_0^{\mathcal{T}} F(c_s)\,ds$ of vector-valued *spectral functions* $F = (F_1,\ldots,F_d)$. Specifically, a spectral function F is defined on a subset of all positive semidefinite matrices, and it must satisfy that $F(X) = F(O^\intercal XO)$ for any positive semidefinite matrix X and any orthogonal and symmetric matrix O.

The concept of spectral function is well documented in Friedland (1981) and Aït-Sahalia and Xiu (2019, sec. 2.5, pp. 291–292), to whom we refer for a review of the concept. It is central to the latter's development of PCA.

A main property of spectral functions F is that they can be decomposed as $F = f \circ \lambda$, where f is a symmetric function on an open symmetric domain in $\mathbb{R}_d^+$, and $\lambda(X)$ is the vector of all nonincreasing eigenvalues of the positive semidefinite matrix X (Aït-Sahalia and Xiu 2019). Building on Aït-Sahalia and Xiu, we impose a continuity and growth condition on f, as well as a condition that eigenvalue processes cannot cross each other (Aït-Sahalia and Xiu 2019, Assumptions 2 and 3, p. 292). We make these assumptions by reference since they are best described in the context of Aït-Sahalia and Xiu (2019, sec. 2.5). Recall that we also assume the dimensionality d be asymptotically finite throughout this section.

To estimate the integrated spectral function, we first create a new equidistant grid as follows:

$$T_{n,i} = i\Delta T_n, \text{ for } 1 \leq i \leq B, \text{ such that } \Delta T_n \text{ satisfies (11) and } B = \mathcal{T}/\Delta T_n. \tag{18}$$

Condition (11) is an initial choice and we will elaborate on the selection of ΔT_n in next subsection.

We construct the estimator as follows:

$$\hat{V}(\Delta T_n, X; F) = \sum_{i=1}^{B} F\left(\hat{c}_{\Delta T_n,T_{n,i-1}}\right)\Delta T_n,$$

where $\hat{c}_{\Delta T_n,T_{n,i-1}}$ is defined in (12). Note that the estimator can also be written as

$$\hat{V}(\Delta T_n, X; F) = \sum_{i=1}^{B} f\left(\hat{\lambda}_{T_{i-1}}\right)\Delta T_n,$$

where $\hat{\lambda}_{T_{i-1}} = \lambda\left(\hat{c}_{\Delta T_n,T_{n,i-1}}\right)$ and $\lambda(X)$ is the vector of all nonincreasing eigenvalues of the positive semidefinite matrix X.

4.1. Selection of ΔT_n

In this subsection, we mainly discuss the selection of ΔT_n. We start from the decomposition of the estimation error:

$$\begin{aligned}\hat{V}(\Delta T_n, X; F) - \int_0^{\mathcal{T}} F(c_s)\,ds &= \underbrace{\sum_{i=1}^{B}\left[F\left(\hat{c}_{\Delta T_n,T_{n,i-1}}\right) - F\left(c_{T_{n,i-1}}\right)\right]\Delta T_n}_{\text{Error due to spot volatility estimator, } R^{\text{Spot}}} \\ &\quad - \underbrace{\sum_{i=1}^{B}\int_{T_{n,i-1}}^{T_{n,i}}\left[F(c_s) - F\left(c_{T_{n,i-1}}\right)\right]ds}_{\text{Discretization error, } R^{\text{Discrete}}}.\end{aligned} \tag{19}$$

By Taylor expansion, for $1 \leq m \leq d$, the mth component of the vector-valued function F can be expanded as follows:

$$\begin{aligned}&F_m\left(\hat{c}_{\Delta T_n,T_{n,i-1}}\right) - F_m\left(c_{T_{n,i-1}}\right) \\ &= \sum_{r_1,s_1=1}^{d}\partial_{r_1s_1}F_m\left(c_{T_{n,i-1}}\right)\beta_{\Delta T_n,T_{n,i-1}}^{(r_1,s_1)} \\ &\quad + \frac{1}{2}\sum_{r_1,s_1,r_2,s_2=1}^{d}\partial_{r_1s_1,r_2s_2}^2F_m\left(c_{T_{n,i-1}}\right)\beta_{\Delta T_n,T_{n,i-1}}^{(r_1,s_1)}\beta_{\Delta T_n,T_{n,i-1}}^{(r_2,s_2)} \\ &\quad + O_p\left(\left\|\boldsymbol{\beta}_{\Delta T_n,T_{n,i-1}}\right\|^3\right),\end{aligned}$$

where $\beta_{\Delta T_n,T_{n,i-1}}^{(r,s)}$ is defined in (13), and consequently, R^{Spot} could be further decomposed as follows:

$$\begin{aligned}&R^{\text{Spot}} \\ &= \underbrace{\Delta T_n\sum_{i=1}^{B}\left[\sum_{r_1,s_1=1}^{d}\partial_{r_1s_1}F_m\left(c_{T_{n,i-1}}\right)\beta_{\Delta T_n,T_{n,i-1}}^{(r_1,s_1)}\right]}_{\text{Main contributor of variance in } R^{\text{Spot}}, \text{ defined as } R^{\text{Spot-V}}} \\ &\quad + \underbrace{\Delta T_n\sum_{i=1}^{B}\left[\frac{1}{2}\sum_{r_1,s_1,r_2,s_2=1}^{d}\partial_{r_1s_1,r_2s_2}^2F_m\left(c_{T_{n,i-1}}\right)\beta_{\Delta T_n,T_{n,i-1}}^{(r_1,s_1)}\beta_{\Delta T_n,T_{n,i-1}}^{(r_2,s_2)}\right]}_{\text{Main contributor of bias in } R^{\text{Spot}}, \text{ defined as } R^{\text{Spot-B}}} \\ &\quad + \underbrace{O_p\left(\Delta T_n\sum_{i=1}^{B}\left\|\boldsymbol{\beta}_{\Delta T_n,T_{n,i-1}}\right\|^3\right)}_{\text{Aggregated remainder of Taylor expansion, defined as } R^{\text{Expansion}}}.\end{aligned} \tag{20}$$

Table 1. Error size comparison under different choices of ΔT_n.

	Types of error				
	R^{Discrete}	$R^{\text{Spot-V}}$	$R^{\text{Spot-B}}$	$E\left(R^{\text{Spot-B}}\right)-\varphi^{\text{Bias}}_{\Delta T_n}$	$R^{\text{Expansion}}$
$\Delta T_n \to 0$ and $\inf_n a_n^{-1}\Delta T_n > 0$	$O_p(\Delta T_n)$	$O_p(\Delta T_n)$	$O_p(\Delta T_n)$	$o_p(\Delta T_n)$	$O_p\left(\Delta T_n^2\right)$
$a_n^{-1}\Delta T_n \to 0$ and $a_n^{-3/2}\Delta T_n \to \infty$	$O_p(\Delta T_n)$	$O_p(a_n)$	$O_p\left(a_n^2\Delta T_n^{-1}\right)$	$O_p\left(a_n^4\Delta T_n^{-2}\right)=o_p(a_n)$	$O_p\left(a_n^3\Delta T_n^{-1}\right)$
$\sup_n a_n^{-3/2}\Delta T_n < \infty$ and $a_n^{-2}\Delta T_n \to \infty$	$O_p(\Delta T_n)$	$O_p(a_n)$	$O_p\left(a_n^2\Delta T_n^{-1}\right)$	$O_p\left(a_n^4\Delta T_n^{-2}\right)$	$O_p\left(a_n^3\Delta T_n^{-1}\right)$

NOTE: ΔT_n is defined in (18). The discretization error R^{Discrete} is defined in (19), the martingale term and bias term $R^{\text{Spot-V}}$ and $R^{\text{Spot-B}}$ and the aggregated remainder term $R^{\text{Expansion}}$ are defined in (20), and $E\left(R^{\text{Spot-B}}\right)-\varphi^{\text{Bias}}_{\Delta T_n}$ is the bias term contributed by the edge effect in the covariance estimator and $\varphi^{\text{Bias}}_{\Delta T_n}$ is defined in (21).

Because the second-order term in R^{Spot} will introduce a bias term into the estimation error, to achieve CLT and optimal convergence rate, we need to consider bias correction. The selection of ΔT_n should make sure not only the optimal convergence rate, but also the ease of estimation of the bias-correction term.

On the other hand, the edge effect (see (7) and (10)) in S-TSRV estimator can also contribute to the bias term in R^{Spot}, whose effect can be measured by $E\left(R^{\text{Spot-B}}\right)-\varphi^{\text{Bias}}_{\Delta T_n}$, where $\varphi^{\text{Bias}}_{\Delta T_n}$ is defined as

$$\varphi^{\text{Bias}}_{\Delta T_n} = \frac{1}{\Delta T_n}\sum_{i=1}^{B} \times\left[\frac{1}{2}\sum_{r_1,s_1,r_2,s_2=1}^{d}\partial^2_{r_1s_1,r_2s_2}F_m\left(c_{T_{n,i-1}}\right)\varphi^{(r_1,r_2,s_1,s_2)}_{\Delta T_n,T_{n,i-1}}\right], \tag{21}$$

with $\varphi^{(r_1,r_2,s_1,s_2)}_{\Delta T_n,T_{n,i-1}}$ being defined in (14).

By summarizing the results of Lemmas 1 and 2, we show the comparison of three cases in Table 1. From Table 1, we observe that to achieve the optimal convergence rate of $R^{\text{Spot-V}}$, that is, $O_p(a_n)$, we need to make sure $\sup_n a_n^{-1}\Delta T_n < \infty$. Moreover, when $\sup_n a_n^{-1}\Delta T_n < \infty$ and $a_n^{-2}\Delta T_n \to \infty$, the bias term $R^{\text{Spot-B}}$ has the order of $O_p\left(a_n^2\Delta T_n^{-1}\right)$, and at the same time, the bias caused by edge effect $E\left(R^{\text{Spot-B}}\right)-\varphi^{\text{Bias}}_{\Delta T_n}$ has the order of $O_p\left(a_n^4\Delta T_n^{-2}\right)$. To reduce the complexity in estimating the bias-correction term $E\left(R^{\text{Spot-B}}\right)$, we also require that $E\left(R^{\text{Spot-B}}\right)-\varphi^{\text{Bias}}_{\Delta T_n}$ have exactly smaller order than a_n, which implies that $\sup_n a_n^{-1}\Delta T_n < \infty$ and $a_n^{-3/2}\Delta T_n \to \infty$. However, when $\inf_n a_n^{-1}\Delta T_n > 0$ (a typical example is $\Delta T_n \asymp a_n$), the asymptotic variance term will include the terms related to $\left\langle c^{(r_1,s_1)}, c^{(r_2,s_2)}\right\rangle_t'$, which will bring much greater complexity to the bias-correction term and the AVAR estimator. Finally, we set the selection of ΔT_n as $a_n^{-1}\Delta T_n \to 0$ and $a_n^{-3/2}\Delta T_n \to \infty$ as $n \to \infty$.

Based on Table 1 and the above discussion, the rest of this article will be organized as follows. We will first state the consistency of $\hat{V}(\Delta T_n, X; F)$ with the assumption (11) and then show its second-order behavior under the assumption $a_n^{-1}\Delta T_n \to 0$ and $a_n^{-2}\Delta T_n \to \infty$ as $n \to \infty$. Finally, we propose the bias-corrected estimator, that is, $\tilde{V}(\Delta T_n, X; F)$ and show its consistency and CLT under the assumption $a_n^{-1}\Delta T_n \to 0$ and $a_n^{-3/2}\Delta T_n \to \infty$ as $n \to \infty$.

