Working Paper No. 15-39

Principal Component Analysis of High Frequency Data

Yacine Ait-Sahalia
Princeton University and NBER

Dacheng Xiu
University of Chicago Booth School of Business

Principal Component Analysis of High Frequency Data*

Yacine Aït-Sahalia†
Department of Economics
Princeton University and NBER

Dacheng Xiu‡
Booth School of Business
University of Chicago

This Version: October 5, 2015

Abstract

We develop the necessary methodology to conduct principal component analysis at high frequency. We construct estimators of realized eigenvalues, eigenvectors, and principal components and provide the asymptotic distribution of these estimators. Empirically, we study the high frequency covariance structure of the constituents of the S&P 100 Index using as little as one week of high frequency data at a time. The explanatory power of the high frequency principal components varies over time. During the recent financial crisis, the first principal component becomes increasingly dominant, explaining up to 60% of the variation on its own, while the second principal component drives the common variation of financial sector stocks.

KEYWORDS: Itô Semimartingale, High Frequency, Spectral Function, Eigenvalue, Eigenvector, Principal Components, Three Factor Model.

JEL CODES: C13, C14, C55, C58.

1 Introduction

Principal component analysis (PCA) is one of the most popular techniques in multivariate statistics, providing a window into any latent common structure in a large dataset. The central idea of PCA is to identify a small number of common or principal components which effectively summarize a large part of the variation of the data, and serve to reduce the dimensionality of the problem and achieve parsimony.

Classical PCA originated in Pearson (1901) and Hotelling (1933) and is widely used in macroeconomics and finance among many other fields. One example is Litterman and Scheinkman (1991), who document a

*We thank Hristo Sendov for valuable discussions on eigenvalues and spectral functions. In addition, we benefited much from comments by Torben Andersen, Tim Bollerslev, Oleg Bondarenko, Marine Carrasco, Gary Chamberlain, Kirill Evdokimov, Jianqing Fan, Christian Hansen, Jerry Hausman, Jean Jacod, Ilze Kalnina, Jia Li, Yingying Li, Oliver Linton, Nour Meddahi, Per Mykland, Ulrich Müller, Andrew Patton, Eric Renault, Jeffrey Russell, Neil Shephard, George Tauchen, Viktor Todorov, Ruey Tsay, and Xinghua Zheng, as well as seminar and conference participants at Brown University, CEMFI, Duke University, Harvard University, MIT, Northwestern University, Peking University, Princeton University, Singapore Management University, University of Amsterdam, University of Chicago, University of Illinois at Chicago, University of Tokyo, the 2015 North American Winter Meeting of the Econometric Society, the CEME Young Econometricians Workshop at Cornell, the NBER-NSF Time Series Conference in Vienna, the 10th International Symposium on Econometric Theory and Applications, the 7th Annual SoFiE Conference, the 2015 Financial Econometrics Conference in Toulouse and the 6th French Econometrics Conference.

†Address: 26 Prospect Avenue, Princeton, NJ 08540, USA. E-mail address: yacine@princeton.edu.

‡Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: dacheng.xiu@chicagobooth.edu. Xiu gratefully acknowledges financial support from the IBM Faculty Scholar Fund at the University of Chicago Booth School of Business.

three-factor structure of the term structure of yields using PCA. PCA has also been applied to analyze the dimension of volatility dynamics, e.g., Egloff, Leippold, and Wu (2010), suggesting a two-factor model of volatility, capturing the long and short-term fluctuations of the volatility term structure. Another example is the development of economic activity and inflation indices by Stock and Watson (1999) using PCA and a factor model. A sentiment measure, reflecting the optimistic or pessimistic view of investors, was created by Baker and Wurgler (2006) using the first principal component of a number of sentiment proxies, while a policy uncertainty index was created by Baker, Bloom, and Davis (2013).

The estimation of the eigenvalues of the sample covariance matrix is the key step towards PCA. Classical asymptotic results, starting with Anderson (1958) and Anderson (1963), show that eigenvalues of the sample covariance matrix are consistent and asymptotically normal estimators of the population eigenvalues, at least when the data follow a multivariate normal distribution. Even under normality, the asymptotic distribution becomes rather involved when repeated eigenvalues are present. Waternaux (1976) showed that a similar central limit result holds for simple eigenvalues as long as the distribution of the data has finite fourth moment, while Tyler (1981) obtains the asymptotic distribution of eigenvectors under more general assumptions; see also discussions in the books by Jolliffe (2002) and Jackson (2003).

The classical PCA approach to statistical inference of eigenvalues suffers from three main drawbacks. First is the curse of dimensionality. For instance, it is well known that even the largest eigenvalue is no longer consistently estimated when the cross-sectional dimension grows at the same rate as the sample size along the time domain. Second, the asymptotic theory is essentially dependent on frequency domain analysis under stationarity (and often additional) assumptions, see, e.g., Brillinger (2001). Third, the principal components are linear combinations of the data, which fail to capture potentially nonlinear patterns therein.

These three drawbacks create difficulties when PCA is employed on asset returns data. A typical portfolio may consist of dozens of stocks. For instance, a portfolio with 30 stocks has 465 parameters in their covariance matrix, if no additional structure is imposed, while one with 100 stocks contain 5,050 parameters. As a result, years of time series data are required for estimation, raising issues of survivorship bias, potential non-stationarity and parameter constancy. Moreover, asset returns are known to exhibit time-varying volatility and heavy tails, leading to deviations from the assumptions required by the classical PCA asymptotic theory. In addition, prices of derivatives, e.g., options, are often nonlinear functions of the underlying asset's price and volatility. Ruling out nonlinear combinations disconnects their principal components from the underlying factors that drive derivative returns.

These issues motivate the development in this paper of the tools necessary to conduct PCA for continuous-time stochastic processes using high frequency data and high frequency or "in-fill" asymptotics, whereby the number of observations grows within a fixed time window. This approach not only adds the important method of PCA to the high frequency toolkit, but it addresses the three drawbacks mentioned above. First, the large amount of intraday time series data vastly improve the time series vs. cross-sectional dimensionality trade-off. And these are not redundant observations for the problem at hand. Unlike for expected returns, it is well established theoretically that increasing the sampling frequency improves variance and covariance estimation, at least until market microstructure concerns start biting. Yet market microstructure noise is not a serious concern at the one minute sampling frequency we will employ empirically, for liquid stocks which typically trade on infra-second time scales. As a result of the large increase in the time series dimension, it is plausible to expect high frequency asymptotic analysis with the cross-sectional dimension fixed to serve as accurate approximations.[1] In fact, we will show both in simulations and empirically that PCA works quite well at high

[1]At low frequency, in the case of large dimensional data, Geman (1980) and Bai, Silverstein, and Yin (1988) investigated the strong law of the largest eigenvalue, when the cross-sectional dimension of the data grows at the same rate as the sample size. Johnstone (2001) further analyzed the distribution of the largest eigenvalue using random matrix theory. These analyses

frequency for a portfolio of 100 stocks using as little as one week of one-minute returns data. Second, the high frequency asymptotic framework enables nonparametric analysis of general stochastic processes, thereby allowing strong dependence, non-stationarity, and heteroscedasticity due to stochastic volatility and jumps, and freeing the analysis from the strong parametric assumptions that are required in a low frequency setting. In effect, the existence of population moments becomes irrelevant and PCA becomes applicable at high frequency for general Itô semimartingales. Third, our principal components are built upon instantaneous (or locally) linear combinations of the stochastic processes, which thereby capture general nonlinear relationships by virtue of Itô's lemma.

Within a continuous-time framework, we first develop the concept of "realized" or high-frequency PCA for data sampled from a stochastic process within a fixed time window. Realized PCA depends on realizations of eigenvalues, which are stochastic processes, evolving over time. Our implementation of realized PCA is designed to extend classical PCA from its low frequency setting into the high frequency one as seamlessly as possible. Therefore, we start by estimating realized eigenvalues from the realized covariance matrix. One technical challenge we must overcome is the fact that when an eigenvalue is simple, it is a smooth function of the instantaneous covariance matrix. On the other hand, when it comes to a repeated eigenvalue, this function is not differentiable, which hinders statistical inference, as the asymptotic theory requires at least second-order differentiability. To tackle the issue of non-smoothness of repeated eigenvalues, we propose an estimator constructed by averaging all repeated eigenvalues. Using the theory of spectral functions, we show how to obtain differentiability of the proposed estimators, and the required derivatives. We therefore construct an estimator of realized spectral functions, and develop the high frequency asymptotic theory for this estimator. Estimation of eigenvalues, principal components and eigenvectors then arise as special cases of this estimator, by selecting a specific spectral function. Aggregating local estimates results in an asymptotic bias, due to the nonlinearity of the spectral functions. The bias is fortunately of higher order, which enables us to construct a bias-corrected estimator, and show that the latter is consistent and achieves stable convergence in law to a mixed normal distribution.

As far as asymptotic theory is concerned, our estimators are constructed by aggregating functions of instantaneous covariance estimates. Other examples of this general type arise in other high frequency contexts. Jacod and Rosenbaum (2013) analyze such estimators with an application of integrated quarticity estimation. Li and Xiu (2014) develop inference theory based on the generalized method of moments to investigate structural economic models which include the instantaneous volatility as an explanatory variable. Kalnina and Xiu (2013) discuss estimation of the leverage effect measured by the integrated correlation with an additional volatility instrument. Li, Todorov, and Tauchen (2013) discuss inference theory for volatility functional dependencies. An alternative inference theory in Mykland and Zhang (2009) is designed for a class of estimators based on an aggregation of local estimates using a finite number of blocks.

Also related to the problem we study is the rank inference developed in Jacod, Lejay, and Talay (2008)

confirmed that the largest eigenvalue is no longer consistently estimated. Moreover, Johnstone and Lu (2009) prove that, for a single factor model, the estimated eigenvector corresponding to the largest eigenvalue is also inconsistent unless the sample size grows at a rate faster than the one at which the cross-sectional dimension increases. To resolve the inconsistency of the PCA in this setting, different estimators of the covariance matrix have been proposed, using banding (Bickel and Levina (2008b)), tapering (Cai and Zhou (2012)), and thresholding (Bickel and Levina (2008a)) methods, or imposing additional assumptions such as sparsity. However, the sparsity requirement of the covariance matrix does not hold empirically or a large cross-section of stock returns, see, e.g., Fan, Furger, and Xiu (2015). Alternatively, there is a new literature on methods that "sparsify" the PCA directly, i.e., imposing the sparsity on eigenvectors, see, e.g., Jolliffe, Trendafilov, and Uddin (2003), Zou, Hastie, and Tibshirani (2006), d'Aspremont, Ghaoui, Jordan, and Lanckriet (2007), Johnstone and Lu (2009), and Amini and Wainwright (2009). None of these methods are currently applicable to dependent data nor are central limit theorems available for these estimators in a large dimensional setting.

and Jacod and Podolskij (2013), where the cross-sectional dimension is also fixed and the processes follow continuous Itô semimartingales. Allowing for an increasing dimension but with an additional sparsity structure, Wang and Zou (2010) consider the estimation of integrated covariance matrix with high-dimensional and high frequency data in the presence of measurement errors. Tao, Wang, and Zhou (2013) investigate the minimax rate of convergence for covariance matrix estimation in the same setting; see also Tao, Wang, and Chen (2013) and Tao, Wang, Yao, and Zou (2011) for related work. Zheng and Li (2011) establish a Marcenko-Pastur type theorem for the spectral distribution of the integrated covariance matrix for a special class of diffusion processes. In a similar setting, Heinrich and Podolskij (2014) study the spectral distribution of empirical covariation matrices of Brownian integrals.

Since it is about PCA, this paper shares all the natural pros and cons of PCA, especially relative to factor models. There is a large low frequency literature in economics and finance on factor analysis. While factor analysis in the classical setting of low frequency and fixed dimension is a very useful tool, as in Ross (1976) for instance, Chamberlain and Rothschild (1983) point out that when the dimension grows with the sample size, PCA may be preferred due to its simplicity, and the fact that it may be employed under weaker assumptions, such as a factor structure that is only approximate. In fact, PCA has been used extensively at low frequency for analyzing factor models, to determine the number of factors, e.g. Bai and Ng (2002), or to construct and use factors for forecasting, e.g. Stock and Watson (2002), with the asymptotic theory developed by Connor and Korajczyk (1988), Stock and Watson (1998), and Bai (2003). The above factor models are static, as opposed to the dynamic factor models discussed in Forni, Hallin, Lippi, and Reichlin (2000), Forni and Lippi (2001), and Forni, Hallin, Lippi, and Reichlin (2004), in which discrete-time lagged values of the unobserved factors may also affect the observed dependent variables. By contrast, this paper develops the theory for high frequency PCA in a continuous-time setting. Our model consists of Itô semimartingales, and is fully nonparametric without any assumptions on the existence of a factor structure. That said, similar techniques as those developed here can in principle be developed in the future to estimate continuous-time models with a factor structure but this is a distinct problem from PCA (just as it is in a low frequency setting).

The paper is organized as follows. Section 2 sets up the notation and the model. Section 3 provides the estimators and the main asymptotic theory. Section 4 reports the results of Monte Carlo simulations. We then implement the method to analyze by PCA the covariance structure of the S&P 100 stocks in Section 5, where we ask whether at high frequency cross-sectional patterns in stock returns are compatible with the well-established low frequency evidence of a low-dimensional common factor structure (e.g., Fama and French (1993)). We find that, excluding jump variation, three Brownian factors explain between 50 and 60% of continuous variation of the stock returns, that their explanatory power varies over time and that during the recent financial crisis, the first principal component becomes increasingly dominant, explaining up to over 60% of the variation on its own, capturing market-wide, systemic, risk. Despite the differences in methods, time periods, and length of observation, these empirical findings at high frequency are surprisingly consistent with the well-established low frequency Fama-French results of three common factors. Of course, the design limitation of PCA is that it does lend itself easily to an identification of what the principal components are in terms of economic factors. Nevertheless, we find evidence that the first principal component shares time series characteristics with the overall market return, and, using biplots, we find that at the height of the crisis the second principal component drives the common variation of financial sector stocks. Section 6 concludes. Proofs are in the appendix.

2 The Setup

2.1 Standard Notations and Definitions

In what follows, $\mathbb{R}_d$ denotes the d-dimensional Euclidean space, and its subset $\mathbb{R}_d^+$ contains non-negative real vectors. Let e^k be the unit vector in $\mathbb{R}_d^+$, with the kth entry equal to 1, $\mathbf{1} = \sum_{k=1}^d e^k$, and $\mathbb{I}$ be the identity matrix. $\mathcal{M}_d$ denotes the Euclidean space of all $d \times d$ real-valued symmetric matrices. $\mathcal{M}_d^+$ is a subset of $\mathcal{M}_d$, which includes all positive-semidefinite matrices. We use $\mathcal{M}_d^{++}$ to denote the set of all positive-definite matrices. The Euclidean space $\mathcal{M}_d$ is equipped with an inner product $\langle A, B\rangle = \mathrm{Tr}(AB)$. We use $\|\cdot\|$ to denote the Euclidean norm for vectors or matrices, and the superscript "+" to denote the Moore-Penrose inverse of a real-matrix, see e.g., Magnus and Neudecker (1999).

All vectors are column vectors. The transpose of any matrix A is denoted by $A^\intercal$. The ith row of A is written as $A_{i,\cdot}$, and the jth column of A is $A_{\cdot,j}$. A_{ij} denotes the (i,j)th entry of A. The operator $\mathrm{diag} : \mathcal{M}_d \to \mathbb{R}_d$ is defined as $\mathrm{diag}(A) = (A_{11}, A_{22}, \ldots, A_{dd})^\intercal$. In addition, we define its inverse operator Diag, which maps a vector x to a diagonal matrix.

Let f be a function from $\mathbb{R}_d^+$ to $\mathbb{R}$. The gradient of f is written as ∂f, and its Hessian matrix is denoted as $\partial^2 f$. The derivative of a matrix function $F : \mathcal{M}_d^+ \to \mathbb{R}$ is denoted by $\partial F \in \mathcal{M}_d$, with each element written as $\partial_{ij} F$, for $1 \leq i, j \leq d$. Note that the derivative is defined in the usual sense, so that for any $A \in \mathcal{M}_d^+$, $\partial_{ij} A = J_{ij}$, where J_{ij} is a single-entry matrix with the (i,j)th entry equal to 1. The Hessian matrix of F is written as $\partial^2 F$, with each entry referred to as $\partial^2_{jk,lm} F$, for $1 \leq j, k, l, m \leq d$. We use ∂^k to denote kth order derivatives and $\delta_{i,j}$ to denote the Kronecker's delta function giving 1 if $i = j$ or 0 otherwise. A function f is Lipchitz if there exists a constant K such that $|f(x+h) - f(x)| \leq K \|h\|$. In the proof, K is a generic constant which may vary from line to line. A function with kth continuous derivatives is denoted a C^k function.

Data are sampled discretely every Δ_n units of time. All limits are taken as $\Delta_n \to 0$. "$\overset{\text{u.c.p.}}{\Longrightarrow}$" denotes uniformly on compacts in probability, and "$\xrightarrow{\text{p}}$" denotes convergence in probability. We use "$\xrightarrow{\mathcal{L}-\text{s}}$" to denote stable convergence in law. We write $a_n \asymp b_n$ if for some $c \geq 1$, $b_n/c \leq a_n \leq c b_n$ for all n. Finally, $[\cdot,\cdot]$ denotes the quadratic covariation between Itô semimartingales, and $[\cdot,\cdot]^c$ the continuous part thereof.

2.2 Eigenvalues and Eigenvectors

We now collect some preliminary results about eigenvalues and eigenvectors. For any vector $x \in \mathbb{R}_d^+$, $\bar{x}$ denotes the vector with the same entries as x, ordered in a non-increasing order. We use $\bar{\mathbb{R}}_d^+$ to denote the subset of $\mathbb{R}_d^+$ containing vectors x satisfying $x_1 \geq x_2 \geq \ldots \geq x_d \geq 0$.

By convention, for any $x \in \bar{\mathbb{R}}_d^+$, we can write:

$$x_1 = \ldots = x_{g_1} > x_{g_1+1} = \ldots = x_{g_2} > \ldots x_{g_{r-1}} > x_{g_{r-1}+1} = \ldots = x_{g_r} \geq 0, \tag{1}$$

where $g_r = d$, and r is the number of distinct element. $\{g_1, g_2, \ldots, g_r\}$ depends on x. We then define a corresponding partition of indices as $I_j = \{g_{j-1}+1, g_{j-1}+2, \ldots, g_j\}$, for $j = 1, 2, \ldots, r$.

For any $A \in \mathcal{M}_d^+$, $\lambda(A) = (\lambda_1(A), \lambda_2(A), \ldots, \lambda_d(A))^\intercal$ is the vector of its eigenvalues in a non-increasing order. This notation also permits us to consider λ as a mapping from $\mathcal{M}_d^+$ to $\bar{\mathbb{R}}_d^+$. An important result establishing the continuity of λ is, see, e.g., Tao (2012):

Lemma 1. *$\lambda : \mathcal{M}_d^+ \to \bar{\mathbb{R}}_d^+$ is Lipchitz.*

Associated with any eigenvalue λ_g of $A \in \mathcal{M}_d^+$, we denote its eigenvector as γ_g, which satisfies $A\gamma_g = \lambda_g \gamma_g$, and $\gamma_g^\intercal \gamma_g = 1$. The eigenvector, apart from its sign, is uniquely defined when λ_g is simple. In such a case, without loss of generality we require the first non-zero element of the eigenvector to be positive. In the presence

of repeated eigenvalues, the eigenvector is determined up to an orthogonal transformation. In any case, we can choose eigenvectors such that for any $g \neq h$, $\gamma_g^\intercal \gamma_h = 0$, see, e.g., Anderson (1958).

When λ_g is a simple root, we regard γ_g as another vector-valued function of A. It turns out in this case, both $\lambda_g(\cdot)$ and $\gamma_g(\cdot)$ are infinitely smooth at A, see, e.g., Magnus and Neudecker (1999).

Lemma 2. *Suppose λ_g is a simple root of $A \in \mathcal{M}_d^+$, then $\lambda_g : \mathcal{M}_d^+ \to \bar{\mathbb{R}}^+$ and $\gamma_g : \mathcal{M}_d^+ \to \mathbb{R}_d$ are C^∞ at A. Moreover, we have*

$$\partial_{jk}\lambda_g(A) = \gamma_{gj}(A)\gamma_{gk}(A), \quad \textit{and} \quad \partial_{jk}\gamma_g(A) = (\lambda_g \mathbb{I} - A)^+_{\cdot,j}\gamma_{gk}(A),$$

where γ_{gk} is the kth entry of γ_g. If in addition all the eigenvalues of A are simple, with $(\gamma_1, \gamma_2, \ldots, \gamma_d)$ being the corresponding eigenvectors, then

$$\begin{aligned}
\partial^2_{jk,lm}\gamma_{gh} = &- \sum_{p\neq g} \frac{1}{(\lambda_g - \lambda_p)^2} \left(\gamma_{gl}\gamma_{gm}\gamma_{ph}\gamma_{pj}\gamma_{gk} - \gamma_{pl}\gamma_{pm}\gamma_{ph}\gamma_{pj}\gamma_{gk}\right) \\
&+ \sum_{p\neq g}\sum_{q\neq p} \frac{1}{(\lambda_g - \lambda_p)(\lambda_p - \lambda_q)} \gamma_{ql}\gamma_{pm}\gamma_{qh}\gamma_{pj}\gamma_{gk} \\
&+ \sum_{p\neq g}\sum_{q\neq p} \frac{1}{(\lambda_g - \lambda_p)(\lambda_p - \lambda_q)} \gamma_{ql}\gamma_{pm}\gamma_{qj}\gamma_{ph}\gamma_{gk} \\
&+ \sum_{p\neq g}\sum_{q\neq g} \frac{1}{(\lambda_g - \lambda_p)(\lambda_g - \lambda_q)} \gamma_{qk}\gamma_{ql}\gamma_{gm}\gamma_{ph}\gamma_{pj}.
\end{aligned}$$

In general, while the eigenvalue is always a continuous function, the eigenvector associated with a repeated root is not necessarily continuous. In what follows, we only consider estimation of an eigenvector when it is associated with a simple eigenvalue.

2.3 Dynamics of the Variable

The process we analyze is a general d-dimensional Itô semimartingale, defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ with the following Grigelionis representation:

$$X_t = X_0 + \int_0^t b_s\, ds + \int_0^t \sigma_s dW_s + (\delta 1_{\|\delta\|\leq 1}) * (\mu - \nu)_t + (\delta 1_{\{\|\delta\|>1\}}) * \mu_t, \tag{2}$$

where W is a d-dimensional Brownian motion, μ is a Poisson random measure on $\mathbb{R}^+ \times \mathbb{R}_d$ with the compensator $\nu(dt, dx) = dt \otimes \bar{\nu}(dx)$, and $\bar{\nu}$ is a σ-finite measure. More details on high frequency models and asymptotics can be found in the book Aït-Sahalia and Jacod (2014).

The volatility process σ_s is càdlàg, $c_s = (\sigma\sigma^\intercal)_s \in \mathcal{M}_d^+$, for any $0 \leq s \leq t$. We denote d non-negative eigenvalues of c_s by $\lambda_{1,s} \geq \lambda_{2,s} \geq \ldots \geq \lambda_{d,s}$, summarized in a vector λ_s. As is discussed above, we sometimes write $\lambda_s = \lambda(c_s)$, regarding $\lambda(\cdot)$ as a function of c_s. By Lemma 1, $\lambda(\cdot)$ is a continuous function so that $\lambda(c_s)$ is càdlàg.

2.4 Principal Component Analysis at High Frequency

Similar to the classical PCA, the PCA procedure in this setting consists in searching repeatedly for instantaneous linear combinations of X, namely principal components, which maximize certain measure of variation, while being orthogonal to the principal components already constructed, at any time between 0 and t. In contrast to the classical setting, where one maximizes the variance of the combination (see, e.g., Anderson (1958)), the criterion here is the continuous part of the quadratic variation.

Lemma 3. *Suppose that X is a d-dimensional vector-valued process described in (2). Then there exists a sequence of $\{\lambda_{g,s}, \gamma_{g,s}\}_{1\leq g\leq d, 0\leq s\leq t}$, such that*

$$c_s\gamma_{g,s} = \lambda_{g,s}\gamma_{g,s}, \quad \gamma_{g,s}^\intercal\gamma_{g,s} = 1, \quad \textit{and} \quad \gamma_{h,s}^\intercal c_s\gamma_{g,s} = 0,$$

where $\lambda_{1,s} \geq \lambda_{2,s} \geq \ldots \geq \lambda_{d,s} \geq 0$. Moreover, for any càdlàg and vector-valued adapted process γ_s, such that $\gamma_s^\intercal\gamma_s = 1$, and $\gamma_s^\intercal c_s\gamma_{h,s} = 0$, $1 \leq h \leq g-1$,

$$\int_0^u \lambda_{g,s}ds \geq \left[\int_0^u \gamma_{s-}^\intercal dX_s, \int_0^u \gamma_{s-}^\intercal dX_s\right]^c, \quad \textit{for any } 0 \leq u \leq t.$$

When $\lambda_{g,s}$ is a simple root of c_s between 0 and t, then $\gamma_{g,s}$ is an adapted càdlàg process, due to the continuity of $\gamma_g(\cdot)$ by Lemma 2, so that we can construct its corresponding principal component $\int_0^t \gamma_{g,s-}^\intercal dX_s$.

Given Lemma 3, we can define our parameters of interest: the realized eigenvalue, i.e. $\int_0^t \lambda_s ds$; the realized principal components associated with some simple root λ_g, i.e. $\int_0^t \gamma_{g,s-}^\intercal dX_s$. It may also be worth investigating the average loading on the principal component, i.e. the realized eigenvector, which is $\int_0^t \gamma_{g,s}ds$. Since the definitions of these quantities all rely on integrals, we call them integrated eigenvalues, integrated principal components, and integrated eigenvectors.

The above analysis naturally leads us to consider inference for $\int_0^t \lambda(c_s)ds$, for which the differentiability of $\lambda(\cdot)$ is critical: we start with the convergence of an estimator $\hat{c}_s$ to c_s, from which we need in a delta method sense to infer the convergence of the integrated eigenvalues estimator. A simple eigenvalue is C^∞-differentiable, but for repeated eigenvalues this is not necessarily the case. For this reason, in order to fully address the estimation problem, we need to introduce spectral functions.

2.5 Spectral Functions

A real-valued function F defined on a subset of $\mathcal{M}_d^+$ is called a spectral function (see, e.g., Friedland (1981)) if for any orthogonal matrix O in $\mathcal{M}_d$ and X in $\mathcal{M}_d^+$, $F(X) = F(O^\intercal XO)$. We describe sets in $\mathbb{R}_d^+$ and functions from $\mathbb{R}_d^+$ to $\mathbb{R}^+$ as symmetric if they are invariant under coordinate permutations. That is, for any symmetric function f with an open symmetric domain in $\mathbb{R}_d^+$, we have $f(x) = f(Px)$ for any permutation matrix $P \in \mathcal{M}_d$. Associated with any spectral function F, we define a function f on $\mathbb{R}_d^+$, so that $f(x) = F(\text{Diag}(x))$. Since permutation matrices are orthogonal, f is symmetric, and $f \circ \lambda = F$, where $\circ$ denotes function composition.

We will need to differentiate the spectral function F. We introduce a matrix function associated with the corresponding symmetric function f of F, which, for any $x \in \bar{\mathbb{R}}_d^+$ in the form of (1), is given by

$$\mathcal{A}_{p,q}^f(x) = \begin{cases} 0 & \text{if} \quad p = q; \\ \partial_{pp}^2 f(x) - \partial_{qq}^2 f(x) & \text{if} \quad p \neq q, \quad \text{and} \quad p, q \in I_l; \\ (\partial_p f(x) - \partial_q f(x))/(x_p - x_q) & \text{otherwise.} \end{cases}$$

for some $l = \{1, 2, \ldots, r\}$.

The following lemma collects some known and useful results regarding the continuous differentiability and convexity of spectral functions, which we will use below:

Lemma 4. *The symmetric function f is twice continuously differentiable at a point $\lambda(A) \in \bar{\mathbb{R}}_d^+$ if and only if the spectral function $F = f \circ \lambda$ is twice continuously differentiable at the point $A \in \mathcal{M}_d^+$. The gradient and the Hessian matrix are given below:*

$$\partial_{jk}(f \circ \lambda)(A) = \sum_{p=1}^d O_{pj}\partial_p f(\lambda(A))O_{pk},$$

$$\partial^2_{jk,lm}(f\circ\lambda)(A)=\sum_{p,q=1}^{d}\partial^2_{pq}f(\lambda(A))O_{pl}O_{pm}O_{qj}O_{qk}+\sum_{p,q=1}^{d}\mathcal{A}^f_{pq}(\lambda(A))O_{pl}O_{pj}O_{qk}O_{qm},$$

where O is any orthogonal matrix that satisfies $A=O^{\intercal}Diag\,(\lambda(A))\,O$. More generally, f is a C^k function at $\lambda(A)$ if and only if F is C^k at A, for any $k=0,1,\ldots,\infty$. In addition, f is a convex function if and only if F is convex.