4.2. Consistency and Second-Order Behavior of $\hat{V}(\Delta T_n, X; F)$

The consistency is stated as following lemma.

Lemma 3 (Consistency of $\hat{V}(\Delta T_n, X; F)$). Suppose that ΔT_n is a sequence of positive real numbers satisfying (11). Assume the dimensionality d to be asymptotically finite. For the basic settings of processes, we assume Conditions 1–4 in Mykland, Zhang, and Chen (2019), and Assumptions 1–3. For the spectral function F, make Assumption 2 of (Aït-Sahalia and Xiu 2019, sec. 3.1, p. 292), see the beginning of (our) Section 4. Then we obtain

$$\hat{V}(\Delta T_n, X; F) \xrightarrow{p} \int_0^{\mathcal{T}} F(c_s)\, ds.$$

Proof. From the results (i) and (ii) in Lemma 1, we obtain

$$\sup_{1\le i\le B}\left|\hat{c}^{(r,s)}_{\Delta T_n,T_i} - c^{(r,s)}_{T_i}\right| \le \sup_{1\le i\le B}\left|\bar{c}^{(r,s)}_{\Delta T_n,T_i} - c^{(r,s)}_{T_i}\right| + \sup_{1\le i\le B}\left|\hat{c}^{(r,s)}_{\Delta T_n,T_i} - \bar{c}^{(r,s)}_{\Delta T_n,T_i}\right| = o_p(1),$$

which implies that $\hat{c}^{(r,s)}_{\Delta T_n,T_i} \xrightarrow{p} c^{(r,s)}_{T_i}$. Then based on this fact, we can show the consistency by following the proof of Theorem 1 in Aït-Sahalia and Xiu (2019). □

Next, we show the second-order behavior of $\hat{V}(\Delta T_n, X; F)$ in following theorem. We first define a quantity:

$$\left[M^{(r_1,s_1)}, M^{(r_2,s_2)}\right]^{(B)}_t = \sum_{T_{n,i}\le t}\left(M^{(r_1,s_1)}_{T_{n,i}} - M^{(r_1,s_1)}_{T_{n,i-1}}\right) \times\left(M^{(r_2,s_2)}_{T_{n,i}} - M^{(r_2,s_2)}_{T_{n,i-1}}\right). \tag{22}$$

Theorem 1 (Second-order behavior of $\hat{V}(\Delta T_n, X; F)$). Suppose that ΔT_n is a sequence of positive real numbers satisfying $a_n^{-1}\Delta T_n \to 0$ and $a_n^{-2}\Delta T_n \to \infty$ as $n \to \infty$. Assume the dimensionality d to be asymptotically finite. For the basic settings of processes, we assume Conditions 1–4 in Mykland, Zhang, and Chen (2019), as well as Assumptions 1–3 (of the current article). Moreover, assume the convergence rate of the S-TSRV estimator is $O_p(a_n)$, that is, see (5) and $a_n^{-2}\left[M^{(r_1,s_1)}, M^{(r_2,s_2)}\right]^{(B)}_u \xrightarrow{p} \text{ACOV}\left(M^{(r_1,s_1)}, M^{(r_2,s_2)}\right)_u$ for all u and $(r_1,s_1), (r_2,s_2)$. For the spectral function F, make Assumptions 2 and 3 of (Aït-Sahalia and Xiu 2019, sec. 3.1, p. 292), see

the beginning of (our) Section 4. Then we obtain

$$a_n^{-2}\Delta T_n\left(\hat{V}\left(\Delta T_n, X; F\right) - \int_0^{\mathcal{T}} F\left(c_s\right) ds\right) \xrightarrow{p} \varphi_{\mathcal{T}},$$

where

$$\varphi_{\mathcal{T}} = \frac{1}{2}\sum_{r_1,s_1,r_2,s_2=1}^{d}\int_0^{\mathcal{T}} \partial^2_{r_1s_1,r_2s_2}F\left(c_u\right) d\text{ACOV} \times \left(M^{(r_1,s_1)}, M^{(r_2,s_2)}\right)_u.$$

Proof. The proof of this theorem is gathered in Appendix D in the supplementary materials. □

Proposition 1. We further assume that the grid (2) is equidistantly spaced, that is, $\tau_i = i\Delta\tau_n$ with $\Delta\tau_n = \mathcal{T}/N$, and suppose that $N\left(K-J\right)^{-2}\left(M_n^-\right)^{-2} = a_n^2\xi$ with positive constant ξ.Then following the result (iii) in Lemma 2, and Theorem 1, we obtain

$$\varphi_t = \frac{1}{2}\sum_{r_1,s_1,r_2,s_2=1}^{d}\int_0^{t} \partial^2_{r_1s_1,r_2s_2}F\left(c_u\right) \times \left(\frac{1}{3}c_u^{(r_1,r_2)}c_u^{(s_1,s_2)}[2][2] + \frac{2\xi}{\mathcal{T}}\varsigma^{(r_1,r_2)}\varsigma^{(s_1,s_2)}[2][2]\right) du,$$

where "[2][2]" means the summation over four terms where r_1 can change place with s_1 and r_2 can change place with s_2.

4.3. Bias Corrected Estimator

In this subsection, we assume all conditions in Theorem 1. Moreover, further assume $a_n^{-1}\Delta T_n \to 0$ and $a_n^{-3/2}\Delta T_n \to \infty$ as $n \to \infty$. We further discuss implementation of the case of non-simple eigenvalues in Section 6.1.

We propose the bias corrected estimator as follows:

$$\tilde{V}\left(\Delta T_n, X; F\right) = \Delta T_n \sum_{i=1}^{B}\left[F\left(\hat{c}_{\Delta T_n, T_{n,i-1}}\right) - \frac{1}{2}\sum_{r_1,s_1,r_2,s_2=1}^{d} \times \partial^2_{r_1s_1,r_2s_2}F\left(\hat{c}_{\Delta T_n, T_{n,i-1}}\right)\hat{\varphi}^{(r_1,r_2,s_1,s_2)}_{\Delta T_n, T_{n,i-1}}\right], \quad (23)$$

where $\hat{c}^{(r,s)}_{\Delta T_n,t}$ is defined in (12) and

$$\hat{\varphi}^{(r_1,r_2,s_1,s_2)}_{\Delta T_n, T_{n,i-1}} = \check{\phi}^{(r_1,s_1)}_{\Delta T_n, T_{n,i-1}}\check{\phi}^{(r_2,s_2)}_{\Delta T_n, T_{n,i-1}}, \quad (24)$$

with

$$\check{\phi}^{(r,s)}_{\Delta T_n, T_{n,i-1}} = \frac{1}{2}\left(\hat{c}^{(r,s)}_{\Delta T_n/2,(i-1/2)\Delta T_n} - \hat{c}^{(r,s)}_{\Delta T_n/2,(i-1)\Delta T_n}\right). \quad (25)$$

We state the CLT of the bias corrected estimator as follows.

Theorem 2 (CLT of bias corrected estimator). Make all assumptions in Theorem 1, and further suppose $a_n^{-1}\Delta T_n \to 0$ and $a_n^{-3/2}\Delta T_n \to \infty$ as $n \to \infty$. Then we obtain

$$a_n^{-1}\left(\tilde{V}\left(\Delta T_n, X; F\right) - \int_0^{\mathcal{T}} F\left(c_s\right) ds\right) \xrightarrow{\mathcal{L}} W_{\mathcal{T}},$$

stably, where W_t is a continuous process defined on an extension of the original probability space, which conditionally on $\mathcal{F}$, is a continuous centered Gaussian martingale with its covariance matrix Σ given by

$$\Sigma_t^{(p,q)} = \sum_{r_1,s_1,r_2,s_2=1}^{d}\int_0^t \partial_{r_1s_1}F_p\left(c_u\right)\partial_{r_2s_2}F_q\left(c_u\right) d\text{ACOV} \times \left(M^{(r_1,s_1)}, M^{(r_2,s_2)}\right)_u.$$

Proof. The proof of this theorem is gathered in Appendix E in the supplementary materials. □

If we further make the assumptions in Proposition 1, we have

$$\Sigma_t^{(p,q)} = \sum_{r_1,s_1,r_2,s_2=1}^{d}\int_0^t \partial_{r_1s_1}F_p\left(c_u\right)\partial_{r_2s_2}F_q\left(c_u\right) \times \left(\frac{1}{3}c_u^{(r_1,r_2)}c_u^{(s_1,s_2)}[2][2] + \frac{2\xi}{\mathcal{T}}\varsigma^{(r_1,r_2)}\varsigma^{(s_1,s_2)}[2][2]\right) du.$$

Remark 1 (Estimator of AVAR). Following the idea of development of the bias-correction term, we propose the AVAR estimator as follows:

$$\widehat{\text{AVAR}}\left(\Delta T_n, X; F\right)^{(p,q)} = \Delta T_n^2\sum_{i=1}^{B}\Bigg[\sum_{r_1,s_1,r_2,s_2=1}^{d}\partial_{r_1s_1}F_p\left(\hat{c}_{\Delta T_n, T_{n,i-1}}\right) \partial_{r_2s_2}F_q\left(\hat{c}_{\Delta T_n, T_{n,i-1}}\right)\hat{\varphi}^{(r_1,r_2,s_1,s_2)}_{\Delta T_n, T_{n,i-1}}\Bigg],$$

where $\hat{\varphi}^{(r_1,r_2,s_1,s_2)}_{\Delta T_n, T_{n,i-1}}$ is defined in (24).

5. Estimation of High-Dimensional Spot Covariance PCA and Precision Matrices

The nonparametric framework of high frequency PCA allows the factor models to have time-varying factor loadings, and also frees the high-order assumptions concerning the common factor and idiosyncratic component. In this section, we first provide the detailed model specification and then propose the new estimation methodology for the high dimensional spot covariance and precision matrices, which can be regarded as the realized version of POET in Fan, Liao, and Mincheva (2013).

5.1. Factor Model With Time-Varying Factor Loadings

The log-price process $X_t = \left(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(d)}\right)$ of d stocks is generated from a factor model:

$$dX_t = \mathbf{B}_t d\mathbf{F}_t + dZ_t, \quad (26)$$

where $\mathbf{F}_t = \left(\mathbf{F}_t^{(1)}, \mathbf{F}_t^{(2)}, \ldots, \mathbf{F}_t^{(q)}\right)$ is a $q \times 1$ vector process, representing a set of unknown and time-varying common factors, $\mathbf{B}_t$ is a $d \times q$ matrix process of time-varying factor loadings, and $Z_t = \left(Z_t^{(1)}, Z_t^{(2)}, \ldots, Z_t^{(d)}\right)$ is a $d \times 1$ vector process of idiosyncratic noise components, satisfying

$$\langle \mathbf{F}, Z\rangle_t = 0 \text{ for all } t. \quad (27)$$

We should mention that the number of common factors $q \in \mathbb{N}^+$ is assumed to be fixed and asymptotically finite over time interval $[0, \mathcal{T}]$.

It is straightforward to see that if X, $\mathbf{F}$, $\mathbf{B}$, and Z are continuous Itô semimartingales, then

$$d\langle X, X\rangle_t = \mathbf{B}_t d\langle \mathbf{F}, \mathbf{F}\rangle_t \mathbf{B}_t^\intercal + d\langle Z, Z\rangle_t . \tag{28}$$

Recall the definition $c_t = \langle X, X\rangle_t'$. If we further define $c_t^{\mathbf{F}} = \langle \mathbf{F}, \mathbf{F}\rangle_t'$ and $\mathbf{s}_t = \langle Z, Z\rangle_t'$, it is obvious that for $0 \le t \le \mathcal{T}$, we have

$$c_t = \mathbf{B}_t c_t^{\mathbf{F}} \mathbf{B}_t^\intercal + \mathbf{s}_t. \tag{29}$$

To assure the asymptotic consistency between PCA and factor analysis, the existing PCA literature concerning high dimensional factor model opts to assume that $d \to \infty$ and that the eigenvalues corresponding to the common factors are spiked, that is, of order $O_p(d)$, while the eigenvalues corresponding to the idiosyncratic component are assumed to be bounded with respect to d, that is, see Bai and Ng (2002) and Fan, Liao, and Mincheva (2013). Note that if the eigenvalue corresponding to a common factor is diverging as $d \to \infty$, this factor is called *pervasive*. It is easy to see that if all common factors are pervasive, the decomposition (28) is asymptotically identifiable.

Because the common factors are unknown, it is necessary to normalize $\mathbf{B}_t$ and $\mathbf{F}_t$ using the following canonical condition:

Assumption 4 (Canonical condition). For all $0 \le t \le \mathcal{T}$, we assume that

$$d\langle \mathbf{F}, \mathbf{F}\rangle_t = \mathbb{I}_q dt \text{ and } \mathbf{B}_t^\intercal \mathbf{B}_t \text{ is diagonal.}$$

Under the canonical Assumption 4, it is natural to study the matrix $\mathbf{B}_t \mathbf{B}_t^\intercal$. Set this matrix to have eigenvalues $\{\mathfrak{l}_t^{(j)}\}_{1\le j\le q}$ (in non-ascending order) and corresponding eigenvectors $\{\mathfrak{g}_t^{(j)}\}_{1\le j\le q}$.

Then the asymptotic consistency between PCA and factor analysis can be rigorously stated in the form of the following proposition.

Proposition 2. Assume that for all $0 \le t \le \mathcal{T}$, all eigenvalues of the $q \times q$ matrix $d^{-1}\mathbf{B}_t^\intercal \mathbf{B}_t$ are distinct and bounded away from 0 and ∞ as $d \to \infty$. Then under Assumption 4, if $\left\{\lambda_t^{(j)}\right\}_{1\le j\le q}$ are the eigenvalues of c_t in a nonascending order and $\left\{\gamma_t^{(j)}\right\}_{1\le j\le q}$ are their corresponding eigenvectors, we have for $1 \le j \le q$: $\liminf_{d\to\infty} \left\|\tilde{\mathbf{b}}_t^{(j)}\right\|^2 / d > 0$ and

$$\left|\lambda_t^{(j)} - \mathfrak{l}_t^{(j)}\right| \le \|\mathbf{s}_t\|,$$
$$\left\|\gamma_t^{(j)} - \mathfrak{g}_t^{(j)}\right\| = O\left(d^{-1}\|\mathbf{s}_t\|\right)$$

and for $j > q$,

$$\left|\lambda_t^{(j)}\right| \le \|\mathbf{s}_t\|.$$

Proof. This proposition follows from the proofs of the Propositions 1 and 2 in Fan, Liao, and Mincheva (2013), which is a direct application of Weyl's theorem and sin (θ) theorem (Davis and Kahan 1970). □

Based on the result of Proposition 2, we know that the asymptotic consistency between PCA and factor analysis is assured by the pervasiveness assumption of common factors and boundedness assumption for the eigenvalues corresponding to the idiosyncratic components.

To assure the boundedness assumption of $\|\mathbf{s}_t\|$, the existing literature usually prespecifies one of several simple structures on $\mathbf{s}_t$, for example, the strict diagonal structure in Fan, Fan, and Lv (2008), the sparsity structure in Fan, Liao, and Mincheva (2011, 2013) and Fan, Liao, and Liu (2016), and the block diagonal structure in Fan, Furger, and Xiu (2016). For factor models with unknown factors, the sparsity structure can be handled by the POET as in Fan, Liao, and Mincheva (2013), while the block-diagonal structure can be treated by the block-diagonalization of principal orthogonal complement based on the Global Industrial Classification Standard (GICS) code. The latter approach was used in Aït-Sahalia and Xiu (2017).

In this article, we adopt the sparsity structure for $\mathbf{s}_t$, which is measured by

$$m_d = \sup_{0\le t\le\mathcal{T}} \max_{1\le i\le d} \sum_{1\le j\le d} \left|\mathbf{s}_t^{(i,j)}\right|^\nu \text{ for some } \nu \in (0,1),$$

and for $\nu = 0$, define $m_d = \sup_t \max_i \sum_j I\left(\mathbf{s}_t^{(i,j)} \ne 0\right)$. This measure is widely used in existing literature, that is, Bickel and Levina (2008) and Cai and Liu (2011). As pointed out by Fan, Liao, and Mincheva (2013), when the diagonal elements of $\mathbf{s}_t$ are bounded and $m_d = o(d)$, then the consistency in Proposition 2 can be achieved because $\|\mathbf{s}_t\| \le \|\mathbf{s}_t\|_1 = O(m_d)$.

5.2. Realized POET

The estimation of large covariance and related precision (inverse covariance) matrices is important in financial econometrics research. For example, the estimation performance of the covariance matrix for a factor model is naturally connected to the risk management problem in portfolio allocation (Fan, Zhang, and Yu 2012). Moreover, estimating the idiosyncratic covariance matrix and related precision (inverse covariance) matrix is the prerequisite for testing the asset pricing model (Sentana 2009; Fan, Liao, and Mincheva 2013).

Because of the time-varying feature of the volatility processes, it is here necessary to develop the estimation methodology for the spot covariance and precision matrices in high dimensionality. Since the new methodology is based on the thresholding of the spot principal orthogonal complement, which could be regarded as the realized version of POET in Fan, Liao, and Mincheva (2013), we call the new estimator *realized principal orthogonal complement thresholding estimator* (*realized POET*).

A new feature of realized POET is that the precision matrices of c_t and $\mathbf{s}_t$ can also be consistently estimated.

5.2.1. Constrained Least Quadratic Variation Method

Let $\lambda_t^{(1)} \ge \lambda_t^{(2)} \ge \cdots \ge \lambda_t^{(d)}$ be the eigenvalues of the spot covariance matrix c_t, and for $1 \le i \le d$, $\gamma_t^{(i)}$ is the eigenvector corresponding to $\lambda_t^{(i)}$. Then by spectral decomposition, it is

straightforward to see that c_t could be further decomposed as

$$c_t = \sum_{i=1}^{q} \lambda_t^{(i)} \gamma_t^{(i)} \left(\gamma_t^{(i)}\right)^\intercal + \mathbf{R}_t,$$

where $\mathbf{R}_t = \sum_{i=q+1}^{d} \lambda_t^{(i)} \gamma_t^{(i)} \left(\gamma_t^{(i)}\right)^\intercal$ is the *spot principal orthogonal complement*.

It is natural to see that under Assumption 4, we have for $0 \le t \le \mathcal{T}$:

$$\mathbf{B}_t \mathbf{B}_t^\intercal = \sum_{i=1}^{q} \lambda_t^{(i)} \gamma_t^{(i)} \left(\gamma_t^{(i)}\right)^\intercal \text{ and } \mathbf{s}_t = \mathbf{R}_t. \tag{30}$$

This approach to estimation is equivalent to a *CLQV optimization*:

$$(\mathbf{B}_t) = \underset{\mathbf{B}_t \in \mathbb{R}^{d\times q}}{\arg\min}\, \mathrm{tr}\, \langle Z, Z\rangle_t',$$

subject to the canonical condition (Assumption 4). The solution of the spot factor loading $\mathbf{B}_t$ in this CLQV optimization problem can be further expressed as

$$\mathbf{B}_t = \Gamma_t \Lambda_t^{1/2}, \tag{31}$$

where $\Lambda_t = \mathrm{diag}\left(\lambda_t^{(1)}, \lambda_t^{(2)}, \ldots, \lambda_t^{(q)}\right)$ and $\Gamma_t = \left(\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(q)}\right)$ for $0 \le t \le \mathcal{T}$. It is easy to check that the decompositions (29)–(31) are equivalent under Assumption 4.

Recall that $\mathrm{tr}\langle Z, Z\rangle_t = \sum_{i=1}^{d} \langle Z^{(i)}, Z^{(i)}\rangle_t$, which implies that this CLQV method is a partial analogy (not an exact equivalence) to the *CLS* method in Section 2.3 of Fan, Liao, and Mincheva (2013). The difference is that the CLQV method can recover neither the factors (i.e., $d\mathbf{F}_t$ term) nor the residuals (i.e., dZ_t term), while the CLS method can obtain both of them innately. The absence of residuals is a barrier to estimating the standard error of $\hat{\mathbf{s}}_t$, which is required in some entry-dependent thresholding approaches.

Although the residuals dZ_t cannot be recovered directly in the CLQV method, the optimization result $\mathbf{R}_t$ can be regard as the asymptotic least squares estimator of $\mathbf{s}_t$ given $\mathbf{B}_t = \Gamma_t \Lambda_t^{1/2}$. This can be briefly shown as follows. Suppose that dX_t and $\mathbf{B}_t$ are observed, based on Equation (26), the OLS solution of $d\mathbf{F}_t$ could be expressed as $\widehat{d\mathbf{F}}_t^{\mathrm{LS}} = \left(\mathbf{B}_t^\intercal \mathbf{B}_t\right)^{-1} \mathbf{B}_t^\intercal dX_t$ and consequently $\widehat{dZ}_t^{\mathrm{LS}} = \mathbf{P}_{\mathbf{B}_t} dX_t$ where $\mathbf{P}_{\mathbf{A}}$ is the projection matrix on $\mathbf{A}$ defined as

$$\mathbf{P}_{\mathbf{A}} := \mathbb{I}_d - \mathbf{A}\left(\mathbf{A}^\intercal \mathbf{A}\right)^{-1} \mathbf{A}^\intercal, \tag{32}$$

with $\mathbb{I}_d$ denoting the d-dimensional identity matrix. Then if we assume that $\mathrm{cov}(dX_t) = c_t dt$ and $\mathrm{cov}\left(\widehat{dZ}_t^{\mathrm{LS}}\right) = \mathbf{s}_t^{\mathrm{LS}} dt$, it is straightforward to see that the spot covariance of residual has the following expression:

$$\mathbf{s}_t^{\mathrm{LS}} = \mathbf{P}_{\mathbf{B}_t} c_t \mathbf{P}_{\mathbf{B}_t}^\intercal. \tag{33}$$

Finally, given $\mathbf{B}_t = \Gamma_t \Lambda_t^{1/2}$, it is straightforward to see that

$$\mathbf{s}_t^{\mathrm{LS}} = c_t - \mathbf{B}_t \mathbf{B}_t^\intercal, \tag{34}$$

which follows from the fact that $\mathbf{P}_{\mathbf{B}_t} \mathbf{B}_t = 0$ and $\mathbf{B}_t^\intercal \left(c_t - \mathbf{B}_t \mathbf{B}_t^\intercal\right) = \left(c_t - \mathbf{B}_t \mathbf{B}_t^\intercal\right) \mathbf{B}_t = 0$.