Next, we note that both simple and repeated eigenvalues can be viewed as special cases of spectral functions.

Example 1 (A Simple Eigenvalue)**.** *Suppose the kth eigenvalue of $A\in\mathcal{M}_d^+$ is simple, that is, $\lambda_{k-1}(A)>\lambda_k(A)>\lambda_{k+1}(A)$. Define for any $x\in\mathbb{R}_d^+$,*

$$f(x)=\textit{the kth largest entry in } x=\bar{x}_k.$$

Apparently, f is a symmetric function, and it is C^∞ at any point $y\in\bar{\mathbb{R}}_d^+$, with $y_{k-1}>y_k>y_{k+1}$. Indeed, $\partial f(y)=e^k$, and $\partial^l f(y)=0$ for $l\geq 2$. By Lemma 4, $\lambda_k(A)=(f\circ\lambda)(A)$ is C^∞ at A.

Example 2 (Non-Simple Eigenvalues)**.** *Suppose the eigenvalues of $A\in\mathcal{M}_d^+$ satisfy:*

$$\lambda_{g_{l-1}}(A)>\lambda_{g_{l-1}+1}(A)\geq\ldots\geq\lambda_{g_l}(A)>\lambda_{g_{l+1}}(A),$$

for some $1\leq g_{l-1}<g_l\leq d$. By convention, when $g_l=d$, the last " $>$" above is not used. Consider the following function f, evaluated at $x\in\mathbb{R}_d^+$, which is given by

$$f(x)=\frac{1}{g_l-g_{l-1}}\sum_{j=g_{l-1}+1}^{g_l}\bar{x}_j.$$

It is easy to verify that f is symmetric, and C^∞ at any point $y\in\bar{\mathbb{R}}_d^+$ that satisfies $y_{g_{l-1}}>y_{g_{l-1}+1}\geq\ldots\geq y_{g_l}>y_{g_{l+1}}$. Moreover, $\partial f(y)=\frac{1}{g_l-g_{l-1}}\sum_{k=g_{l-1}+1}^{g_l}e^k$, and $\partial^l f(y)=0$, for any $l\geq 2$. As a result of Lemma 4, the corresponding spectral function,

$$F(A)=(f\circ\lambda)(A)=\frac{1}{g_l-g_{l-1}}\sum_{j=g_{l-1}+1}^{g_l}\lambda_j(A),$$

is C^∞ at A. In the special case where

$$\lambda_{g_{l-1}}(A)>\lambda_{g_{l-1}+1}(A)=\ldots=\lambda_{g_l}(A)>\lambda_{g_{l+1}}(A),$$

i.e., there is a repeated eigenvalue, and $F(A)=\lambda_{g_{l-1}+1}(A)=\ldots=\lambda_{g_l}(A)$. By contrast, $\lambda_j(A)$ is not differentiable, for any $g_{l-1}+1\leq j\leq g_l$.

Example 3 (Trace and Determinant)**.** *For any $x\in\mathbb{R}_d^+$, define*

$$f_1(x)=\sum_{j=1}^{d}x_j,\quad\textit{and}\quad f_2(x)=\prod_{j=1}^{d}x_j.$$

Both f_1 and f_2 are symmetric and C^∞. Therefore, for any $A\in\mathcal{M}_d^+$, $\mathrm{Tr}(A)=(f_1\circ\lambda)(A)$ and $det(A)=(f_2\circ\lambda)(A)$ are C^∞ at A.

The previous examples make it clear that the objects of interest, eigenvalues, are special cases of spectral functions, whether they are simple or repeated. (We are also able to estimate the trace and determinant "for

free".) Lemma 4 links the differentiability of a spectral function F, which is needed for statistical inference, to that of its associated symmetric function f. The key advantage is that differentiability of f is easier to establish.

Before turning to inference, we prove a useful result that characterizes the topology of the set of matrices with special eigenvalue structures. The domain of spectral functions we consider will be confined to this set in which these spectral functions are smooth.

Lemma 5. *For any $1 \leq g_1 < g_2 < \ldots < g_r \leq d$, the set*

$$\mathcal{M}(g_1, g_2, \ldots, g_r) = \left\{ A \in \mathcal{M}_d^{++} \mid \lambda_{g_l}(A) > \lambda_{g_l+1}(A), \text{ for any } \ l = 1, 2, \ldots, r-1 \right\} \tag{3}$$

is dense and open in $\mathcal{M}_d^{++}$. In particular, the set of positive-definite matrices with distinct eigenvalues, i.e., $\mathcal{M}(1, 2, \ldots, d)$, is dense and open in $\mathcal{M}_d^{++}$.

We conclude this section with some additional notations. First, we introduce a symmetric open subset of $\bar{\mathbb{R}}^+/\{0\}$, the image of $\mathcal{M}(g_1, g_2, \ldots, g_r)$ under $\lambda(\cdot)$:

$$\mathcal{D}(g_1, g_2, \ldots, g_r) = \left\{ x \in \mathbb{R}_d^+/\{0\} \mid \bar{x}_{g_l} > \bar{x}_{g_l+1}, \text{ for any } \ l = 1, 2, \ldots, r-1 \right\}. \tag{4}$$

We also introduce a subset of $\mathcal{M}(g_1, g_2, \ldots, g_r)$, in which our spot covariance matrix c_t will take values.

$$\begin{aligned} &\mathcal{M}^*(g_1, g_2, \ldots, g_r) = \Big\{ A \in \mathcal{M}_d^{++} \mid \\ &\lambda_1(A) = \ldots = \lambda_{g_1}(A) > \lambda_{g_1+1}(A) = \ldots = \lambda_{g_2}(A) > \ldots \lambda_{g_{r-1}}(A) > \ \lambda_{g_{r-1}+1}(A) = \ldots = \lambda_{g_r}(A) \Big\}. \end{aligned}$$

Finally, we introduce the following set, which becomes relevant if we are only interested in a simple eigenvalue λ_g:

$$\mathcal{M}(g) = \left\{ A \in \mathcal{M}_d^{++} \mid \lambda_{g-1}(A) > \lambda_g(A) > \lambda_{g+1}(A) \right\}, \text{ and } \mathcal{D}(g) = \lambda(\mathcal{M}(g)).$$

By convention, we ignore the first (resp. second) inequality in the definition of $\mathcal{M}(g)$ when $g = 1$ (resp. $g = d$).

3 Estimators and Asymptotic Theory

3.1 Assumptions

We start with the standard assumption on the process X:[2]

Assumption 1. *The drift term b_t is progressively measurable and locally bounded. The spot covariance matrix $c_t = (\sigma\sigma^\intercal)_t$ is an Itô semimartingale. Moreover, for any $\gamma \in [0, 1)$, there is a sequence of stopping times (τ_n) increasing to ∞, and a deterministic function $\bar{\delta}_n$ such that $\int_{\mathbb{R}_d} \bar{\delta}_n(x)^\gamma \bar{\nu}(dx) < \infty$ and that $\|\delta(\omega, t, x)\| \wedge 1 \leq \bar{\delta}_n(x)$, for all (ω, t, x) with $t \leq \tau_n(\omega)$.*

In view of the previous examples, our main theory is tailored to the statistical inference on $\int_0^t F(c_s)ds$. Thanks to Lemma 4, we can make assumptions directly on f instead of F, which are much easier to verify.

Assumption 2. *Suppose F is a vector-valued spectral function, and f is the corresponding vector-valued symmetric function such that $F = f \circ \lambda$. f is a continuous function, and satisfies $\|f(x)\| \leq K(1 + \|x\|^\zeta)$, for some $\zeta > 0$.*

[2]Using the notation on Page 583 of Jacod and Protter (2011), Assumption 1 states that the process X satisfies Assumption (H-1) and that c_t satisfies Assumption (H-2).

In all the examples of Section 2.5, this assumption holds. By Lemma 1 and Assumption 2, $F(c_s)$ is càdlàg, and the integral $\int_0^t F(c_s)ds$ is well-defined.

The above assumptions are sufficient to ensure the desired consistency of estimators we will propose. Additional assumptions are required to establish their asymptotic distribution.

Assumption 3. *There exists some open and convex set $\mathcal{C}$, such that its closure $\bar{\mathcal{C}} \subset \mathcal{M}(g_1, g_2, \ldots, g_r)$, where $1 \leq g_1 < g_2 < \ldots < g_r \leq d$, and that for any $0 \leq s \leq t$, $c_s \in \mathcal{C} \cap \mathcal{M}^*(g_1, g_2, \ldots, g_r)$. Moreover, f is C^3 on $\mathcal{D}(g_1, g_2, \ldots, g_r)$.*

Assumption 3, in particular, $c_s \in \mathcal{M}^*(g_1, g_2, \ldots, g_r)$, guarantees that different groups of eigenvalues do not cross over within $[0, t]$. This condition is mainly used to deliver the joint central limit theorem for spectral functions that depend on all eigenvalues, although it may not be necessary for some special cases, such as $\det(\cdot)$ and $\mathrm{Tr}(\cdot)$, which are smooth everywhere. The convexity condition on $\mathcal{C}$ is in principle not difficult to satisfy, given that $\mathcal{M}(g_1, g_2, \ldots, g_r)$ can be embedded into some Euclidean space of real vectors. This condition is imposed to ensure that the domain of the spectral function can be restricted to certain neighborhood of $\{c_s\}_{0 \leq s \leq t}$, in which the function is smooth and the mean-value theorem can be applied. This assumption is not needed if c_t is continuous.

It is also worth mentioning that all eigenvalues of c_t being distinct is a special case, which is perhaps the most relevant scenario in practice, as is clear from Lemma 5. Even in this scenario, Assumption 3 is required (with $g_1, g_2, \ldots, g_r$ being chosen as $1, 2, \ldots, d$), because the eigenvalue function $\lambda(\cdot)$ is not everywhere differentiable.

Assumption 3 resembles the spacial localization assumption in Li, Todorov, and Tauchen (2014) and Li and Xiu (2014), which is different from the polynomial growth conditions proposed by Jacod and Rosenbaum (2013). The growth conditions are not satisfied in our setting, when the difference between two groups of repeated eigenvalues approaches zero.

If we are only interested in the spectral function that depends on one simple eigenvalue, e.g., the largest and simple integrated eigenvalue, $\int_0^t \lambda_1(c_s)ds$, then it is only necessary to ensure that $\lambda_1(c_s) > \lambda_2(c_s)$, for $0 \leq s \leq t$, regardless of whether the remaining eigenvalues are simple or not. We thereby introduce the following assumption for this scenario, which is much weaker than Assumption 3.

Assumption 4. *There exists some open and convex set $\mathcal{C}$, such that $\bar{\mathcal{C}} \subset \mathcal{M}(g)$, for some $g = 1, 2, \ldots, d$, and that for any $0 \leq s \leq t$, $c_s \in \mathcal{C}$. Moreover, f is C^3 on $\mathcal{D}(g)$.*

3.2 Realized Spectral Functions

We now turn to the construction of the estimators. To estimate the integrated spectral function, we start with estimation of the spot covariance matrix. Suppose we have equidistant observations on X over the interval $[0, t]$, separated by a time interval Δ_n. We form non-overlapping blocks of length $k_n\Delta_n$. At each $ik_n\Delta_n$, we estimate $c_{ik_n\Delta_n}$ by

$$\widehat{c}_{ik_n\Delta_n} = \frac{1}{k_n\Delta_n} \sum_{j=1}^{k_n} \left(\Delta^n_{ik_n+j}X\right)\left(\Delta^n_{ik_n+j}X\right)^{\intercal} 1_{\left\{\|\Delta^n_{ik_n+j}X\| \leq u_n\right\}}, \tag{5}$$

where $u_n = \alpha\Delta_n^{\varpi}$, and $\Delta^n_l X = X_{l\Delta_n} - X_{(l-1)\Delta_n}$. Choices of α and ϖ are standard in the literature (see, e.g., Aït-Sahalia and Jacod (2014)) and are discussed below when implemented in simulations.