5.2.2. Estimators and Convergence Rates

First of all, we shall make some technical assumptions. In contrast to Bai and Ng (2002) (see Assumptions A and C(2,4,5)) and Fan, Liao, and Mincheva (2013) (see Assumptions 2(c), 4(b), and 4(c)), there is no need to make assumptions about the higher order behaviors of the common factor and the idiosyncratic component in our theory development. With the help of identities (33) and (34), we only impose some very basic assumptions on the spot factor loadings $\mathbf{B}_t^\intercal$ and the spot idiosyncratic covariance matrix $\mathbf{s}_t$, by following the Assumptions 2(b) and 4(a) in Fan, Liao, and Mincheva (2013).

Assumption 5. We denote the columns of $\mathbf{B}_t^\intercal$ as $\mathbf{b}_t^{(1)}$, $\mathbf{b}_t^{(2)}, \ldots, \mathbf{b}_t^{(d)}$. We assume that there exists $C_0 > 0$ such that for all $d \ge 1, 0 \le t \le \mathcal{T}$ and for all $i \le d$,

$$\left\|\mathbf{b}_t^{(i)}\right\|_{\max} < C_0.$$

There are constants $\vartheta_1, \vartheta_2 > 0$ such that $\lambda_{\min}(\mathbf{s}_t) > \vartheta_1$ and $\|\mathbf{s}_t\|_1 < \vartheta_2$ almost surely for all $0 \le t \le \mathcal{T}$.

We denote that spot covariance estimator for c_t by $\hat{c}_t$, that is, $\hat{c}_t = \left\{\hat{c}_{\Delta T_n, t}^{(r,s)}\right\}_{1\le r,s\le d}$ which is defined in (12). Moreover, we set $\Delta T_n \asymp a_n$ where a_n is defined in (6), which implies that the spot covariance matrix estimator $\hat{c}_t$ reaches the optimal convergence rate $O_p\left(a_n^{1/2}\right)$, based on the results of Lemma 1.

For some $k \le d$, we define

$$\hat{\mathbf{B}}_{k,t} = \hat{\Gamma}_{k,t} \hat{\Lambda}_{k,t}^{1/2}, \tag{35}$$

where $\hat{\Lambda}_{k,t} = \mathrm{diag}\left(\hat{\lambda}_t^{(1)}, \hat{\lambda}_t^{(2)}, \ldots, \hat{\lambda}_t^{(k)}\right)$, $\hat{\Gamma}_{k,t} = \left(\hat{\gamma}_t^{(1)}, \hat{\gamma}_t^{(2)}, \ldots, \hat{\gamma}_t^{(k)}\right)$ and $\hat{\lambda}_t^{(i)}$ is the ith largest eigenvalue of $\hat{c}_t$, and $\hat{\gamma}_t^{(i)}$ is the corresponding eigenvector.

The estimator of *the number of factors* q at time t is defined as

$$\hat{q}_t = \underset{1\le k\le q_{\max}}{\arg\min} \left\{ d^{-1}\, \mathrm{tr}\left(\hat{c}_t - \hat{\mathbf{B}}_{k,t} \hat{\mathbf{B}}_{k,t}^\intercal\right) + k\mathcal{G}(\Delta T_n, d)\right\}, \tag{36}$$

where $q_{\max}$ is a prespecified upper bound, and $\mathcal{G}(\Delta T_n, d)$ is a penalty function such that

$$\mathcal{G}(\Delta T_n, d) \to 0 \text{ and } \left(\left(\Delta T_n \log d\right)^{1/2} + d^{-1}\right)^{-1} \times \mathcal{G}(\Delta T_n, d) \to \infty \text{ as } n, d \to \infty. \tag{37}$$

In analogy with the similar idea of Theorem 2 in Bai and Ng (2002), we obtain the following result.

Theorem 3. Define $\hat{c}_t = \left\{\hat{c}_{\Delta T_n, t}^{(r,s)}\right\}_{1\le r,s\le d}$ with $\Delta T_n \asymp a_n$ and a_n is defined in (6). For basic settings about the observations, we assume Conditions 1–4 in Mykland, Zhang, and Chen (2019), and Assumptions 1–3 (in the current article). Suppose the assumptions in Proposition 2 and Assumption 5 hold. Assume that $\log d = o\left(\Delta T_n^{-1}\right)$ as $n \to \infty$ and $d \to \infty$. Let the estimator be defined as in (36) and the penalty function satisfying (37), then we have

$$P\left(\hat{q}_t = q\right) \to 1.$$

Proof. The proofs of Theorems 3–5 are in Appendix F in the supplementary materials. □

Based on the above theorem, we define the penalty function as follows:

$$\mathcal{G}(\Delta T_n, d) = \varkappa \left(\left(\Delta T_n \log d \right)^{1/2} + d^{-1} \right)^{1-\varepsilon_0}$$

for constants $\varkappa > 0$ and $0 < \varepsilon_0 < 1$. The estimator for spot factor loading $\mathbf{B}_t$ is defined as

$$\hat{\mathbf{B}}_{\hat{q}_t,t} = \hat{\Gamma}_{\hat{q}_t,t} \hat{\Lambda}_{\hat{q}_t,t}^{1/2}, \tag{38}$$

which is based on the definition (35). Then we could define the estimator of spot principal orthogonal complement as follows:

$$\hat{\mathbf{s}}_{\hat{q}_t,t} = \hat{c}_t - \hat{\mathbf{B}}_{\hat{q}_t,t} \hat{\mathbf{B}}_{\hat{q}_t,t}^{\intercal}, \tag{39}$$

which is equivalent to the expression $\hat{\mathbf{s}}_{\hat{q}_t,t} = \sum_{i=\hat{q}_t+1}^{d} \hat{\lambda}_t^{(i)} \hat{\gamma}_t^{(i)} \left(\hat{\gamma}_t^{(i)} \right)^{\intercal}$. Before introducing the main theorems, we first define the quantity:

$$\omega_n = \left(\Delta T_n \log d \right)^{1/2} + d^{-1/2}.$$

Theorem 4. Assume all the conditions in Theorem 3. Then we obtain

$$\left\| \hat{c}_t - c_t \right\|_{\max} = O_p \left(\left(\Delta T_n \log d \right)^{1/2} \right),$$

and

$$\left\| \hat{\mathbf{B}}_{\hat{q}_t,t} \hat{\mathbf{B}}_{\hat{q}_t,t}^{\intercal} - \mathbf{B}_t \mathbf{B}_t^{\intercal} \right\|_{\max} = O_p(\omega_n),$$
$$\left\| \hat{\mathbf{s}}_{\hat{q}_t,t} - \mathbf{s}_t \right\|_{\max} = O_p(\omega_n).$$

Proof. The proofs of Theorems 3–5 are in Appendix F in the supplementary materials. □

Now we apply the adaptive thresholding on $\hat{\mathbf{s}}_{\hat{q}_t,t}$. Denote the thresholding estimator by $\hat{\mathbf{s}}_{\hat{q}_t,t}^{*}$, defined as follows:

$$\hat{\mathbf{s}}_{\hat{q}_t,t}^{*} \triangleq \begin{cases} \hat{\mathbf{s}}_{\hat{q}_t,t}^{(i,j)}, & i = j, \\ \phi_{ij} \left(\hat{\mathbf{s}}_{\hat{q}_t,t}^{(i,j)} \right), & i \neq j, \end{cases}$$

where ϕ_{ij} is the adaptive thresholding rule, for $z \in \mathbb{R}$,

$$\phi_{ij}(z) = 0 \text{ when } |z| \leq \chi_{ij}, \text{ otherwise } \left| \phi_{ij}(z) - z \right| \leq \chi_{ij}.$$

Examples of the adaptive thresholding rule include the hard thresholding $\phi_{ij}(z) = zI\left(|z| \geq \chi_{ij}\right)$, soft thresholding, SCAD, and the adaptive lasso, see Rothman, Levina, and Zhu (2009) and Fan, Liao, and Liu (2016). Because of the absence of residuals, the standard error estimator of $\hat{\mathbf{s}}_{\hat{q}_t,t}^{(i,j)}$ cannot be easily obtained. Thus, in contrast to the choice of χ_{ij} in Fan, Liao, and Mincheva (2013), the thresholding parameter is set to be elementwise constant, that is, defined as

$$\chi_{ij} = C\omega_n, \tag{40}$$

with a sufficiently large $C > 0$.

Based on the result in Theorem 4, we obtain the following proposition.

Proposition 3. Assume all conditions in Theorem 3. Then for a sufficiently large $C > 0$ in thresholding parameter (40), the realized POET estimator satisfies

$$\left\| \hat{\mathbf{s}}_{\hat{q}_t,t}^{*} - \mathbf{s}_t \right\| = O_p \left(\omega_n^{1-\nu} m_d \right).$$

If $\omega_n^{1-\nu} m_d = o_p(1)$, then the eigenvalues of $\hat{\mathbf{s}}_{\hat{q}_t,t}^{*}$ are all bounded away from 0 with probability approaching 1, and

$$\left\| \left(\hat{\mathbf{s}}_{\hat{q}_t,t}^{*} \right)^{-1} - \mathbf{s}_t^{-1} \right\| = O_p \left(\omega_n^{1-\nu} m_d \right).$$

Proof. The proof of this proposition follows directly from the similar discussions in the proof of Theorem 5 of Fan, Liao, and Mincheva (2013). □

Next, define the spot covariance matrix estimator based on the realized POET as follows:

$$\hat{c}_{\hat{q}_t,t}^{*} := \hat{\mathbf{B}}_{\hat{q}_t,t} \hat{\mathbf{B}}_{\hat{q}_t,t}^{\intercal} + \hat{\mathbf{s}}_{\hat{q}_t,t}^{*}.$$

We then consider the estimation performance of the precision matrix based on $\left(\hat{c}_{\hat{q}_t,t}^{*} \right)^{-1}$. The theoretical development is based on the Sherman–Morrison–Woodbury formula, that is

$$\begin{aligned} \left(\hat{c}_{\hat{q}_t,t}^{*} \right)^{-1} = & \left(\hat{\mathbf{s}}_{\hat{q}_t,t}^{*} \right)^{-1} - \left(\hat{\mathbf{s}}_{\hat{q}_t,t}^{*} \right)^{-1} \\ & \times \hat{\mathbf{B}}_{\hat{q}_t,t} \left(\mathbb{I}_{\hat{q}_t} + \hat{\mathbf{B}}_{\hat{q}_t,t}^{\intercal} \left(\hat{\mathbf{s}}_{\hat{q}_t,t}^{*} \right)^{-1} \hat{\mathbf{B}}_{\hat{q}_t,t} \right)^{-1} \\ & \times \hat{\mathbf{B}}_{\hat{q}_t,t}^{\intercal} \left(\hat{\mathbf{s}}_{\hat{q}_t,t}^{*} \right)^{-1}. \end{aligned}$$

We show that the convergence rate for the estimator of the precision matrix is as follows.