We then estimate eigenvalues of $\widehat{c}_{ik_n\Delta_n}$ by solving for the roots of $|\widehat{c}_{ik_n\Delta_n} - \lambda\mathbb{I}| = 0$. Using the notation in Lemma 1, we have $\lambda(\widehat{c}_{ik_n\Delta_n}) = \widehat{\lambda}_{ik_n\Delta_n}$. Almost surely, the eigenvalues stacked in $\widehat{\lambda}_{ik_n\Delta_n}$ are distinct (see,

e.g., Okamoto (1973)) so that we have $\widehat{\lambda}_{1,ik_n\Delta_n} > \widehat{\lambda}_{2,ik_n\Delta_n} > \ldots > \widehat{\lambda}_{d,ik_n\Delta_n}$. Our proposed estimator of the integrated spectral function is then given by[3]

$$V(\Delta_n, X; F) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} f\left(\widehat{\lambda}_{ik_n\Delta_n}\right). \tag{6}$$

We start by establishing consistency of this estimator:

Theorem 1. *Suppose Assumptions 1 and 2 hold. Also, either $\zeta \leq 1$, or $\zeta > 1$ with $\varpi \in [\frac{\zeta-1)}{2\zeta-\gamma}, \frac{1}{2})$ holds. Then the estimator (6) is consistent. As $k_n \to \infty$ and $k_n\Delta_n \to 0$,*

$$V(\Delta_n, X; F) \overset{\text{u.c.p.}}{\Longrightarrow} \int_0^t F(c_s)ds. \tag{7}$$

Next, obtaining a central limit theorem for the estimator is more involved, as there is a second-order asymptotic bias associated with the estimator (6), a complication that is also encountered in other situations such as the quarticity estimator in Jacod and Rosenbaum (2013), the GMM estimator in Li and Xiu (2014), or the leverage effect estimator by Kalnina and Xiu (2013). The bias here is characterized as follows:

Proposition 1. *Suppose Assumptions 1, 2, and 3 hold. In addition, $k_n \asymp \Delta_n^{-\varsigma}$ and $u_n \asymp \Delta_n^{\varpi}$ for some $\varsigma \in (\frac{\gamma}{2}, \frac{1}{2})$ and $\varpi \in [\frac{1-\varsigma}{2-\gamma}, \frac{1}{2})$. As $\Delta_n \to 0$, we have*

$$k_n\left(V(\Delta_n, X; F) - \int_0^t F(c_s)ds\right) \xrightarrow{\text{p}} \frac{1}{2}\sum_{j,k,l,m=1}^{d}\int_0^t \partial^2_{jk,lm}F(c_s)\left(c_{jl,s}c_{km,s} + c_{jm,s}c_{kl,s}\right)ds. \tag{8}$$

The characterization of the bias in (8) suggests a bias-corrected estimator as follows:

$$\begin{aligned}\widetilde{V}(\Delta_n, X; F) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \Big\{ & F(\widehat{c}_{ik_n\Delta_n}) \\ & -\frac{1}{2k_n}\sum_{j,k,l,m=1}^{d}\partial^2_{jk,lm}F(\widehat{c}_{ik_n\Delta_n})\left(\widehat{c}_{jl,ik_n\Delta_n}\widehat{c}_{km,ik_n\Delta_n} + \widehat{c}_{jm,ik_n\Delta_n}\widehat{c}_{kl,ik_n\Delta_n}\right)\Big\}.\end{aligned} \tag{9}$$

We then derive the asymptotic distribution of the bias-corrected estimator:

Theorem 2. *Suppose Assumptions 1, 2, and 3 hold. In addition, $k_n \asymp \Delta_n^{-\varsigma}$ and $u_n \asymp \Delta_n^{\varpi}$ for some $\varsigma \in (\frac{\gamma}{2}, \frac{1}{2})$ and $\varpi \in [\frac{1-\varsigma}{2-r}, \frac{1}{2})$. As $\Delta_n \to 0$, we have*

$$\frac{1}{\sqrt{\Delta_n}}\left(\widetilde{V}(\Delta_n, X; F) - \int_0^t F(c_s)ds\right) \xrightarrow{\mathcal{L}-\text{s}} \mathcal{W}_t, \tag{10}$$

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

$$\text{E}(\mathcal{W}_{p,t}\mathcal{W}_{q,t}|\mathcal{F}) = \int_0^t \sum_{j,k,l,m=1}^{d}\partial_{jk}F_p(c_s)\partial_{lm}F_q(c_s)\left(c_{jl,s}c_{km,s} + c_{jm,s}c_{kl,s}\right)ds. \tag{11}$$

Remark 1. *As previously noted, in the case when the spectral function F only depends on a simple eigenvalue λ_g, the same result holds under the weaker Assumption 4 instead of 3.*

[3]We prefer this estimator to the alternative one using overlapping windows, because the overlapping implementation runs much slower. In a finite sample, both estimators have a decent performance.

A feasible implementation of this distribution requires an estimator of the asymptotic variance, which we construct as follows:

Proposition 2. *The asymptotic variance of $\widetilde{V}(\Delta_n, X; F)$ can be estimated consistently by:*

$$k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \sum_{j,k,l,m=1}^{d} \partial_{jk}F_p(\widehat{c}_{ik_n\Delta_n})\partial_{lm}F_q(\widehat{c}_{ik_n\Delta_n}) \left(\widehat{c}_{jl,ik_n\Delta_n}\widehat{c}_{km,ik_n\Delta_n} + \widehat{c}_{jm,ik_n\Delta_n}\widehat{c}_{kl,ik_n\Delta_n}\right)$$

$$\xrightarrow{\text{p}} \int_0^t \sum_{j,k,l,m=1}^{d} \partial_{jk}F_p(c_s)\partial_{lm}F_q(c_s)\left(c_{jl,s}c_{km,s} + c_{jm,s}c_{kl,s}\right) ds. \tag{12}$$

3.3 Realized Eigenvalues

We next specialize the previous theorem to obtain a central limit theorem for the realized eigenvalue estimators. With the structure of eigenvalues in mind, we use a particular vector-valued spectral function F^λ, tailor-made to deliver the asymptotic theory we need for realized eigenvalues:

$$F^\lambda(\cdot) = \left(\frac{1}{g_1}\sum_{j=1}^{g_1}\lambda_j(\cdot), \frac{1}{g_2-g_1}\sum_{j=g_1+1}^{g_2}\lambda_j(\cdot), \ldots, \frac{1}{g_r-g_{r-1}}\sum_{j=g_{r-1}+1}^{g_r}\lambda_j(\cdot)\right)^\intercal. \tag{13}$$

Apparently, if a group contains only one λ_j, then the corresponding entry of F^λ is equal to this single eigenvalue; if within certain group, all eigenvalues are identical, then the corresponding entry of F^λ yields the common eigenvalue of the group.

Corollary 1. *Suppose $k_n \asymp \Delta_n^{-\varsigma}$ and $u_n \asymp \Delta_n^{\varpi}$ for some $\varsigma \in (\frac{\gamma}{2}, \frac{1}{2})$ and $\varpi \in [\frac{1-\varsigma}{2-r}, \frac{1}{2})$.*
(i) Under Assumption 1, the estimator of integrated eigenvalue vector given by

$$V(\Delta_n, X; \lambda) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \lambda(\widehat{c}_{ik_n\Delta_n}) \tag{14}$$

is consistent.
(ii) If Assumption 3 further holds, the estimator corresponding to the pth entry of F^λ given by (9) can be written explicitly as:

$$\widetilde{V}(\Delta_n, X; F_p^\lambda) = \frac{k_n\Delta_n}{g_p - g_{p-1}} \sum_{i=0}^{[t/(k_n\Delta_n)]} \sum_{h=g_{p-1}+1}^{g_p} \left\{\widehat{\lambda}_{h,ik_n\Delta_n} - \frac{1}{k_n}\text{Tr}\left((\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n})^+\widehat{c}_{ik_n\Delta_n}\right)\widehat{\lambda}_{h,ik_n\Delta_n}\right\}.$$

The joint central limit theorem for $\widetilde{V}(\Delta_n, X; F^\lambda) = \left(\widetilde{V}(\Delta_n, X; F_1^\lambda), \ldots, \widetilde{V}(\Delta_n, X; F_r^\lambda)\right)^\intercal$ is given by

$$\frac{1}{\sqrt{\Delta_n}}\left(\widetilde{V}(\Delta_n, X; F^\lambda) - \int_0^t F^\lambda(c_s)ds\right) \xrightarrow{\mathcal{L}-\text{s}} \mathcal{W}_t^\lambda, \tag{15}$$

where $\mathcal{W}^\lambda$ is a continuous process defined on an extension of the original probability space, which conditionally on $\mathcal{F}$, is continuous centered Gaussian martingale with a diagonal covariance matrix given by

$$\text{E}(\mathcal{W}_t^\lambda(\mathcal{W}_t^\lambda)^\intercal|\mathcal{F}) = \begin{pmatrix} \frac{2}{g_1}\int_0^t \lambda_{g_1,s}^2 ds & & & \\ & \frac{2}{g_2-g_1}\int_0^t \lambda_{g_2,s}^2 ds & & \\ & & \ddots & \\ & & & \frac{2}{g_r-g_{r-1}}\int_0^t \lambda_{g_r,s}^2 ds \end{pmatrix}, \tag{16}$$

where $F^{\lambda}(c_s) = (\lambda_{g_1,s}, \lambda_{g_2,s}, \ldots, \lambda_{g_r,s})^{\intercal}$.
(iii) Under Assumptions 1 and 4, our estimator with respect to the gth simple eigenvalue $\lambda_g(\cdot)$*, is given by*

$$\widetilde{V}(\Delta_n, X; \lambda_g) = k_n \Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \left\{ \widehat{\lambda}_{g,ik_n\Delta_n} - \frac{1}{k_n} \mathrm{Tr}\left((\widehat{\lambda}_{g,ik_n\Delta_n} - \widehat{c}_{ik_n\Delta_n})^{+} \widehat{c}_{ik_n\Delta_n} \right) \widehat{\lambda}_{g,ik_n\Delta_n} \right\}, \tag{17}$$

which satisfies:

$$\frac{1}{\sqrt{\Delta_n}} \left(\widetilde{V}(\Delta_n, X; \lambda_g) - \int_0^t F^{\lambda}(c_s) ds \right) \xrightarrow{\mathcal{L}-\mathrm{s}} \mathcal{W}_t^{\lambda_g}, \tag{18}$$

where $\mathcal{W}^{\lambda_g}$ *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 variance* $\int_0^t \lambda_{g,s}^2 ds$*.*

Remark 2. *Corollary 1 is the analogue in our context to the classical results on asymptotic theory for PCA by Anderson (1963), who showed that*

$$\sqrt{n}(\widehat{\lambda} - \lambda) \xrightarrow{d} N\left(0, 2Diag\left(\lambda_1^2, \lambda_2^2, \ldots, \lambda_d^2\right)\right). \tag{19}$$

where $\widehat{\lambda}$ *and* λ *are the vectors of eigenvalues of the sample and population covariance matrices and* λ *is simple. Note the similarity between (19) and (16) in the special case where the eigenvalues are simple (in which case* $g_i = i$*) and are constant instead of being stochastic. The classical setting requires the strong assumption that the covariance matrix follows a Wishart Distribution. By contrast, our results are fully nonparametric, relying instead on high frequency asymptotics.*

Remark 3. *If there are eigenvalues of* c_s *that are equal to 0 for any* $0 \leq s \leq t$*, then our estimator is super-efficient. This is an interesting case as the rank of the covariance matrix is determined by the number of non-zero eigenvalues, see, e.g., the rank inference in Jacod, Lejay, and Talay (2008) and Jacod and Podolskij (2013).*

3.4 Realized Principal Components

As we have remarked before, when eigenvalues are simple, the corresponding eigenvectors are uniquely determined. Therefore, we can estimate the instantaneous eigenvectors together with eigenvalues, which eventually leads us to construct the corresponding principal component:

Proposition 3. *Suppose Assumptions 1 and 4 hold. In addition,* $\gamma_{g,s}$ *is a vector-valued function that corresponds to the eigenvector of* c_s *with respect to a simple root* $\lambda_{g,s}$*. Suppose* $k_n \asymp \Delta_n^{-\varsigma}$ *and* $u_n \asymp \Delta_n^{\varpi}$ *for some* $\varsigma \in (\frac{\gamma}{2}, \frac{1}{2})$ *and* $\varpi \in [\frac{1-\varsigma}{2-r}, \frac{1}{2})$*. Then we have as* $\Delta_n \to 0$

$$\sum_{i=1}^{[t/(k_n\Delta_n)]-1} \widehat{\gamma}_{g,(i-1)k_n\Delta_n}^{\intercal} (X_{(i+1)k_n\Delta_n} - X_{ik_n\Delta_n}) \xrightarrow{\text{u.c.p}} \int_0^t \gamma_{g,s-}^{\intercal} dX_s.$$

3.5 Realized Eigenvectors

So far we have constructed the principal components and estimated integrated eigenvalues. It remains to figure out the loading of each entry of X on each principal component, i.e., the eigenvector. As the eigenvector is stochastic, we estimate the integrated eigenvector. Apart from its sign, an eigenvector is uniquely identified if its associated eigenvalue is simple. We determine the sign hence identify the eigenvector, by requiring a priori certain entry of the eigenvector to be positive, e.g., the first non-zero entry.

Corollary 2. *Suppose Assumptions 1 and 4 hold. In addition, $\gamma_{g,s}$ is a vector-valued function that corresponds to the eigenvector of c_s with respect to a simple root $\lambda_{g,s}$, for each $s \in [0,t]$. Then we have as $\Delta_n \to 0$*

$$\frac{1}{\sqrt{\Delta_n}}\left(k_n\Delta_n\sum_{i=0}^{[t/(k_n\Delta_n)]}\left(\widehat{\gamma}_{g,ik_n\Delta_n}+\frac{1}{2k_n}\sum_{p\neq g}\frac{\widehat{\lambda}_{g,ik_n\Delta_n}\widehat{\lambda}_{p,ik_n\Delta_n}}{(\widehat{\lambda}_{g,ik_n\Delta_n}-\widehat{\lambda}_{p,ik_n\Delta_n})^2}\widehat{\gamma}_{g,ik_n\Delta_n}\right)-\int_0^t\gamma_{g,s}ds\right)\xrightarrow{\mathcal{L}-\mathrm{s}}\mathcal{W}_t^{\gamma},$$

where $\mathcal{W}^{\gamma}$ 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 given by

$$\mathrm{E}(\mathcal{W}_t^{\gamma}(\mathcal{W}_t^{\gamma})^{\intercal}|\mathcal{F})=\int_0^t\lambda_{g,s}(\lambda_{g,s}\mathbb{I}-c_s)^{+}c_s(\lambda_{g,s}\mathbb{I}-c_s)^{+}ds.$$

3.6 "PCA" on the Integrated Covariance Matrix

One may wonder why we chose to estimate the integrated eigenvalues of the spot covariance matrix rather than the eigenvalues of the integrated covariance matrix. Because eigenvalues are complicated functionals of the covariance matrix, the two quantities are not much related, and it turns out that the latter procedure is less informative than the former. For instance, the rank of the instantaneous covariance matrix is determined by the number of non-zero instantaneous eigenvalues. The eigenstructure of the integrated covariance matrix is rather opaque due to the aggregation of instantaneous covariances. Nevertheless, it is possible to construct a "PCA" procedure on the integrated covariance matrix, with the following results:

Proposition 4. *Suppose the eigenvalues of the integrated covariance matrix $\int_0^t c_s ds$ are given by*

$$\lambda_1=\ldots=\lambda_{g_1}>\lambda_{g_1+1}=\ldots=\lambda_{g_2}>\ldots\lambda_{g_{r-1}}>\lambda_{g_{r-1}+1}=\ldots=\lambda_{g_r}>0,\quad \textit{for } 1\leq r\leq d.$$

Then, we have, for $\varpi\in[1/(4-2\gamma),1/2)$,

$$\frac{1}{\sqrt{\Delta_n}}\left(F^{\lambda}\left(\sum_{i=0}^{[t/\Delta_n]}(\Delta_i^nX)(\Delta_i^nX)^{\intercal}1_{\{\|\Delta_i^nX\|\leq\alpha\Delta_n^{\varpi}\}}\right)-F^{\lambda}\left(\int_0^t c_sds\right)\right)\xrightarrow{\mathcal{L}-\mathrm{s}}\bar{\mathcal{W}}_t^{\lambda},$$

where $\bar{\mathcal{W}}^{\lambda}$ 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 given by

$$\mathrm{E}(\bar{\mathcal{W}}_{i,t}^{\lambda}\bar{\mathcal{W}}_{j,t}^{\lambda}|\mathcal{F})=\sum_{u,v,k,l=1}^{d}\partial_{uv}F_i^{\lambda}\left(\int_0^t c_sds\right)\left(\int_0^t(c_{uk,s}c_{vl,s}+c_{ul,s}c_{vk,s})ds\right)\partial_{kl}F_j^{\lambda}\left(\int_0^t c_sds\right),$$

$$\partial_{uv}F_i^{\lambda}(A)=\frac{1}{g_i-g_{i-1}}\sum_{k=g_{i-1}+1}^{g_i}O_{ku}O_{kv}.$$

where O is any orthogonal matrix that satisfies $A=O^{\intercal}Diag\left(\lambda(A)\right)O$.