Theorem 5. Assume all conditions in Theorem 3, as well as $\omega_n^{1-\nu} m_d = o_p(1)$, then for a sufficiently large $C > 0$ in thresholding parameter (40), $\left(\hat{c}_{\hat{q}_t,t}^{*} \right)^{-1}$ is nonsingular with probability approaching 1, and

$$\left\| \left(\hat{c}_{\hat{q}_t,t}^{*} \right)^{-1} - c_t^{-1} \right\| = O_p \left(\omega_n^{1-\nu} m_d \right).$$

Proof. The proofs of Theorems 3–5 are in Appendix F in the supplementary materials. □

6. Monte Carlo Evidence

In this section, we use Monte Carlo simulation to show the numerical validity of our methodology. We will take the estimation of eigenvalues as an example, where the eigenvalues are allowed to be nonsimple. Further simulation results are presented in Appendix G in the supplementary materials.

6.1. Bias Corrected Estimator for Nonsimple Eigenvalues

Suppose the eigenvalues of a d-dimensional positive semidefinite matrix X satisfy

$$\begin{aligned} \lambda^{(1)}(X) = \cdots = \lambda^{(g_1)}(X) > \lambda^{(g_1+1)}(X) = \cdots = \lambda^{(g_2)}(X) \\ > \cdots \lambda^{(g_{r-1})}(X) > \lambda^{(g_{r-1}+1)}(X) = \lambda^{(g_r)}(X) \geq 0, \end{aligned}$$

where $g_r = d$ and r is the number of distinct eigenvalues. We would like to estimate:

$$\int_0^{\mathcal{T}} F^{\lambda}(c_s)\, ds,$$

where

$$F^{\lambda}(\cdot) = \left(\frac{1}{g_1} \sum_{j=1}^{g_1} \lambda^{(j)}(\cdot), \frac{1}{g_2 - g_1} \sum_{j=g_1+1}^{g_2} \lambda^{(j)}(\cdot), \ldots, \frac{1}{g_r - g_{r-1}} \sum_{j=g_{r-1}+1}^{g_r} \lambda^{(j)}(\cdot) \right)^{\intercal}.$$

We can also write $F^{\lambda}(\cdot)$ using its components: $F_p^{\lambda}(\cdot)$ with $p = 1, 2, \ldots, r$. Without loss of generality, we set $g_0 = 0$.

Following from the similar calculations in Corollary 1 and related proof in Aït-Sahalia and Xiu (2019), for $1 \le p \le r$, we know that the consistent estimator is

$$\hat{V}\left(\Delta T_n, X; F_p^{\lambda}\right) = \Delta T_n \sum_{i=1}^{B} \left\{ \frac{1}{g_p - g_{p-1}} \sum_{h=g_{p-1}+1}^{g_p} \hat{\lambda}^{(h)}_{\Delta T_n, T_{i-1}} \right\},$$

and the bias-corrected estimator can be expressed as

$$\begin{aligned} \tilde{V}\left(\Delta T_n, X; F_p^{\lambda}\right) = \Delta T_n \sum_{i=1}^{B} \Bigg\{ \frac{1}{g_p - g_{p-1}} \sum_{h=g_{p-1}+1}^{g_p} \Bigg[& \hat{\lambda}^{(h)}_{\Delta T_n, T_{i-1}} \\ & - \left(\hat{O}_{\Delta T_n, T_{i-1}}\right)_{h,\bullet} \check{\phi}_{\Delta T_n, T_{i-1}} \\ & \times \left(\hat{\lambda}^{(h)}_{\Delta T_n, T_{i-1}} \mathbb{I}_d - \hat{c}_{\Delta T_n, T_{i-1}}\right)^{+} \check{\phi}_{\Delta T_n, T_{i-1}} \\ & \times \left(\hat{O}_{\Delta T_n, T_{i-1}}\right)^{\intercal}_{h,\bullet} \Bigg] \Bigg\}, \end{aligned} \tag{41}$$

where $\hat{\lambda}^{(h)}_{\Delta T_n, T_{i-1}} = \lambda^{(h)}\left(\hat{c}_{\Delta T_n, T_{i-1}}\right)$ (the hth largest eigenvalue of matrix $\hat{c}_{\Delta T_n, T_{i-1}}$), $\hat{O}_{\Delta T_n, T_{i-1}}$ is the orthogonal matrix such that

$$\hat{O}_{\Delta T_n, T_{i-1}} \hat{c}_{\Delta T_n, T_{i-1}} \hat{O}_{\Delta T_n, T_{i-1}} = \operatorname{diag}\left(\lambda\left(\hat{c}_{\Delta T_n, T_{i-1}}\right)\right),$$

$\check{\phi}_{\Delta T_n, T_{i-1}} = \left\{\check{\phi}^{(r,s)}_{\Delta T_n, T_{i-1}}\right\}_{1 \le r,s \le d}$ defined in (25), $\mathbb{I}_d$ is the d-dimensional identity matrix and the superscript "+" denotes the Moore–Penrose inverse of a real matrix.

Moreover, for $1 \le p \le r$, the estimator for the asymptotic variance of $\tilde{V}\left(\Delta T_n, X; F_p^{\lambda}\right)$ can be expressed as

$$\widehat{\mathrm{AVAR}}\left(\Delta T_n, X; F_p^{\lambda}\right) = \Delta T_n^2 \sum_{i=1}^{B} \hat{\Psi}^{(p)}_{\Delta T_n, T_{i-1}}, \tag{42}$$

where

$$\hat{\Psi}^{(p)}_{\Delta T_n, T_{i-1}} = \frac{1}{\left(g_p - g_{p-1}\right)^2} \sum_{v=g_{p-1}+1}^{g_p} \left(\vartheta^{(v)}\right)^2$$

with ϑ being the vector of diagonal elements in the matrix $\hat{O}_{\Delta T_n, T_{i-1}} \check{\phi}_{\Delta T_n, T_{i-1}} \hat{O}^{\intercal}_{\Delta T_n, T_{i-1}}$, that is, for $1 \le v \le d$,

$$\vartheta^{(v)} = \left(\hat{O}_{\Delta T_n, T_{i-1}} \check{\phi}_{\Delta T_n, T_{i-1}} \hat{O}^{\intercal}_{\Delta T_n, T_{i-1}}\right)^{(v,v)}.$$

On the other hand, we denote the nonoverlapping estimator which is proposed by Aït-Sahalia and Xiu (2019) (i.e., see (ii) in Corollary 1) by $\hat{\theta}\left(k_n, \Delta_n, F_p^{\lambda}\right)$, where we set $\Delta_n = \Delta\tau_n$ and k_n to be the closest divisors of $[\mathcal{T}/\Delta\tau_n]$ to $\frac{1}{2}\Delta\tau_n^{-1/2}\sqrt{\log(d)}$ here d is the dimension of X. Moreover, we can construct the AVAR estimator of $\hat{}\left(k_n, \Delta_n, F_p^{\lambda}\right)$ in two ways. The first way is based on formula (16) of Aït-Sahalia and Xiu (2019), by plugging in the estimators $\hat{\lambda}_{T_i}$. The second way is to construct the "observed AVAR" by formula (42). These are used in Figure G.1 and Tables G.1–G.3 in Appendix G in the supplementary materials.

6.2. Simulation Settings

Following the factor model defined in (26) and (27), we further define

$$d\mathbf{F}_t^{(j)} = \mu_j dt + \sigma_t^{(j)} d\mathcal{W}_t^{(j)} \text{ and } dZ_t^{(i)} = \nu_t d\mathcal{B}_t^{(i)},$$

where $i = 1, 2, \ldots, d$ and $j = 1, 2, \ldots, q$.

In this simulation, the first component of $\mathbf{F}$ is set as the market factor. Thus, its factor loadings $\mathbf{B}_{\bullet,1}$ are positive. Therefore, we simulate the factor loading in the following scheme:

$$d\mathbf{B}_t^{(i,j)} = \begin{cases} \tilde{\kappa}_1\left(\tilde{\theta}_{i1} - \mathbf{B}_t^{(i,j)}\right) dt + \tilde{\xi}_1 \sqrt{\mathbf{B}_t^{(i,j)}}\, d\tilde{\mathcal{B}}_t^{(i,j)} & \text{if } j = 1, \\ \tilde{\kappa}_j\left(\tilde{\theta}_{ij} - \mathbf{B}_t^{(i,j)}\right) dt + \tilde{\xi}_j d\tilde{\mathcal{B}}_t^{(i,j)} & \text{if } j \ge 2. \end{cases}$$

The correlation matrix of $d\mathcal{W}$ is defined as $\rho^{\mathbf{F}}$. The volatility processes of $\mathbf{F}$ and Z are simulated as follows:

$$\begin{aligned} & d\left(\sigma_t^{(j)}\right)^2 = \kappa_j\left(\theta_j - \left(\sigma_t^{(j)}\right)^2\right) dt + \eta_j \sigma_t^{(j)} d\tilde{\mathcal{W}}_t^{(j)} \text{ and} \\ & dv_t^2 = \kappa\left(\theta - v_t^2\right) dt + \eta v_t d\bar{\mathcal{B}}_t, \end{aligned}$$

where the correlation between $d\mathcal{W}^{(j)}$ and $d\tilde{\mathcal{W}}^{(j)}$ is ρ_j.

For comparison purposes, all parameters in the simulation are set to be the same as Table 1 in Aït-Sahalia and Xiu (2019), except that $\theta = 0.06$ and $\eta = 0.3$.

The processes are sampled at an equidistant grid with $\Delta t_n = 1$ sec. And the observed processes are contaminated by microstructure noise:

$$Y_{t_j} = X_{t_j} + \epsilon_{t_j},$$

where ε_{t_j} are iid d-dimensional random vectors, sampled from $\mathbf{N}_d(0, \Sigma^{\epsilon})$, with $\Sigma^{\epsilon} = \Phi\Phi^{\intercal}$ and $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_d)^{\intercal}$. Note that $\Phi_1, \Phi_2, \ldots, \Phi_d$ are iid random variables from $N\left(0, (0.0005)^2\right)$. It is worth to mention that we purposely set the size of noise to be very small.

The time horizon in the simulation experiment is set as: $\mathcal{T} =$ 1 week (assume 1 week consists of 5 trading days). We assume that a trading day consists of 6.5 hrs of trading.

6.3. Simulation Results

We apply the realized PCA procedure with both $\hat{\theta}\left(k_n, \Delta_n, F_p^{\lambda}\right)$ and $\tilde{V}\left(\Delta T_n, X; F_p^{\lambda}\right)$. We first examine the effect of market

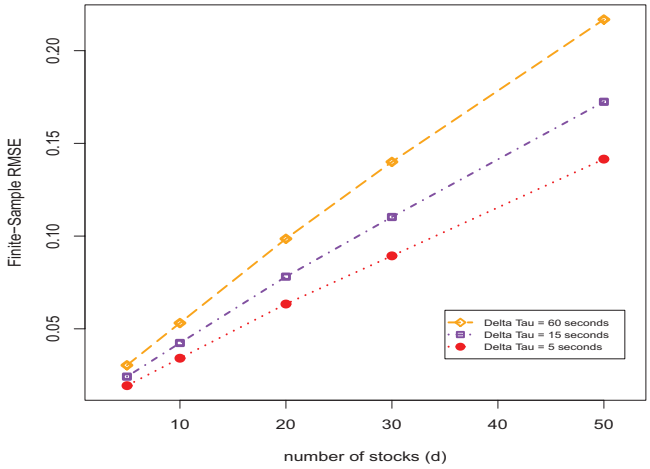


Figure 1. Finite sample root mean squared error (RMSE) of integrated largest eigenvalue estimates based on the S-TSRV, that is, $\tilde{V}\left(\Delta T_n, X; F_1^{\lambda}\right)$, with 1000 simulation trials and $\Delta\tau_n = 5, 15, 60$ sec, $d = 5, 10, 20, 30, 50$. Note that "Delta Tau" in the plot denoting $\Delta\tau_n$, which is the pre-averaging window of the S-TSRV.