Similarly, we have:

Proposition 5. *Suppose $\varpi\in[1/(4-2\gamma),1/2)$. For the eigenvector γ_g corresponding to some simple eigenvalue λ_g of $\int_0^t c_sds$,*

$$\frac{1}{\sqrt{\Delta_n}}\left(\gamma_g\left(\sum_{i=0}^{[t/\Delta_n]}(\Delta_i^nX)(\Delta_i^nX)^{\intercal}1_{\{\|\Delta_i^nX\|\leq\alpha\Delta_n^{\varpi}\}}\right)-\gamma_g\left(\int_0^t c_sds\right)\right)\xrightarrow{\mathcal{L}-\mathrm{s}}\bar{\mathcal{W}}_t^{\gamma},$$

where $\bar{\mathcal{W}}^{\gamma}$ 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 given by

$$\mathrm{E}(\bar{\mathcal{W}}^{\gamma}_{i,t}(\bar{\mathcal{W}}^{\gamma}_{j,t})^{\intercal}|\mathcal{F}) = \sum_{u,v,k,l=1}^{d} \partial_{uv}\gamma_{gi}\left(\int_0^t c_s ds\right)\left(\int_0^t (c_{uk,s}c_{vl,s} + c_{ul,s}c_{vk,s})ds\right)\partial_{kl}\gamma_{gj}\left(\int_0^t c_s ds\right),$$
$$\partial_{kl}\gamma_{gi}(A) = (\lambda_g(A)\mathbb{I} - A)^{+}_{ik}\,\gamma_{gl}(A).$$

The corresponding principal component is then defined as $\gamma_g^{\intercal}(X_t - X_0)$. However, this eigenvalue and the principal component do not satisfy the fundamental relationship between the eigenvalue and the variance of the principal component:

$$\lambda_g\left(\int_0^t c_s ds\right) \neq \left[\gamma_g^{\intercal}(X_t - X_0), \gamma_g^{\intercal}(X_t - X_0)\right]^c, \tag{20}$$

which is a desired property of any sensible PCA procedure. This fact alone makes this second procedure much less useful than the one we proposed above, based on integrated eigenvalues of the spot covariance. Another feature that distinguishes this second "PCA" procedure with both the classical one and our realized PCA is that the asymptotic covariance matrix of eigenvalues of $\int_0^t c_s ds$ is no longer diagonal.

An analogy of the relationship between the two PCA procedures is that between integrated quarticity $t\int_0^t \sigma_s^4 ds$ and the squared integrated volatility $(\int_0^t \sigma_s^2 ds)^2$. If the covariance matrix is constant, the two procedures are equivalent but not in general.

4 Simulation Evidence

We now investigate the small sample performance of the estimators in conditions that approximate the empirical setting of PCA for large stock portfolios. In order to generate data with a common factor structure, we simulate a factor model in which a vector of log-stock prices X follows the continuous-time dynamics

$$dX_t = \beta_t dF_t + dZ_t, \tag{21}$$

where F is a collection of unknown factors, β is the matrix of factor loadings, and Z is a vector of idiosyncratic components, which is orthogonal to F. This model is a special case of the general semimartingale model (2).

By construction, the continuous part of the quadratic variation has the following local structure:

$$[dX, dX]^c_t = \beta_t [dF, dF]^c_t \beta_t^{\intercal} + [dZ, dZ]^c_t.$$

When the idiosyncratic components have smaller magnitude, the dominating eigenvalues of X are close to the non-zero eigenvalues of $\beta_t[dF, dF]^c_t\beta_t^{\intercal}$, by Weyl's inequality. As a result, we should be able to detect the number of common factors in F from observations on X. That said, our goal here is to conduct nonparametric PCA rather than to estimate a parametric factor model, for which we need to resort to a large panel of X, whose dimension increases with the sample size, see, e.g., Bai and Ng (2002), and so (21) is only employed as the data-generating process on which PCA is employed without any knowledge of the underlying factor structure.

Specifically, we simulate

$$dX_{i,t} = \sum_{j=1}^{r} \beta_{ij,t} dF_{j,t} + dZ_{i,t}, \quad dF_{j,t} = \mu_j dt + \sigma_{j,t} dW_{j,t} + dJ^F_{j,t}, \quad dZ_{i,t} = \gamma_t dB_{i,t} + dJ^Z_{i,t},$$

where $i = 1, 2, \ldots, d$, and $j = 1, 2, \ldots, r$. In the simulations, one of the Fs plays the role of the market factor, so that its associated βs are positive. The correlation matrix of dW is denoted as ρ^F. We allow for time-varying $\sigma_{j,t}$, γ_t, and $\beta_{ij,t}$, which evolve according to the following system of equations:

$$d\sigma^2_{j,t} = \kappa_j(\theta_j - \sigma^2_{j,t})dt + \eta_j\sigma_{j,t}d\widetilde{W}_{j,t} + dJ^{\sigma^2}_{j,t}, \quad d\gamma^2_t = \kappa(\theta - \gamma^2_t)dt + \eta\gamma_t d\bar{B}_t,$$
$$d\beta_{ij,t} = \begin{cases} \tilde{\kappa}_j(\tilde{\theta}_{ij} - \beta_{ij,t})dt + \tilde{\xi}_j\sqrt{\beta_{ij,t}}d\widetilde{B}_{ij,t} & \text{if the } j\text{th factor is the ``market''}, \\ \tilde{\kappa}_j(\tilde{\theta}_{ij} - \beta_{ij,t})dt + \tilde{\xi}_j d\widetilde{B}_{ij,t} & \text{otherwise.} \end{cases},$$

where the correlation between $dW_{j,\cdot}$ and $d\widetilde{W}_{j,\cdot}$ is ρ_j, $\{J^F_j\}_{1\leq j\leq r}$ and $\{J^Z_i\}_{1\leq i\leq d}$ are driven by two Poisson Processes with arrival rates λ^F and λ^Z, respectively. Their jump sizes follow double exponential distributions with means denoted by μ^F_+, μ^F_-, μ^Z_+, and μ^Z_-, respectively. $\{J^{\sigma^2}_j\}_{1\leq j\leq r}$ co-jumps with J^F, and their jump sizes follow exponential distributions with the mean equal to μ^{σ^2}. All the model parameters are given in Table 1. They are chosen to be realistic given the empirical characteristics of the cross-section of stock returns and their volatilities.

Anticipating our empirical application to the S&P 100 Index constituents, we simulate intraday returns of $d = 100$ stocks at various frequencies and with horizons T spanning 1 week to 1 month. When $r = 3$, the model implies 3 distinct eigenvalues reflecting the local factor structure of the simulated data. The remaining population eigenvalues are identical, due to idiosyncratic variations. Throughout, we fix k_n to be the closest divisors of $[t/\Delta_n]$ to $\theta\Delta_n^{-1/2}\sqrt{\log(d)}$, with $\theta = 0.5$ and d is the dimension of X. Our choice of $\sqrt{\log(d)}$ is motivated from the literature on high-dimensional covariance matrix estimation, although our asymptotic design does not take into account an increasing dimensionality. Other choices of θ, e.g., 0.05 - 0.5, or functions of d, e.g., $\sqrt{d\log d}$, deliver the same results. To truncate off jumps from spot covariance estimates, we adopt the usual procedure in the literature, i.e., choosing $u_{i,n} = 3(\int_0^t c_{ii,s}ds/t)^{0.5}\Delta_n^{0.47}$, for $1 \leq i \leq d$, where $\int_0^t c_{ii,s}ds$ can be estimated by, for instance, bipower variations.

We then apply the realized PCA procedure. We first examine the curse of dimensionality – how increasing number of stocks affects the estimation – and how the sampling frequency affects the estimation. In light of Corollary 1, we estimate 3 simple integrated eigenvalues as well as the average of the remaining identical eigenvalues, denoted as $\int_0^t \lambda_{is}ds$, $i = 1, 2, 3$, and 4. We report the mean and standard errors of the estimates as well as the root-mean-square errors of the standardized estimates with $d = 5$, 10, 15, 20, 30, 50, and 100 stocks, respectively, using returns sampled every $\Delta_n = 5$ seconds, 1 minute and 5 minutes over one week and one month horizons. The results, as shown from Tables 2 - 5, suggest that, as expected, the estimation is more difficult as the dimensionality increases, but the large amount of high frequency data and in-fill asymptotic techniques deliver very satisfactory finite sample approximations. The eigenvalues are accurately recovered. The repeated eigenvalues are estimated with smaller biases and standard errors, due to the extra averaging taken at the estimation stage. In Tables 6 - 7, we provide estimates for the first eigenvectors. Similar to the estimation for eigenvalues, the estimates are very accurate, even with 100 stocks.

To further verify the accuracy of the asymptotic distribution in small samples as the sample size increases, we provide in Figure 1 histograms of the standard estimates of integrated eigenvalues using 30 stocks with 5-second returns. Finally, we examine the finite sample accuracy of the integrated simple eigenvalues, in the more challenging scenario with 100 stocks sampled every minute, which matches the setup of the empirical analysis we will conduct below. To verify that the detection of 3 eigenvalues is not an artifact, we vary the number of common factors in the data generating process from $r = 2$ to 4, by removing the third factor or adding another factor that shares the same parameters as the third factor. The histograms are provided in Figure 2. The results collectively show that the finite sample performance of the method is quite good even for a 100-stock portfolio with one-minute returns and one-week horizon.

5 High-Frequency Principal Components in the S&P 100 Stocks

Given the encouraging Monte Carlo results, we now conduct PCA on intraday returns of S&P 100 Index (OEX) constituents. We collect stock prices over the 2003 - 2012 period from the Trade and Quote (TAQ) database of the New York Stock Exchange (NYSE).[4] Due to entry and exit from the index, there are in total 154 tickers over this 10-year period. We conduct PCA on a weekly basis. One of the key advantages of the large amount of high frequency data is that we are effectively able to create a ten-year long time series of eigenvalues and principal components at the weekly frequency by collating the results obtained over each week in the sample.

After removing those weeks during which the index constituents are switching, we are left with 482 weeks. To clean the data, for each ticker on each trading day, we only keep the intraday prices from the single exchange which has the largest number of transaction records. We provide in Figure 3 the quantiles of the number of transactions between 9:35 a.m. EST and 3:55 p.m. EST, across 100 stocks each day. These stocks have excellent liquidity over the sampling period. We thereby employ 1-minute subsamples for the most liquid 90 stocks in the index (for simplicity, we still refer to this 90-stock subsample as the S&P 100) in order to address any potential concerns regarding microstructure noise and asynchronous trading.[5] These stocks trade multiple times per sampling interval and we compute returns using the latest prices recorded within each sampling interval. Overnight returns are removed so that there is no concern of price changes due to dividend distributions or stock splits. The 158 time series of cumulative returns are plotted in Figure 4.

5.1 Scree Plot

We report the time series average of the eigenvalue estimates against their order in Figure 5, a plot known in classical PCA analysis as the "scree plot" (see e.g. Jolliffe (2002)). This plot classically constitutes the main graphical aid employed to determine the appropriate number of components in empirical applications. The graph reveals that there are a handful of eigenvalue estimates, three, which are separated from the rest. The fact that three factors on average explain the cross-sectional variation of stock returns is surprisingly consistent with the well-established low frequency Fama-French common factor analysis, despite the large differences in methods, time periods, sample frequencies and length of observation.

We also plot the estimates of the percentage variation explained by the first three integrated eigenvalues in Figure 6, along with the average variation explained by the remaining eigenvalues. There are substantial time variation of the first three eigenvalues, all of which are at peak around the recent financial crisis, indicating an increased level of comovement, which in this context is the definition of systemic risk. The idiosyncratic factors become relatively less important and even more dominated by the common factors during the crisis.

The first eigenvalue accounts for on average 30-40% of the total variations of the 90 constituents, capturing the extent of the market variation in the sample. The second and third components together account for an additional 15%-20% of the total variation. Therefore, there exists significant amount of remaining idiosyncratic variation, beyond what can be explained by a three-common-factor model. The average variation explained by the remaining 87 components are around 0.4%-0.6%, which corresponds to the idiosyncratic contributions relative to the first three principal components.

[4] While the constituents of the OEX Index change over time, we keep track of the changes to ensure that our choice of stocks is always in line with the index constituents.

[5] Estimators that are robust to both noise and asynchronicity are available for integrated covariance estimation in Aït-Sahalia, Fan, and Xiu (2010), Christensen, Kinnebrock, and Podolskij (2010), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011), and Shephard and Xiu (2012). Extending the current method to a noisy and asynchronous setting is theoretically interesting but not empirically necessary for the present paper; it is left for future work.

Next, we compare the cumulative returns of the first three cumulative principal components. The time series plots are shown in Figure 7. The empirical correlations among the three components are -0.020, 0.005, and -0.074, respectively, which agrees with the design that the components are orthogonal. The first principal component accounts for most of the common variation, and it shares the time series features of the overall market return. This is further reinforced by the fact that the loadings of all the stocks on the first principal component, although time-varying, remain positive throughout the sample period, which is not the case for the additional principal components. It is worth pointing out however that these principal components are not constrained to be portfolio returns, as their weights at each point in time are nowhere constrained to add up to one. This means that the first principal component cannot be taken directly to be the market portfolio, or more generally identified with additional Fama-French or additional mimicking portfolios.

5.2 Biplots

Although PCA is by design not suited to identifying the underlying economic factors corresponding to the principal components, the time series evidence in Figure 7 suggests that the first principal component captures the features of the overall market return, as already noted. Can we extract more information about the economic nature of the principal components? For this purpose, we can use our results to produce biplots (see Gabriel (1971)) and see which assets line up together in the basis of the principal components. A biplot gives the relative position of the d variables on a plot where two principal components are the axes. The length of a vector from the origin pointing towards the position of a variable on the biplot tends to reflect the magnitude of the variation of the variable. Moreover, vectors associated with similar variables tend to point towards the same direction, hence a biplot can be used for classification of the assets. While it is possible to compute biplots for the spot eigenvectors, the biplots for the integrated eigenvectors are more stable.

We report two representative snapshots of the biplots of the first two integrated eigenvectors in Figures 8 (March 7-11, 2005, preceding the crisis) and 9 (March 10-14, 2008, during the crisis, although not at its most extreme), in order to interpret the second principal component. We identify differently in the figures the vectors that correspond to financial stocks (identified using the Global Industrial Classification Standard (GICS) codes). The Federal Reserve made several important announcement during the week of March 10 -14, 2008 to address heightened liquidity pressures of the financial system, including the creation of the Term Securities Lending Facility, and the approval of a financing arrangement between J.P. Morgan Chase and Bear Stearns.

We find that the financial sector is clearly separated from the rest of the stocks during March 10-14, 2008, in sharp contrast with the biplot for the week of March 7 - 11, 2005, during which no such separation occurs. This suggests that PCA based on low-frequency data is likely to overlook these patterns. Interestingly, we can also see from the March 2008 biplot that the Lehman Brothers (LEH) stands out, having the largest loadings on both components. The corresponding scree plots for these two weeks in Figures 10 and 11 both suggest the presence of at least three factors, but the eigenvalues in March 10-14, 2008 are much larger in magnitude, consistently with a higher systematic component to asset returns. For each week, the figures report 95% confidence intervals for the first three eigenvalues computed based on the distribution given in Corollary 1.

6 Conclusions

This paper develops the tools necessary to implement PCA at high frequency, constructing and estimating realized eigenvalues, eigenvectors and principal components. This development complements the classical PCA theory in a number of ways. Compared to its low frequency counterpart, PCA becomes feasible over

short windows of observation (of the order of one week), relatively large dimensions (90 in our application) and further are free from the need to impose strong parametric assumptions on the distribution of the data, applying instead to a broad class of semimartingales. The estimators perform well in simulations and reveal that the joint dynamics of the S&P 100 stocks at high frequency are well explained by a three-factor model, a result that is broadly consistent with the Fama-French factor model at low frequency, surprisingly so given the large differences in time scale, sampling and returns horizon.

This paper represents a necessary first step to bring PCA tools to a high frequency setting. Although not empirically relevant in the context of the analysis above of a portfolio of highly liquid stocks at the one-minute frequency, the next steps in the development of the theory will likely include the incorporation of noise-robust and asynchronicity-robust covariance estimation methods. We hope to pursue these extensions in future work.

References

Aït-Sahalia, Y., J. Fan, and D. Xiu (2010): "High-Frequency Covariance Estimates with Noisy and Asynchronous Data," *Journal of the American Statistical Association*, 105, 1504–1517.

Aït-Sahalia, Y., and J. Jacod (2014): *High Frequency Financial Econometrics*. Princeton University Press.

Amini, A. A., and M. J. Wainwright (2009): "High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components," *Annals of Statistics*, 37(5B), 2877–2921.

Anderson, T. W. (1958): *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

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

Bai, J. (2003): "Inferential Theory for Factor models of Large Dimensions," *Econometrica*, 71, 135–171.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

Bai, Z., J. W. Silverstein, and Y. Q. Yin (1988): "A Note on the Largest Eigenvalue of a Large Dimensional Covariance Matrix," *Journal of Multivariate Analysis*, 26, 166–168.

Baker, M., and J. Wurgler (2006): "Investor Sentiment and the Cross-Section of Stock Returns," *The Journal of Finance*, 61(4), 1645–1680.

Baker, S. R., N. Bloom, and S. J. Davis (2013): "Measuring Economic Policy Uncertainty," Discussion paper, Stanford University and University of Chicago.

Ball, J. M. (1984): "Differentiability Properties of Symmetric and Isotropic Functions," *Duke Mathematical Journal*, 51, 699–728.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011): "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading," *Journal of Econometrics*, 162, 149–169.

Bickel, P. J., and E. Levina (2008a): "Covariance Regularization by Thresholding," *Annals of Statistics*, 36(6), 2577–2604.

——— (2008b): "Regularized Estimation of Large Covariance Matrices," *Annals of Statistics*, 36, 199–227.

Brillinger, D. R. (2001): *Time Series: Data Analysis and Theory*, Classics in Applied Mathematics (Book 36). SIAM: Society for Industrial and Applied Mathematics.

Cai, T. T., and H. H. Zhou (2012): "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," *Annals of Statistics*, 40(5), 2389–2420.

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

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

Connor, G., and R. Korajczyk (1988): "Risk and Return in an Equilibrium APT: Application of a New Test Methodology," *Journal of Financial Economics*, 21, 255–289.

d'Aspremont, A., L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet (2007): "A Direct Formulation for Sparse PCA Using Semidefinite Programming," *SIAM Review*, 49(3), 434–448.

Davis, C. (1957): "All Convex Invariant Functions of Hermitian Matrices," *Archiv der Mathematik*, 8(4), 276–278.

Egloff, D., M. Leippold, and L. Wu (2010): "The term structure of variance swap rates and optimal variance swap investments," *Journal of Financial and Quantitative Analysis*, 45, 1279–1310.

Fama, E. F., and K. R. French (1993): "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3–56.

Fan, J., A. Furger, and D. Xiu (2015): "Incorporating Global Industrial Classification Standard into Portfolio Allocation: A Simple Factor-Based Large Covariance Matrix Estimator with High Frequency Data," *Journal of Business and Economic Statistics, forthcoming*.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The Generalized Dynamic-Factor Model: Identification and Estimation," *The Review of Economics and Statistics*, 82, 540–554.

——— (2004): "The Generalized Dynamic Factor Model: Consistency and Rates," *Journal of Econometrics*, 119(2), 231–255.

Forni, M., and M. Lippi (2001): "The Generalized Dynamic Factor Model: Representation Theory," *Econometric Theory*, 17, 1113–1141.

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

Gabriel, K. (1971): "The biplot graphic display of matrices with application to pricipal component analysis," *Biometrika*, 58, 453–467.

Geman, S. (1980): "A Limit Theorem for the Norm of Random Matrices," *Annals of Probability*, 8, 252–261.

Heinrich, C., and M. Podolskij (2014): "On Spectral Distribution of High Dimensional Covaraition Matrices," Discussion paper, University of Aarhus.

Horn, R. A., and C. R. Johnson (2013): *Matrix Analysis*. Cambridge University Press, second edn.

Hotelling, H. (1933): "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24, 417–441, 498–520.

Jackson, J. E. (2003): *A User's Guide to Principal Components*. Wiley.

Jacod, J., A. Lejay, and D. Talay (2008): "Estimation of the Brownian dimension of a continuous Itô process," *Bernoulli*, 14, 469–498.

Jacod, J., and M. Podolskij (2013): "A test for the rank of the volatility process: The Random Perturbation Approach," *Annals of Statistics*, 41, 2391–2427.

Jacod, J., and P. Protter (2011): *Discretization of Processes*. Springer-Verlag.

Jacod, J., and M. Rosenbaum (2013): "Quarticity and Other Functionals of Volatility: Efficient Estimation," *Annals of Statistics*, 41, 1462–1484.

Jacod, J., and A. N. Shiryaev (2003): *Limit Theorems for Stochastic Processes*. Springer-Verlag, second edn.

Johnstone, I. M. (2001): "On the distribution of the largest eigenvalue in principal components analysis," *Annals of Statistics*, 29, 295–327.

Johnstone, I. M., and A. Y. Lu (2009): "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104(486), 682–693.

Jolliffe, I. T. (2002): *Principal Component Analysis*. Springer-Verlag.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003): "A Modified Principal Component Technique Based on the LASSO," *Journal of Computational and Graphical Statistics*, 12(3), 531–547.

Kalnina, I., and D. Xiu (2013): "Model-Free Leverage Effect Estimators at High Frequency," Discussion paper, Université de Montréal and University of Chicago.

Lewis, A. S. (1996a): "Convex Analysis on the Hermitian Matrices," *SIAM Journal of Optimizaiton*, 6(1), 164–177.

——— (1996b): "Derivatives of Spectral Functions," *Mathematics of Operations Research*, 21, 576–588.

Lewis, A. S., and H. S. Sendov (2001): "Twice Differentiable Spectral Functions," *SIAM Journal on Matrix Analysis and Applications*, 23, 368–386.

Li, J., V. Todorov, and G. Tauchen (2013): "Inference Theory on Volatility Functional Dependencies," Discussion paper, Duke University.

——— (2014): "Adaptive Estimation of Continuous-Time Regression Models using High-Frequency Data," Discussion paper, Duke University.

Li, J., and D. Xiu (2014): "Generalized Method of Integrated Moments for High-Frequency Data," Discussion paper, Duke University and The University of Chicago.

Litterman, R., and J. Scheinkman (1991): "Common factors affecting bond returns," *Journal of Fixed Income*, June, 54–61.

Magnus, J. R., and H. Neudecker (1999): *Matrix Differential Calculus with Applications in Statistics and Economics*. Wiley.

Mykland, P. A., and L. Zhang (2009): "Inference for continuous semimartingales observed at high frequency," *Econometrica*, 77, 1403–1445.

Okamoto, M. (1973): "Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample," *Annals of Statistics*, 1, 763–765.

Pearson, K. (1901): "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 2, 559–572.

Protter, P. (2004): *Stochastic Integration and Differential Equations: A New Approach*. Springer-Verlag, second edn.

Rockafellar, R. T. (1997): *Convex Analysis*. Princeton University Press.

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

Shephard, N., and D. Xiu (2012): "Econometric analysis of multivariate realized QML: Estimation of the covariation of equity prices under asynchronous trading," Discussion paper, University of Oxford and University of Chicago.

Silhavý, M. (2000): "Differentiability Properties of Isotropic Functions," *Duke Mathematical Journal*, 104, 367–373.

Stock, J. H., and M. W. Watson (1998): "Diffusion Indexes," Discussion paper, NBER.

——— (1999): "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293–335.

——— (2002): "Forecasting using Principal Components from a Large Number of Predictors," *Journal of American Statistical Association*, 97, 1167–1179.

Sylvester, J. (1985): "On the Differentiablity of O(n) Invaraint Functions of Symmetric Matrices," *Duke Mathematical Journal*, 52.

Tao, M., Y. Wang, and X. Chen (2013): "Fast Convergence Rates in Estimating Large Volatility Matrices Using High-Frequency Financial Data," *Econometric Theory*, 29(4), 838–856.

Tao, M., Y. Wang, Q. Yao, and J. Zou (2011): "Large Volatility Matrix Inference via Combining Low-Frequency and High-Frequency Approaches," *Journal of the American Statistical Association*, 106, 1025–1040.

Tao, M., Y. Wang, and H. H. Zhou (2013): "Optimal Sparse Volatility Matrix Estimation for High-Dimensional Itô Processes with Measurement Errors," *Annals of Statistics*, 41, 1816–1864.

Tao, T. (2012): *Topics in Random Matrix Theory*. American Mathematical Society.

Tyler, D. E. (1981): "Asymptotic Inference for Eigenvectors," *Annals of Statistics*, 9, 725–736.

Wang, Y., and J. Zou (2010): "Vast volatility matrix estimation for high-frequency financial data," *Annals of Statistics*, 38, 943–978.

Waternaux, C. M. (1976): "Asymptotic Distribution of the Sample Roots for a Nonnormal Population," *Biometrika*, 63, 639–645.

Zheng, X., and Y. Li (2011): "On the Estimation of Integrated Covariance Matrices of High Dimensional Diffusion Processes," *Annals of Statistics*, 39, 3121–3151.

Zou, H., T. Hastie, and R. Tibshirani (2006): "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, 15(2), 265–286.

Appendix A Mathematical Proofs

Appendix A.1 Proof of Lemma 1

Proof. By the classical Weyl inequalities in e.g., Horn and Johnson (2013) and Tao (2012), we have $|\lambda_j(A+\epsilon)-\lambda_j(A)| \leq K\|\epsilon\|$, for all $j=1,2,\ldots,d$, where $A,\epsilon \in \mathcal{M}_d^+$, and K is some constant. This establishes the Lipchitz property. ∎

Appendix A.2 Proof of Lemma 2

Proof. It is straightforward to show (by the implicit function theorem, see Magnus and Neudecker (1999) Theorem 8.7) that any simple eigenvalue and its corresponding eigenvector, written as functions of A, $\lambda_g(A)$ and $\gamma_g(A)$, are C^∞. To calculate their derivatives, note that $A\gamma_g = \lambda_g\gamma_g$, hence we have $(\partial_{jk}A)\gamma_g + A(\partial_{jk}\gamma_g) = (\partial_{jk}\lambda_g)\gamma_g + \lambda_g(\partial_{jk}\gamma_g)$. Pre-multiplying $\gamma_g^\intercal$ on both sides yields $\partial_{jk}\lambda_g = \gamma_g^\intercal(\partial_{jk}A)\gamma_g = \gamma_{gj}\gamma_{gk}$. Rewrite it into $(\lambda_g\mathbb{I}-A)\partial_{jk}\gamma_g = (\partial_{jk}A)\gamma_g - (\partial_{jk}\lambda_g)\gamma_g$, which leads to $(\lambda_g\mathbb{I}-A)^+(\lambda_g\mathbb{I}-A)\partial_{jk}\gamma_g = (\lambda_g\mathbb{I}-A)^+(\partial_{jk}A)\gamma_g$. As a result, $\partial_{jk}\gamma_g = (\lambda_g\mathbb{I}-A)^+ J_{jk}\gamma_g$.