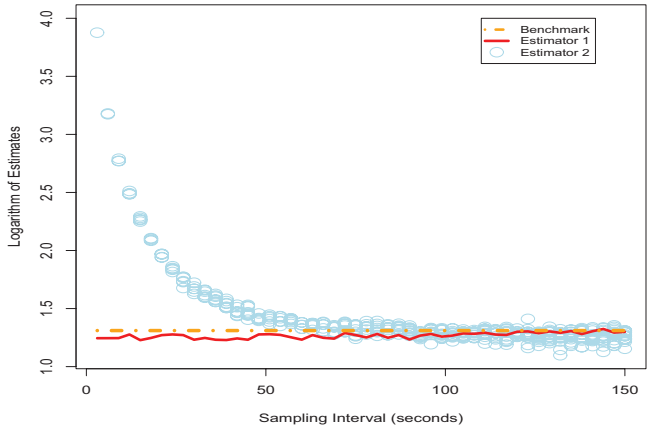


Figure 2. Signature plot for the estimates of integrated largest eigenvalue in logarithmic scale. "Estimator 1" (red solid curve) denotes the estimates $\tilde{V}\left(\Delta T_n, X; F_1^{\lambda}\right)$, and the sampling interval in the plot corresponds to the length of the pre-averaging window $\Delta\tau_n$. "Estimator 2" (lightblue dots) denotes the estimates $\hat{\theta}\left(k_n, \Delta_n, F_1^{\lambda}\right)$ computed with the different sampling intervals and different sampling starting points. The plot suggests that microstructure noise induces substantially more bias and variability on eigenvalue estimators than on regular volatility estimators. The *y*-axis is on the log scale.

microstructure noise in the estimation of integrated eigenvalues by estimator $\hat{\theta}\left(k_n, \Delta_n, F_p^{\lambda}\right)$. The examination is conducted under different combinations of stocks number and sampling frequency. The number of stocks $d = 5, 10, 20, 30$, and 50, while the sampling frequency is set in three scenarios:

1. $\Delta\tau_n = 5$ sec and $\Delta T_n = 2000\Delta\tau_n$, with $K = 20, J = 10$.
2. $\Delta\tau_n = 15$ sec and $\Delta T_n = 500\Delta\tau_n$, with $K = 10, J = 5$.
3. $\Delta\tau_n = 1$ min and $\Delta T_n = 160\Delta\tau_n$, with $K = 4, J = 2$.

Second, we show the estimation performance of $\tilde{V}\left(\Delta T_n, X; F_p^{\lambda}\right)$ with noisy observations, under the same settings of stock number and sampling frequency. Third, the performance of standard error estimators are also examined.

Overall, the simulation results show that, in the presence of microstructure noise, $\hat{\theta}\left(k_n, \Delta_n, F_p^{\lambda}\right)$ becomes inconsistent. More specifically, $\hat{\theta}\left(k_n, \Delta_n, F_p^{\lambda}\right)$ tends to over-estimate the eigenvalues. In particular, the higher the sampling frequency (smaller Δ_n), the larger the estimation bias; while the larger the number of stocks (higher d), the larger the estimation bias. Furthermore, the estimation bias seems to be greater for larger eigenvalues (smaller p). Detailed results are summarized in the tables in Appendix G in the supplementary materials.

In Figure 1, we show the finite sample RMSE of the first integrated eigenvalue estimates, that is, $\tilde{V}\left(\Delta T_n, X; F_p^{\lambda}\right)$ with $p = 1$. It is obvious that the RMSE value increases as the pre-averaging window $\Delta\tau_n$ increases. Moreover, it is evident that the increment of cross-sectional dimension d can magnify the absolute value of the differences in the RMSE values corresponding to different choices of $\Delta\tau_n$.

Figure 2 uses $d = 50$. For any fixed sampling interval Δ_n, one can (sub-)sample the data with varying starting point (e.g., starting from 9:01 a.m., or 9:02 a.m., etc.). Each lightblue circle in the graph represents an estimated $\hat{\theta}\left(k_n, \Delta_n, F_1^{\lambda}\right)$ based on a particular subsample. As seen in Figure 2, $\tilde{V}$ stays reasonably close to the true value even as the sampling interval shrinks to below 15 sec. On the other hand, $\hat{\theta}$ displays positive bias as sampling interval dips below 1 min. If one chooses to sample more sparsely (say, once every 3 min or longer), $\hat{\theta}$ based on a particular (sub-)sample displays greater estimation uncertainty. The distributional behavior of the bias-corrected estimate $\tilde{V}$ is validated, see the histograms in Appendix G in the supplementary materials. We emphasize that the invention of $\hat{\theta}\left(k_n, \Delta_n, F_1^{\lambda}\right)$ remains a seminal contribution to high dimensional analysis with high-frequency data. In applied work, the authors have selected sparse sampling intervals.

7. Empirical Study

7.1. Realized Eigenvalues and PCs

As an empirical study, we implement the high frequency PCA on the intraday returns of the S&P 100 Index (OEX) constituents. The stock prices are extracted from the Trade and Quote (TAQ) database of the New York Stock Exchange (NYSE). As illustrated by Figure 3 of Aït-Sahalia and Xiu (2019), it is easy to see that starting from 2007, more than 75% of trading intervals are less than 5 sec. We collect the intraday stock prices of 70 most actively traded stocks among the S&P 100 Index constituents, between 9:45 a.m. EST and 4:00 p.m. EST of each trading day, ranging from January 2007 to December 2017 (2769 trading days in total).

We estimate the integrated eigenvalues and -vectors in nine intervals of 2500 sec each, for every trading day, for a total of $2769 \times 9 = 24{,}921$ realizations over 11 years.[1] We show the percentages of the total variation explained by PCs corresponding to the first four eigenvalues in Figure 3. The graph shows that the first PC (PC1) explains about half (46.7 %) of the variation in the data. We shall assume all 70 eigenvalues are distinct. At least for the first eigenvalue, this is borne out by Figure 3.

[1] The estimators are as defined in Section 4. The tuning parameters are taken to be $\Delta_n = \Delta\tau_n = 5$ sec and $\Delta T_n = 500\Delta\tau_n, K = 100, J = 1$.

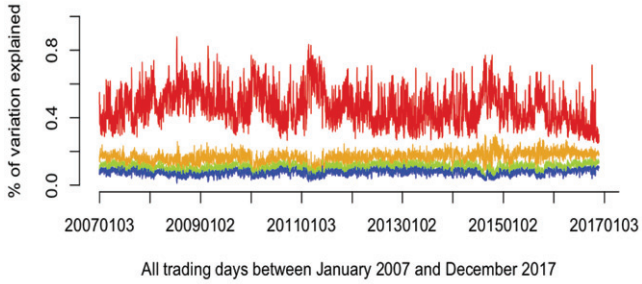


Figure 3. Percentage of the total variation explained by principal components, specifically the 1st to 4th eigenvalues of the S&P 100 index constituents during January 2007–December 2017. The values are rolling means over nine estimation periods of 2500 sec.

To compare investment strategies, we estimate the realized PCs corresponding to the first five eigenvalues using the S-TSRV. The hth realized PC is an estimate of $\int_0^t (\gamma_{s-}^{(h)})^\intercal dX_s$, where $\gamma_s^{(h)}$ is the d-dimensional ($d = 70$) hth eigenvector at time s, cf. Section 3.4 of Aït-Sahalia and Xiu (2019). With the following construction, the realized PCs become the log profit or loss (P/L) of an actual trading strategy.

To achieve this, the realized hth PC is estimated as follows:

$$\sum_{i=1}^{B} \log\left(1 + \left(\hat{\gamma}^{(h)}_{\Delta T_n, T_{i-1}}\right)^\intercal r_{T_i}\right), \tag{43}$$

where r_{T_i} is a column vector with jth element $r_{T_i}^{(j)} = (S_{T_i}^{(j)} - S_{T_{i-1}}^{(j)})/S_{T_{i-1}}^{(j)}$. These are the returns on the stocks $S^{(j)}$, $j = 1, \ldots, d$. The quantity (43) is therefore a log P/L on a strategy that invests a fraction

$$\delta_{i-1}^{(h)} = \sum_{j=1}^{d} \hat{\gamma}^{(h,j)}_{\Delta T_n, T_{i-1}} \tag{44}$$

of the accumulated wealth $w_{T_{i-1}}$ in stocks in the period from T_{i-1} to T_i, and keeps a fraction $1 - \delta_{i-1}^{(h)}$ in cash. Specifically, the strategy holds $w_{T_{i-1}} \hat{\gamma}^{(h,j)}_{\Delta T_n, T_{i-1}} / S^{(j)}_{T_{i-1}}$ units of stock $S^{(j)}$ in this time period. For simplicity, we take interest rates on cash to be zero; this was nearly the case for most of the time period under consideration.

We use the estimate $S_{T_i}^{(j)} = \exp(\bar{Y}_{N^*(T_i)})$, where $\bar{Y}_i$ and $N^*(t)$ are defined in (4) and (3), respectively, and $\hat{\gamma}^{(h)}_{\Delta T_n, T_{i-1}}$ is the eigenvector corresponding to the hth largest eigenvalue of $\hat{c}_{\Delta T_n, T_{i-1}} = \left\{\hat{c}^{(r,s)}_{\Delta T_n, T_{i-1}}\right\}_{1 \le r,s \le d}$ (as defined in formula (12), with the normalizations described in Sections 7.2 and 7.3). We have chosen to use $r_{T_i}^{(j)}$ instead of $\left(\bar{Y}^{(j)}_{N^*(T_i)} - \bar{Y}^{(j)}_{N^*(T_{i-1})}\right)$ (log returns) since the former give rise to a *feasible* trading strategy, whereas log prices cannot be traded. By Itô's formula, the two are approximations to each other.[2]

7.2. The Index and the First PC

The first PC is special, in that it is natural to compare it to the value weighted index, in our case the S&P 100, for the reasons discussed in Section 1. It is also special because the sum of the elements in first eigenvector (the weights given to the stocks) is away from zero, whereas the eigenvectors corresponding to the smaller eigenvalues have sums that follow a (somewhat skewed) bell shaped curve with mode around zero, see Figure 4. For the first PC to try to mimic the index, it seems natural to standardize the first eigenvector to have sum equal to one. The reason for this is that requiring $\delta_{i-1}^{(1)} = 1$ in (44) makes the investment strategy self-financing with no holdings in cash. This is analogous to the strategy of holding the index through futures or via an ETF which tracks the index.