In the case when all eigenvalues are simple, by direct calculation we have

$$\partial_{jk}\gamma_{gh} = \sum_{p\neq g}\frac{1}{\lambda_g-\lambda_p}\gamma_{ph}\gamma_{pj}\gamma_{gk},$$

where we use the fact that $(\gamma_1^\intercal,\gamma_2^\intercal,\ldots,\gamma_d^\intercal)^\intercal A(\gamma_1,\gamma_2,\ldots,\gamma_d) = \mathrm{Diag}(\lambda(A))$. Further,

$$\begin{aligned}
\partial^2_{jk,lm}\gamma_{gh} =& -\sum_{p\neq g}\frac{1}{(\lambda_g-\lambda_p)^2}(\partial_{lm}\lambda_g-\partial_{lm}\lambda_p)\gamma_{ph}\gamma_{pj}\gamma_{gk} + \sum_{p\neq g}\frac{1}{\lambda_g-\lambda_p}\partial_{lm}\left(\gamma_{ph}\gamma_{pj}\gamma_{gk}\right)\\
=& -\sum_{p\neq g}\frac{1}{(\lambda_g-\lambda_p)^2}\left(\gamma_{gl}\gamma_{gm}\gamma_{ph}\gamma_{pj}\gamma_{gk}-\gamma_{pl}\gamma_{pm}\gamma_{ph}\gamma_{pj}\gamma_{gk}\right)\\
&+\sum_{p\neq g}\sum_{q\neq p}\frac{1}{(\lambda_g-\lambda_p)(\lambda_p-\lambda_q)}\gamma_{ql}\gamma_{pm}\gamma_{qh}\gamma_{pj}\gamma_{gk}\\
&+\sum_{p\neq g}\sum_{q\neq p}\frac{1}{(\lambda_g-\lambda_p)(\lambda_p-\lambda_q)}\gamma_{ql}\gamma_{pm}\gamma_{qj}\gamma_{ph}\gamma_{gk}\\
&+\sum_{p\neq g}\sum_{q\neq g}\frac{1}{(\lambda_g-\lambda_p)(\lambda_g-\lambda_q)}\gamma_{qk}\gamma_{ql}\gamma_{gm}\gamma_{ph}\gamma_{pj},
\end{aligned}$$

which concludes the proof. ∎

Appendix A.3 Proof of Lemma 3

Proof. The proof is by induction. Consider, at first the following optimization problem:

$$\max_{\gamma_s}\int_0^u \gamma_s^\intercal c_s\gamma_s ds,\ \text{s.t.}\quad \gamma_s^\intercal\gamma_s = 1,\quad 0\leq s\leq u\leq t.$$

Using a sequence of Lagrange multipliers λ_s, the problem can be written as solving

$$c_s\gamma_s = \lambda_s\gamma_s,\quad \text{and}\quad \gamma_s^\intercal\gamma_s = 1,\ \text{for any } 0\leq s\leq t.$$

Hence, the original problem is translated into eigenanalysis.

Suppose the eigenvalues of c_s are ordered as in $\lambda_{1,s} \geq \lambda_{2,s} \geq \ldots \geq \lambda_{d,s}$. Note that $\gamma_s^\intercal c_s \gamma_s = \lambda_s$, so that $\lambda_s = \lambda_{1,s}$, and $\gamma_s = \gamma_{1,s}$ is one of the corresponding eigenvectors (if $\lambda_{1,s}$ is not unique), and the maximal variation is $\int_0^t \lambda_{1,s} ds$.

Suppose that we have found $\gamma_{1,s}, \ldots, \gamma_{k,s}$, for $1 \leq k < d$ and $0 \leq s \leq t$, the $(k+1)$th principal component is defined by solving the following problem:

$$\max_{\gamma_s} \int_0^u \gamma_s^\intercal c_s \gamma_s ds, \quad \text{s.t.} \quad \gamma_s^\intercal \gamma_s = 1, \quad \text{and} \quad \gamma_{j,s}^\intercal c_s \gamma_s = 0, \text{ for } 1 \leq j \leq k, \quad 0 \leq u \leq t.$$

Using similar technique of Lagrange multipliers, λ_s, and $\nu_{1,s}, \ldots, \nu_{k,s}$, we find

$$c_s \gamma_s = \lambda_s \gamma_s + \sum_{j=1}^{k} \nu_{j,s} c_s \gamma_{j,s}.$$

Multiplying on the left $\gamma_{l,s}^\intercal$, for some $1 \leq l \leq k$, we can show that $\nu_{l,s} c_s \gamma_{l,s} = 0$. Indeed,

$$0 = \lambda_{l,s} \gamma_{l,s}^\intercal \gamma_s = \gamma_{l,s}^\intercal c_s \gamma_s = \gamma_{l,s}^\intercal \lambda_s \gamma_s + \sum_{j=1}^{k} \nu_{j,s} \gamma_{l,s}^\intercal c_s \gamma_{j,s} = \nu_{l,s} \lambda_{l,s}.$$

Therefore, since l is an arbitrary number between 1 and k, we have $c_s \gamma_s = \lambda_s \gamma_s$. Hence, $\lambda_s = \lambda_{k+1,s}$, $\gamma_s = \gamma_{k+1,s}$ is one of the eigenvectors associated with the eigenvalue $\lambda_{k+1,s}$. This establishes the first part of the theorem.

For any càdlàg and adapted process γ_s,

$$\left[\int_0^u \gamma_{s-}^\intercal dX_s, \int_0^u \gamma_{s-}^\intercal dX_s \right]^c = \int_0^t \gamma_s^\intercal c_s \gamma_s ds.$$

Hence the statement follows from the gth-step optimization problem. Note that the validity of the integrals above is warranted by the continuity of λ given by Lemma 1. ∎

Appendix A.4 Proof of Lemma 4

Proof. The first statement of the proof follows by immediate calculations from Theorem 1.1 in Lewis (1996b) and Theorem 3.3 in Lewis and Sendov (2001). The second statement is discussed and proved in, e.g., Ball (1984), Sylvester (1985), and Silhavý (2000). Finally, the last statement on convexity is proved in Davis (1957) and Lewis (1996a). ∎

Appendix A.5 Proof of Lemma 5

Proof. Obviously, for any $1 \leq g_1 < g_2 < \ldots < g_r \leq d$, the set defined in (4), $\mathcal{D}(g_1, g_2, \ldots, g_r)$, is an open set in $\mathbb{R}_d^+ / \{0\}$. Define $f(x) = |\bar{x}_{g_r}| + \sum_{i \neq j} |\bar{x}_{g_i} - \bar{x}_{g_j}|$, which is a continuous and convex function. It is differentiable at x if and only if $x \in \mathcal{D}(g_1, g_2, \ldots, g_r)$. Therefore, by Lemma 4, $f \circ \lambda$ is convex, and it is differentiable at A if and only if $\lambda(A) \in \mathcal{D}(g_1, g_2, \ldots, g_r)$, i.e., $A \in \mathcal{M}(g_1, g_2, \ldots, g_r)$. On the other hand, a convex function is almost everywhere differentiable, see Rockafellar (1997), which implies that $\mathcal{M}(g_1, g_2, \ldots, g_r)$ is dense in $\mathcal{M}_d^{++}$. Moreover, $\mathcal{M}(g_1, g_2, \ldots, g_r)$ is the pre-image of the open set $\mathbb{R}^+ / \{0\}$ under a continuous function $h \circ \lambda$, where $h(x) = \prod_{i \neq j} |\bar{x}_{g_i} - \bar{x}_{g_j}| |\bar{x}_{g_r}|$. Therefore, it is open. ∎

Appendix A.6 Proof of Theorem 1

Proof. Note that

$$V(\Delta_n, X; F) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} f\left(\widehat{\lambda}_{ik_n\Delta_n}\right) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} (f\circ\lambda)(\widehat{c}_{ik_n\Delta_n}).$$

By Assumption 2 and Lemma 4, $f\circ\lambda$ is a continuous vector-valued function. Moreover, for $c\in\mathcal{M}_d^+$, $\|f\circ\lambda(c)\|\le K(1+\|\lambda(c)\|^\zeta)\le K(1+\|c\|^\zeta)$. Below we prove this theorem for any spectral function F that is bounded by $K(1+\|c\|^\zeta)$.

We start with a function F bounded by K everywhere. We extend the definition of $\widehat{c}$ to the entire interval $[0,t]$ by letting:

$$\widehat{c}_s = \widehat{c}_{(i-1)k_n\Delta_n}, \text{ for } \quad (i-1)k_n\Delta_n \le s < ik_n\Delta_n.$$

Note that for any $t>0$, we have

$$\begin{aligned}
&\mathrm{E}\left\|V(\Delta_n, X; F) - \int_0^t F(c_s)ds\right\| \\
\le& k_n\Delta_n \mathrm{E}\left\|F(\widehat{c}_{[t/(k_n\Delta_n)]k_n\Delta_n})\right\| + \int_0^{[t/(k_n\Delta_n)]k_n\Delta_n} \mathrm{E}\left\|F(\widehat{c}_s) - F(c_s)\right\| ds + \int_{[t/(k_n\Delta_n)]k_n\Delta_n}^t \mathrm{E}\left\|F(c_s)\right\| ds \\
\le& Kk_n\Delta_n + \int_0^{[t/(k_n\Delta_n)]k_n\Delta_n} \mathrm{E}\left\|F(\widehat{c}_s) - F(c_s)\right\| ds.
\end{aligned}$$

By the fact that $\widehat{c}_s - c_s \xrightarrow{\mathrm{p}} 0$, it follows that $\mathrm{E}\left\|F(\widehat{c}_s) - F(c_s)\right\|\to 0$, which is bounded uniformly in s and n because F is bounded. Therefore, by the dominated convergence theorem, we obtain the desired convergence.

Next we show the convergence holds under the polynomial bound on F. Denote ψ to be a C^∞ function on $\mathbb{R}^+$ such that $1_{[1,\infty)}(x)\le\psi(x)\le 1_{[1/2,\infty]}(x)$. Let $\psi_\varepsilon(c) = \psi(\|c\|/\varepsilon)$, and $\psi'_\varepsilon(c) = 1-\psi_\varepsilon(c)$. Since the function $F\cdot\psi'_\varepsilon$ is continuous and bounded, the above argument implies that $V(\Delta_n, X; F\cdot\psi'_\varepsilon)\xrightarrow{\mathrm{p}}\int_0^t F\cdot\psi'_\varepsilon(c_s)ds$, for any fixed ε. When ϵ is large enough, we have $\int_0^t F\cdot\psi'_\varepsilon(c_s)ds = \int_0^t F(c_s)ds$ by localization, since c_s is locally bounded. On the other hand, $F\cdot\psi_\varepsilon(c)\le K\|c\|^\zeta 1_{\{\|c\|\ge\varepsilon\}}$, for $\varepsilon>1$. So it remains to show that

$$\lim_{\varepsilon\to\infty}\limsup_{n\to\infty}\mathrm{E}\left(k_n\Delta_n\sum_{i=0}^{[t/(k_n\Delta_n)]}\|\widehat{c}_{ik_n\Delta_n}\|^\zeta 1_{\{\|\widehat{c}_{ik_n\Delta_n}\|>\varepsilon\}}\right) = 0.$$

By (9.4.7) of Jacod and Protter (2011), there exists some sequence a_n going to 0, such that

$$\mathrm{E}\left(\|\widehat{c}_{ik_n\Delta_n}\|^\zeta 1_{\{\|\widehat{c}_{ik_n\Delta_n}\|>\varepsilon\}}|\mathcal{F}_{ik_n\Delta_n}\right)\le\frac{K}{\varepsilon^\zeta} + Ka_n\Delta_n^{(1-\zeta+\varpi(2\zeta-\gamma))},$$

which establishes the desired result. ∎

Appendix A.7 Proof of Proposition 1

Proof. We divide the proof into several steps. To start, we need some additional notations. Let X' and c' denote the continuous parts of the processes X and c, respectively. Also, we introduce $\widehat{c}'_{ik_n\Delta_n}$ to denote the estimator constructed similarly as in (5) with X replaced by X' and without truncation, namely

$$\widehat{c}'_{ik_n\Delta_n} = \frac{1}{k_n\Delta_n}\sum_{j=1}^{k_n}(\Delta^n_{ik_n+j}X')(\Delta^n_{ik_n+j}X')^\intercal.$$

In addition, $\widehat{\lambda}'_{ik_n\Delta_n}$ corresponds to the vector of eigenvalues of $\widehat{c}'_{ik_n\Delta_n}$, and

$$V'(\Delta_n, X; F) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} f\left(\widehat{\lambda}'_{ik_n\Delta_n}\right).$$

We also define

$$\bar{c}_{ik_n\Delta_n} = \frac{1}{k_n\Delta_n}\int_{ik_n\Delta_n}^{(i+1)k_n\Delta_n} c_s ds, \quad \beta^n_{ik_n} = \widehat{c}'_{ik_n\Delta_n} - c_{ik_n\Delta_n}, \quad \alpha^n_l = (\Delta^n_l X)(\Delta^n_l X)^\intercal - c_{l\Delta_n}\Delta_n, \quad \text{and}$$
$$\eta^n_i = \left(\mathrm{E}\left(\sup_{i\Delta_n \le u \le i\Delta_n + k_n\Delta_n} |b_{i\Delta_n+u} - b_{i\Delta_n}|^2 | \mathcal{F}_{i\Delta_n}\right)\right)^{1/2}.$$

We first collect some known estimates in the next lemma:

Lemma 6. *Under the assumptions of Proposition 1, we have*

$$\mathrm{E}\left(\sup_{0\le u\le s} \|c_{t+u} - c_t\|^q | \mathcal{F}_t\right) \le Ks^{1\wedge q/2}, \quad \|\mathrm{E}\left(c_{t+s} - c_t | \mathcal{F}_t\right)\| \le Ks, \tag{A.1}$$

$$\mathbb{E}\left\|(\Delta^n_i X)(\Delta^n_i X)^\intercal 1_{\{\|\Delta^n_i X\| \le u_n\}} - (\Delta^n_i X')(\Delta^n_i X')^\intercal\right\| \le Ka_n\Delta_n^{(2-r)\varpi+1}, \text{ for some } a_n \to 0, \tag{A.2}$$

$$\mathrm{E}\left(\left\|\widehat{c}_{ik_n\Delta_n} - \widehat{c}'_{ik_n\Delta_n}\right\|^q\right) \le Ka_n\Delta_n^{(2q-r)\varpi+1-q}, \text{ for some } q \ge 1, \text{ and } a_n \to 0, \tag{A.3}$$

$$\mathrm{E}\left\|\widehat{c}'_{ik_n\Delta_n} - \bar{c}_{ik_n\Delta_n}\right\|^p \le Kk_n^{-p/2}, \text{ for some } p \ge 1, \tag{A.4}$$

$$\mathrm{E}\left(\|\alpha^n_i\|^q | \mathcal{F}_{i\Delta_n}\right) \le K\Delta^q_n, \text{ for some } q \ge 0, \tag{A.5}$$

$$\|\mathrm{E}\left(\alpha^n_i | \mathcal{F}_{i\Delta_n}\right)\| \le K\Delta_n^{3/2}\left(\Delta_n^{1/2} + \eta^n_i\right), \tag{A.6}$$

$$\left|\mathrm{E}\left(\alpha^{n,jk}_i \alpha^{n,lm}_i - \left(c^{jl}_{i\Delta_n} c^{km}_{i\Delta_n} + c^{jm}_{i\Delta_n} c^{kl}_{i\Delta_n}\right)\Delta^2_n | \mathcal{F}_{i\Delta_n}\right)\right| \le K\Delta_n^{5/2}, \tag{A.7}$$

$$\left\|\mathrm{E}\left(\beta^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n}\right)\right\| \le K{\Delta_n}^{1/2}\left(k_n\Delta_n^{1/2} + \eta^n_i\right), \tag{A.8}$$

$$\mathrm{E}\left(\left\|\beta^n_{ik_n}\right\|^q | \mathcal{F}_{ik_n\Delta_n}\right) \le K\left(k_n^{-q/2} + k_n\Delta_n\right), \text{ for some } q \ge 2, \tag{A.9}$$

$$\Delta_n \mathrm{E}\sum_{i=1}^{[t/\Delta_n]} \eta^n_i \to 0. \tag{A.10}$$

Proof of Lemma 6. These estimates are given by Lemma A.2 in Li and Xiu (2014), (4.3), (4.8), (4.10), (4.11), (4.12), (4.18), Lemmas 4.2 and 4.3 of Jacod and Rosenbaum (2013), and Lemma 13.2.6 of Jacod and Protter (2011). ∎

Now we return to the proof of Proposition 1.