With this choice of standardization, the first PC does indeed resemble the index, as documented in Figure 5. In fact, from the blue curve in Figure 5, it looks like the first PC may actually outperform the S&P 100 index. This is tantalizing, and one can speculate that the faster updating of the PC (relative to the index) is an advantage in a crisis. To construct portfolio weights for Figure 5, we use a rolling mean of the (70-dimensional) eigenvectors from the most recent nine periods of 2500 sec. Recall that there are nine such periods between 9:45 a.m. and 4 p.m. In other words, the portfolio weights are updated nine times every 24 hrs. (The overnight period has the same weight as the first period of the following trading day.) The purpose of rolling means is 2-fold: On the one hand, it reduces idiosyncratic statistical error in each estimated eigenvector. On the other hand, it reduces transaction cost by turning over only about 1/9th of the portfolio every 2500 sec. The choice of nine rather than, say, eight or ten, is based on the pragmatic advantages of following the daily cycle, and is also supported by the acf plot in the left panel of Figure 7. This figure also shows the idiosyncratic error at lag zero.

If transaction costs are larger than those used in Figure 5, it would be natural to update the portfolio less often, or to use a rolling mean over a longer period. As an experiment in this direction, we study the PC1 portfolio that is based on weekly (45 periods) rolling mean eigenvectors in Figure 6. With 20 basis points of transaction cost at each sale, the weekly PC1 portfolio again gets close to the S&P 100 index. An interesting finding is that for the PC1 portfolio without cost, the loss in going from daily to weekly rolling portfolio weights is small compared to the potential impact of transaction cost. Meanwhile, given the high-frequency data that goes into estimators, we have very high precision for the estimated weekly rolling portfolio weights, see, for example, the discussion of negative weights at the end of this section. For a given level of cost, there may be an optimal choice of this tuning parameter.

The idea that the first PC is close to the index has been around for some time (and forms the basis of our "index test"), but this degree of closeness has not been shown. Avellaneda and Lee (2010) conclude that the PC underperforms the index. The closest previous finding is that of Pelger (2019a, 2019b), who

[2]Note that under continuity the trading weights should be the same for our PCA and for a PCA conducted on the original scale. This is because the Itô correction does not alter the quadratic variation. Jumps would make a difference, and this remains to be explored, but for this article we take the view that it is more robust to carry out the PCA on the log scale, even if one wishes the tradable PC. Also, when estimating eigenvalues and -vectors which are actually used as forecasts for near future time periods, it may furthermore not be desirable to include a large jump that has already occurred in the near past.

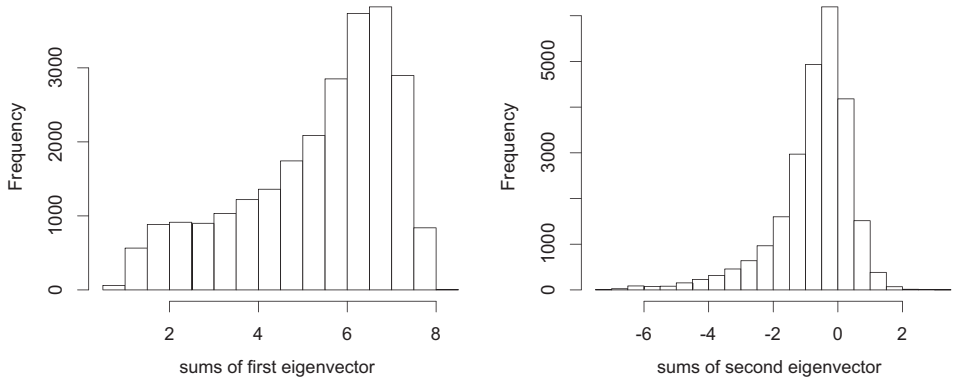


Figure 4. Distribution over time of the sums of the elements in the eigenvector. The next several eigenvectors have distributions that resemble the one for the second eigenvector. The element sums in the first eigenvector are always positive with the three smallest values being 0.74, 0.77, and 0.81 in (a), whereas the element sums in second (and later) eigenvectors may be positive or negative.

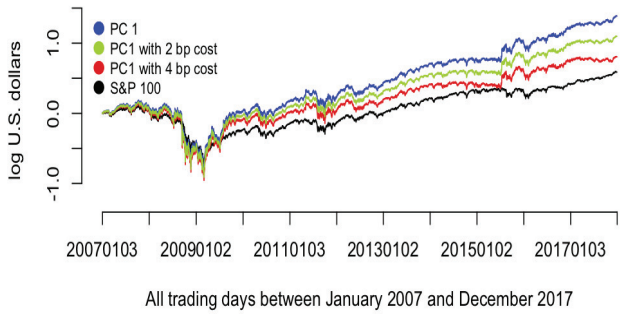


Figure 5. Plot of PC1 and log(OEF) as proxy for log(S&P100). Both are standardized to have value zero at the beginning of 2007. PC1 is constructed as described in the text in Sections 7.1 and 7.2. The green and red curves are also PC1, but incorporate a cost of 2 and 4 basis points (bp) of the value of each sales transaction. The graph corroborates the close relationship between, on the one hand, the STSRV covariance matrix and the resulting PCA, and, on the other hand, the economic arguments behind the value weighted index. This is a main empirical finding of this article.

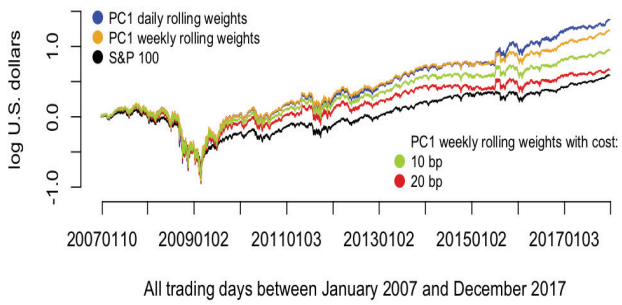


Figure 6. Plot of PC1 and log(OEF) as proxy for log(S&P100). Both are standardized to have value zero at the beginning of 2007. The weights from weekly rolling are equal weights from the proceeding 45 periods (9 periods per day $\times$ 5 days). The orange curve is PC1 without transaction cost. The green and red curves are also PC1, but incorporate a cost of 10 and 20 basis points (bp) of the value of each sales transaction.

concludes that the first PC and the index have "total correlation" equal to one. This is an important result, but total correlation is a measure of aggregated local behavior, and need not correspond to the very long term match demonstrated in Figure 5.

Table 2. Basic financial measures.

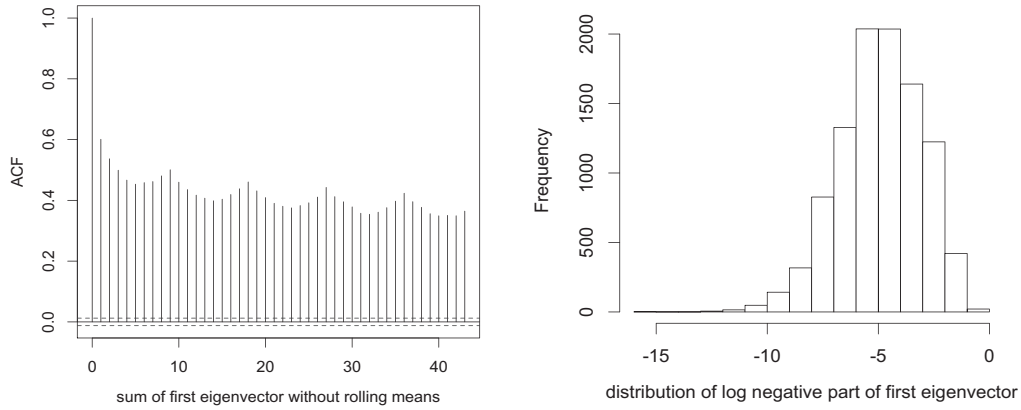
	S&P 100	PC1 daily rolling	PC1 weekly rolling
Annual returns	5.3%	12.5%	11.1%
Cumulative returns	58.8%	138.0%	122.2%
Annual volatility	15.6%	24.3%	23.2%
Sharpe ratio	34.0%	51.4%	47.8%
Sortino ratio	43.5%	72.0%	67.7%
Daily turnover	0	58.3%	11.2%
Maximum drawdown	56.2%	65.3%	65.5%
Alpha	0	0	0
Beta	1	1.44	1.40

NOTE: Annual returns are based on (43) with no transaction cost. Cumulative returns (without transaction cost) and maximum drawdowns are over the entire 11 years from 2007 to 2017. The risk-free rate is assumed to be zero. Volatilities were computed using the S-TSRV. For the computation of alpha and beta, S&P 100 is used as market proxy, and monthly returns have been used in the regression. For all the three series, the maximum drawdown occurred at market close on March 5, 2009.

We report standard financial measures of portfolio performance in Table 2. The weekly rolling PC1 seems to have reasonable performance in terms of risk adjusted return (Sharpe, Sortino ratios). By rolling weekly, we reduce the daily turnover to 11.2%. It is an open question how long we can make the rolling window without sacrificing financial gain.

We emphasize that there are a number of issues to be explored, and this is not a definitive study of the relationship between the index and the first PC. In the case where the asset returns have only one factor, the theoretical prediction would be that the PC should closely match the one factor (going back to Chamberlain and Rothschild (1983), and as discussed in Sections 1.1–1.3) and therefore (by CAPM) the index. In the multifactor environment, similar behavior may be related to the dominance of the index factor in stock prices, cf. Figure 3, but we leave further theory development for another paper. Meanwhile, the empirics is quite compelling. This is the "index test" discussed in Section 1.

Finally, we turn to some additional technical details involved in constructing the PCs. First of all, recall that the sign of the eigenvectors is arbitrary. If γ is an eigenvector, then so is $-\gamma$. For PC1, the natural solution is to require that $\delta_{i-1}^{(1)} = \sum_{j=1}^{d} \hat{\gamma}_{\Delta T_n, T_{i-1}}^{(1,j)}$ be positive. We impose this on all nine eigenvectors from each day. To obtain a self-financing trading strategy,



(a) (b)

Figure 7. Diagnostics for PC1. (a) Autocorrelation plot of the sums of the first eigenvector *without using the rolling mean*. It is clear that there is substantial idiosyncratic variance. There is also a period of 9, corresponding to the daily cycle. The same phenomenon applies to the first and higher order eigenvalues. The phenomenon disappears by using the rolling mean. (b) Distribution over time of $\log(n_i)$, as defined in (45), for the case of the rolling first eigenvector.

however, the requirement that $\delta_{i-1}^{(1)} = 1$ is imposed on the relevant rolling means of 9 or 45 periods.

There is a potential worry that the PC method produces substantial negative (short) positions in some stock. This is potentially a major difference with the value weighted index. For PC1, however, these negative positions are quite minor. If we define the negative fraction of the first eigenvector as

$$n_i = \sum_{j=1}^{d} \left(\hat{\gamma}^{(1,j)}_{\Delta T_n, T_{i-1}}\right)^- \Big/ \sum_{j=1}^{d} \hat{\gamma}^{(1,j)}_{\Delta T_n, T_{i-1}} \text{ where } x^- = \max(-x, 0), \tag{45}$$

where $\hat{\gamma}^{(1)}_{\Delta T_n, T_{i-1}}$ is the 9-period (daily) rolling first eigenvector. We find that the mean of n_i is 0.011, the 95th percentile is 0.067. The histogram of $\log(n_i)$ is given in Figure 7. The maximum over the 11 years is 0.538. For comparison, without the rolling mean, the maximum is 1480.94. Also for comparison, for the similar 45 period (weekly) rolling mean eigenvectors, the mean negative part is 0.0012, the 95th percentile is 0.0053, and the maximum over 11 years is 0.0778. Note that it is always a possibility to build limits on the negative part into the portfolio selection.

7.3. Other PCs

For the higher order eigenvectors, it is not natural to standardize in the same way as for PC1. The sums $\delta_i^{(h)}$ (from Equation (44)) of the eigenvectors straddles zero, as evidenced for $\delta_i^{(2)}$ in the right panel of Figure 4, meaning that the corresponding trading strategies in (43) are naturally market neutral. This is desirable since PC1 is meant to mimic the market index. The time series of higher order PCs are shown in Figure 8.

There remains the problem of choosing a sign for the higher order eigenvectors, since from the PCA this sign is arbitrary. We have here chosen to require that the sign of $\hat{\gamma}^{(h)}_{\Delta T_n, T_{i-1}}$ (the hth eigenvector for time period T_i) be chosen so that this eigenvector is as close as possible to eigenvector at T_{i-1}. This is the so-called "continuity method" which guarantees that the

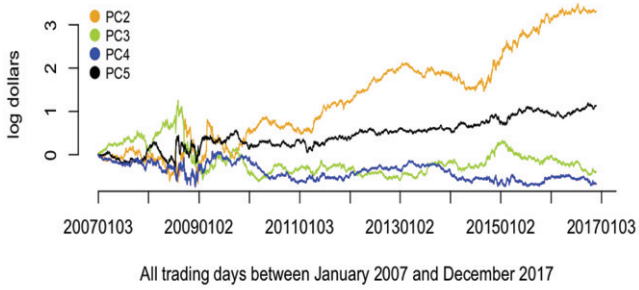


Figure 8. Time series of PC2–PC5.

hth eigenvector rotates no more than $\pi/2$ (clockwise or counterclockwise) from one period to the next. Specifically, proceed as follows.

Algorithm 1. Choice of sign of eigenvectors for $h \geq 2$

$$\text{assign sign}(\hat{\gamma}^{(h)}_{\Delta T_n, T_i}) \text{ so that sign}\{(\hat{\gamma}^{(h)}_{\Delta T_n, T_i})^\intercal \hat{\gamma}^{(h)}_{\Delta T_n, T_{i-1}}\} \geq 0. \tag{46}$$

The sign requirement follows from the geometric interpretation of the dot product. In this case, we require the cosine of the angle between $\hat{\gamma}^{(h)}_{\Delta T_n, T_{i-1}}$ and $\hat{\gamma}^{(h)}_{\Delta T_n, T_i}$ to be nonnegative. The $\delta_i^{(2)}$ in Figure 4(b) is based on Algorithm 1. If we had instead chosen the (arbitrarily signed) raw output from statistical package R, Figure 4(b) would have been more spread out.

As we have seen in Section 7.2, there are two sets of choices that have to be made about the eigenvectors. Algorithm 1 provides a systematic approach to choosing sign. It remains to choose the size of the eigenvectors. Once again, our approach for the first eigenvector (set $\delta_i^{(1)} = 1$ seems inappropriate for $h > 1$, as the natural choice of a market neutral trading strategy may be to start with approximately zero dollars, and then approximately balance short and long positions. This would be consistent with Figure 4(b). Specifically, for $\delta_i^{(2)}$, the mean over time is 0.72 while the standard deviation is 1.93.

We have here chosen the approach in the literature of requiring that $||\hat{\gamma}^{(h)}_{\Delta T_n, T_{i-1}}||_2 = 1$ for $h \geq 2$, cf. Aït-Sahalia and Xiu (2017, 2019) and Dai, Lu, and Xiu (2019). The latter papers also use this approach for $h = 1$.

An alternative would be to standardize the eigenvectors so that the corresponding PCs would have constant volatility. This is an appealing principle. This is not the case, however, for either the S&P 100, or for the PC1 that we have constructed above. For the moment, we conclude that the choice of this normalization is an open problem, and we hope to pursue this in a later article.

Supplementary Materials

The supplement contains the proofs of the theorems and other mathematical results in the main body of the paper (Appendix A-F), as well as additional simulation results (Appendix G).

Acknowledgments

We would like to thank the editors, associate editor, and the referees, for comments that substantially improved the article.

Funding

Financial support from the National Science Foundation under grants DMS 14-07812 and DMS 17-13129 (Mykland), and DMS 14-07820 and DMS 17-13118 (Zhang) is gratefully acknowledged.

References

Aït-Sahalia, Y., and Xiu, D. (2017), "Using Principal Component Analysis to Estimate a High Dimensional Factor Model With High-Frequency Data," *Journal of Econometrics*, 201, 384–399. [1961,1962,1968,1976]

——— (2019), "Principal Component Analysis of High Frequency Data," *Journal of the American Statistical Association*, 114, 287–303. [1961,1965,1966,1971,1972,1973,1976]

Anderson, T. W. (1958), *An Introduction to Multivariate Statistical Analysis* (Vol. 2), New York: Wiley. [1961]

——— (1963), "Asymptotic Theory for Principal Component Analysis," *The Annals of Mathematical Statistics*, 34, 122–148. [1961]

Avellaneda, M., and Lee, J.-H. (2010), "Statistical Arbitrage in the US Equities Market," *Quantitative Finance*, 10, 761–782. [1973]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [1961,1962,1968,1969]

Bickel, P. J., and Levina, E. (2008), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [1968]

Black, F. (1972), "Capital Market Equilibrium With Restricted Borrowing," *The Journal of Business*, 45, 444–455. [1961]

Buffett, W. E. (2014), *Letter to the Shareholders of Berkshire Hathaway Inc.*, New Bedford, MA: Berkshire Hathaway Inc. [1960]

Cai, T., and Liu, W. (2011), "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 106, 672–684. [1968]

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton, NJ: Princeton University Press. [1961]

Carhart, M. M. (1997), "On Persistence in Mutual Fund Performance," *Journal of Finance*, 52, 57–82. [1961]

Chamberlain, G., and Rothschild, M. (1983), "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304. [1961,1974]

Cochrane, J. H. (2005), *Asset Pricing* (2nd ed.), Princeton, NJ: Princeton University Press. [1961]

Connor, G., and Korajczyk, R. A. (1986), "Performance Measurement With the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, 15, 373–394. [1961]

Dai, C., Lu, K., and Xiu, D. (2019), "Knowing Factors or Factor Loadings, or Neither? Evaluating Estimators of Large Covariance Matrices With Noisy and Asynchronous Data," *Journal of Econometrics*, 208, 43–79. [1962,1976]

Davis, C., and Kahan, W. M. (1970), "The Rotation of Eigenvectors by a Perturbation. III," *SIAM Journal on Numerical Analysis*, 7, 1–46. [1968]

Fama, E. F., and French, K. R. (1992), "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47, 427–465. [1961]

——— (2017), "International Tests of a Five-Factor Asset Pricing Model," *Journal of Financial Economics*, 123, 441–463. [1961]

Fan, J., Fan, Y., and Lv, J. (2008), "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197. [1968]

Fan, J., Furger, A., and Xiu, D. (2016), "Incorporating Global Industrial Classification Standard Into Portfolio Allocation: A Simple Factor-Based Large Covariance Matrix Estimator With High-Frequency Data," *Journal of Business & Economic Statistics*, 34, 489–503. [1968]

Fan, J., Liao, Y., and Liu, H. (2016), "An Overview of the Estimation of Large Covariance and Precision Matrices," *The Econometrics Journal*, 19, C1–C32. [1968,1970]

Fan, J., Liao, Y., and Mincheva, M. (2011), "High Dimensional Covariance Matrix Estimation in Approximate Factor Models," *Annals of Statistics*, 39, 3320. [1968]

——— (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society*, Series B, 75, 603–680. [1961,1962,1967,1968,1969,1970]

Fan, J., Zhang, J., and Yu, K. (2012), "Vast Portfolio Selection With Gross-Exposure Constraints," *Journal of the American Statistical Association*, 107, 592–606. [1968]

Friedland, S. (1981), "Convex Spectral Functions," *Linear and Multilinear Algebra*, 9, 299–316. [1965]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd ed.), New York: Springer-Verlag. [1961]

Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables Into Principal Components," *Journal of Educational Psychology*, 24, 417. [1961]

Johnstone, I. M. (2001), "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *Annals of Statistics*, 29, 295–327. [1961]

Kong, X.-B. (2017), "On the Number of Common Factors With High-Frequency Data," *Biometrika*, 104, 397–410. [1961]

Lintner, J. (1965), "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, 47, 13–37. [1961]

Mardia, K. V., Kent, J., and Bibby, J. (1979), *Multivariate Analysis*, London: Academic Press. [1961]

Markowitz, H. (1952), "Portfolio Selection," *Journal of Finance*, 7, 77–91. [1961]

——— (1959), *Portfolio Selection*, New York: Wiley. [1961]

Mykland, P. A., and Zhang, L. (2006), "ANOVA for Diffusions and Ito Processes," *The Annals of Statistics*, 34, 1931–1963. [1962,1963]

——— (2017), "Assessment of Uncertainty in High Frequency Data: The Observed Asymptotic Variance," *Econometrica*, 85, 197–231. [1962]

Mykland, P. A., Zhang, L., and Chen, D. (2019), "The Algebra of Two Scales Estimation, and the S-TSRV: High Frequency Estimation That Is Robust to Sampling Times," *Journal of Econometrics*, 208, 101–119. [1961,1962,1963,1966,1969]

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 559–572. [1961]

Pelger, M. (2019a), "Large-Dimensional Factor Modeling Based on High-Frequency Observations," *Journal of Econometrics*, 208, 23–42. [1961,1962,1973]

——— (2019b), "Understanding Systematic Risk: A High-Frequency Approach," *Journal of Finance* (to appear). [1961,1962,1973]

Popper, K. R. (1959), *The Logic of Scientific Discovery*, Hutchinson & Co., *Logik der Forschung* first published 1935 by Verlag von Julius Springer, Vienna, Austria. [1961]

Ross, S. A. (1976), "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360. [1961]

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [1970]

Sentana, E. (2009), "The Econometrics of Mean-Variance Efficiency Tests: A Survey," *The Econometrics Journal*, 12, C65–C101. [1968]

Sharpe, W. F. (1964), "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance*, 19, 425–442. [1961]

Stock, J. H., and Watson, M. W. (1998), "Diffusion Indexes," Tech. Rep., National Bureau of Economic Research. [1961]

——— (2002), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [1961]

Tao, T. (2012), *Topics in Random Matrix Theory* (Vol. 132), Providence, RI: American Mathematical Society. [1961]

Tsing, N.-K., Fan, M. K., and Verriest, E. (1994), "On Analyticity of Functions Involving Eigenvalues," *Linear Algebra and Its Applications*, 207, 159–180. [1963]

Zhang, L., Mykland, P. A., and Y. Aït-Sahalia (2005), "A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100, 1394–1411. [1962]