1) We show that we can restrict the domain of function f to some compact set, where both the estimates $\{\widehat{c}_{ik_n\Delta_n}\}_{i=0,1,2,\ldots,[t/(k_n\Delta_n)]}$ and the sample path of $\{c_s\}_{s\in[0,t]}$ take values. By (A.4), we have for $p \ge 1$,

$$\mathbb{E}\left\|\widehat{c}'_{ik_n\Delta_n} - \bar{c}_{ik_n\Delta_n}\right\|^p \le Kk_n^{-p/2}.$$

Therefore, by the maximal inequality, we deduce, by picking $p > 2/\varsigma - 2$,

$$\mathbb{E}\left|\sup_{0\le i\le[t/(k_n\Delta_n)]} \left\|\widehat{c}'_{ik_n\Delta_n} - \bar{c}_{ik_n\Delta_n}\right\|^p\right| \le K\Delta_n^{-1}k_n^{-p/2-1} \to 0,$$

therefore, $\sup_{0\le i\le[t/(k_n\Delta_n)]} \left\|\widehat{c}'_{ik_n\Delta_n} - \bar{c}_{ik_n\Delta_n}\right\| = o_p(1)$. Moreover, by (A.2) we have

$$\mathbb{E}\left|\sup_{0\le i\le[t/(k_n\Delta_n)]} \left\|\widehat{c}_{ik_n\Delta_n} - \widehat{c}'_{ik_n\Delta_n}\right\|\right| \le \frac{1}{k_n\Delta_n}\sum_{i=0}^{[t/\Delta_n]-k_n} \mathbb{E}\left\|(\Delta^n_i X)(\Delta^n_i X)^\intercal 1_{\{\|\Delta^n_i X\| \le u_n\}} - (\Delta^n_i X')(\Delta^n_i X')^\intercal\right\|$$

$$\leq K a_n \Delta_n^{(2-\gamma)\varpi - 1 + \varsigma} \to 0.$$

As a result, we have as $\Delta_n \to 0$,

$$\sup_{0 \leq i \leq [t/(k_n\Delta_n)]} \|\widehat{c}_{ik_n\Delta_n} - \bar{c}_{ik_n\Delta_n}\| \xrightarrow{\mathrm{P}} 0. \tag{A.11}$$

Note that by Assumption 3, for $0 \leq s \leq t$, $c_s \in \mathcal{C} \cap \mathcal{M}^*(g_1, g_2, \ldots, g_r)$, where $\mathcal{C}$ is a convex and open set. Therefore, $\{\bar{c}_{ik_n\Delta_n}\}_{i=0,1,2,\ldots,[t/(k_n\Delta_n)]} \in \mathcal{C}$ by convexity. For n large enough, $\{\widehat{c}_{ik_n\Delta_n}\}_{i=0,1,2,\ldots,[t/(k_n\Delta_n)]} \in \mathcal{C}$, with probability approaching 1, by (A.11). Since $\bar{\mathcal{C}} \subset \mathcal{M}(g_1, g_2, \ldots, g_r)$, we can restrict the domain of f to the compact set $\lambda(\bar{\mathcal{C}}) \subset \mathcal{D}(g_1, g_2, \ldots, g_r)$, in which f is C^∞ with bounded derivatives. Moreover, because $\lambda_{g_j}(\cdot), 1 \leq j \leq r$ are continuous functions, $\min_{1\leq j\leq r-1}(\lambda_{g_j}(\cdot) - \lambda_{g_{j+1}}(\cdot))$ is hence continuous, so that $\inf_{c\in\mathcal{C}}\{\min_{1\leq j\leq r-1}(\lambda_{g_j}(c) - \lambda_{g_{j+1}}(c))\} \geq \epsilon > 0$. It follows from Lemma 4 and Theorem 3.5 of Silhavý (2000) that $F(\cdot)$ is C^∞ with bounded derivatives on $\mathcal{C}$.

2) Next, we have

$$\begin{aligned}\|V(\Delta_n, X; F) - V'(\Delta_n, X; F)\| &\leq k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \left\|F(\widehat{c}_{ik_n\Delta_n}) - F(\widehat{c}'_{ik_n\Delta_n})\right\| \\ &\leq K k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \left\|\widehat{c}_{ik_n\Delta_n} - \widehat{c}'_{ik_n\Delta_n}\right\|.\end{aligned}$$

By (A.3), we have

$$\mathrm{E}\left(\left\|\widehat{c}_{ik_n\Delta_n} - \widehat{c}'_{ik_n\Delta_n}\right\|\right) \leq K a_n \Delta_n^{(2-\gamma)\varpi},$$

where a_n is some sequence going to 0, as $n \to \infty$, which implies

$$V(\Delta_n, X; F) - V'(\Delta_n, X; F) = O_p(a_n\Delta_n^{(2-r)\varpi}). \tag{A.12}$$

As a result, given the conditions on ϖ, we have

$$k_n\left(V(\Delta_n, X; F) - V'(\Delta_n, X; F)\right) = o_p(1),$$

hence we can proceed with V' in the sequel.

3) Then we show for each $1 \leq h \leq d$, we have

$$\begin{aligned}k_n\Bigg(V'(\Delta_n, X; F_h) - k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \Big(&F_h(c_{ik_n\Delta_n}) \\ &- \frac{1}{2k_n} \sum_{j,k,l,m=1}^{d} \partial^2_{jk,lm} F_h(c_{ik_n\Delta_n}) \left(c_{jl,ik_n\Delta_n} c_{km,ik_n\Delta_n} + c_{jm,ik_n\Delta_n} c_{kl,ik_n\Delta_n}\right)\Big)\Bigg) = o_p(1).\end{aligned}$$

where F_h is the hth entry of the vector-valued function F.

To prove it, we decompose the left hand side into 4 terms:

$$\begin{aligned}R^n_{1,h} =& k_n^2\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \Big(F_h(c_{ik_n\Delta_n} + \beta^n_{ik_n}) - F_h(c_{ik_n\Delta_n}) - \sum_{l,m=1}^{d} \partial_{lm} F_h(c_{ik_n\Delta_n})\beta^{n,lm}_{ik_n} \\ &- \frac{1}{2} \sum_{j,k,l,m=1}^{d} \partial^2_{jk,lm} F_h(c_{ik_n\Delta_n})\beta^{n,lm}_{ik_n}\beta^{n,jk}_{ik_n}\Big),\end{aligned} \tag{A.13}$$

$$R^n_{2,h} = k_n^2\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \frac{1}{2} \sum_{j,k,l,m=1}^{d} \partial^2_{jk,lm} F_h(c_{ik_n\Delta_n}) \left(\beta^{n,lm}_{ik_n} \beta^{n,jk}_{ik_n} - \frac{1}{k_n} \left(c_{jl,ik_n\Delta_n} c_{km,ik_n\Delta_n} + c_{jm,ik_n\Delta_n} c_{kl,ik_n\Delta_n} \right) \right), \tag{A.14}$$

$$R^n_{3,h} = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \sum_{l,m=1}^{d} \partial_{lm} F_h(c_{ik_n\Delta_n}) \sum_{u=1}^{k_n} (c_{lm,(ik_n+u)\Delta_n} - c_{lm,ik_n\Delta_n}), \tag{A.15}$$

$$R^n_{4,h} = k_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \sum_{l,m=1}^{d} \partial_{lm} F_h(c_{ik_n\Delta_n}) \sum_{u=1}^{k_n} \alpha^{n,lm}_{ik_n+u}, \tag{A.16}$$

We first consider $R^n_{1,h}$. By (A.9), we have

$$\begin{aligned} \mathrm{E}(|R^n_{1,h}|) &\le K k_n^2 \Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \mathrm{E} \left\| \beta^n_{ik_n} \right\|^3 \le K k_n^2 \Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} (k_n^{-3/2} + k_n\Delta_n) \\ &\le K k_n^2 \Delta_n + K k_n^{-1/2} \to 0. \end{aligned}$$

As to $R^n_{2,h}$, we denote the term inside the summation of $R^n_{2,h}$ as $\nu^n_{ik_n}$. So we have

$$R^n_{2,h} = k_n^2 \Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \left(\nu^n_{ik_n} - \mathrm{E}(\nu^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n}) + \mathrm{E}(\nu^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n}) \right).$$

By (A.8), we have

$$|\mathrm{E}(\nu^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n})| \le K\sqrt{\Delta_n}.$$

On the other hand, by (A.9), we can derive

$$\mathrm{E} \left(\nu^n_{ik_n} - \mathrm{E}(\nu^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n}) \right)^2 \le K k_n \Delta_n.$$

Then Doob's inequality implies that

$$\mathrm{E} \left(\sup_{s \le t} \left| \sum_{i=0}^{[s/(k_n\Delta_n)]} \left(\nu^n_{ik_n} - \mathrm{E}(\nu^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n}) \right) \right| \right) \le Kt.$$

As a result,

$$\begin{aligned} \mathrm{E} \left(|R^n_{2,h}| \right) \le & k_n^2 \Delta_n \mathrm{E} \left(\sup_{s \le t} \left| \sum_{i=0}^{[s/(k_n\Delta_n)]} \left(\nu^n_{ik_n} - \mathrm{E}(\nu^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n}) \right) \right| \right) + k_n^2 \Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \left| \mathrm{E}(\nu^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n}) \right| \\ \le & K k_n^2 \Delta_n + K k_n \sqrt{\Delta_n} \to 0. \end{aligned}$$

The proof for $\mathrm{E}(|R^n_{3,h}|) \to 0$ is similar. Denote the term inside the summand as $\xi^n_{ik_n}$. By (A.1) and the Cauchy-Schwarz inequality, we have

$$|\mathrm{E}(\xi^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n})| \le K k_n^2 \Delta_n, \quad \mathrm{E} \left(|\xi^n_{ik_n}|^2 | \mathcal{F}_{ik_n\Delta_n} \right) \le K k_n^3 \Delta_n.$$

By Doob's inequality again,

$$\mathrm{E} \left(|R^n_{3,h}| \right) \le k_n \Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \mathrm{E} \left(\left| \mathrm{E} \left(\xi^n_{ik_n} | \mathcal{F}_{ik_n\Delta_n} \right) \right| \right) + k_n \Delta_n \left(\sum_{i=0}^{[t/(k_n\Delta_n)]} \mathrm{E} \left(|\xi^n_{ik_n}|^2 \right) \right)^{1/2}$$

$$\leq Kk_n^2\Delta_n \to 0.$$

For $R_{4,h}^n$, it can be shown in the proof of Theorem 2 below that $R_{4,h}^n = O_p(k_n\sqrt{\Delta_n}) = o_p(1)$.

4) Finally, it is sufficient to show that

$$k_n\left(\sum_{i=0}^{[t/(k_n\Delta_n)]}\left(\int_{ik_n\Delta_n}^{(i+1)k_n\Delta_n} F_h(c_{ik_n\Delta_n})ds - \int_{ik_n\Delta_n}^{(i+1)k_n\Delta_n} F_h(c_s)ds\right) - \int_{[t/(k_n\Delta_n)]k_n\Delta_n}^{t} F_h(c_s)ds\right) \xrightarrow{\text{p}} 0,$$

as the similar result holds if we replace $F_h(c_{ik_n\Delta_n})$ by $\partial^2_{jk,lm}F_h(c_{ik_n\Delta_n})\left(c_{jl,ik_n\Delta_n}c_{km,ik_n\Delta_n} + c_{jm,ik_n\Delta_n}c_{kl,ik_n\Delta_n}\right)$. Since F_h is bounded, the second term is bounded by $Kk_n^2\Delta_n \to 0$. As to the first term, we notice that

$$\zeta_{ik_n}^n = \int_{ik_n\Delta_n}^{(i+1)k_n\Delta_n} F_h(c_{ik_n\Delta_n})ds - \int_{ik_n\Delta_n}^{(i+1)k_n\Delta_n} F_h(c_s)ds$$

is measurable with respect to $\mathcal{F}_{(i+1)k_n\Delta_n}$, and that

$$|\mathrm{E}\left(\zeta_{ik_n}^n|\mathcal{F}_{ik_n\Delta_n}\right)| \leq K(k_n\Delta_n)^2, \quad \mathrm{E}\left(|\zeta_{ik_n}^n|^2|\mathcal{F}_{ik_n\Delta_n}\right) \leq K(k_n\Delta_n)^3,$$

so the same steps as in (2) and (3) yield the desired results. ∎

Appendix A.8 Proof of Theorem 2

Proof. To start, we decompose

$$\frac{1}{\sqrt{\Delta_n}}\left(\widetilde{V}(\Delta_n, X; F) - k_n\Delta_n\sum_{i=0}^{[t/(k_n\Delta_n)]}F(c_{ik_n\Delta_n})\right) = \frac{1}{k_n\sqrt{\Delta_n}}\left(R_1^n + R_2^n + R_3^n + R_4^n + R_5^n + R_6^n\right), \quad \text{(A.17)}$$

where $R_i^n = \left(R_{i,1}^n, R_{i,2}^n, \ldots, R_{i,d}^n\right)^\intercal$, for $i = 1, 2, 3, 4$, and 5, with $R_{1,h}^n$, $R_{2,h}^n$, $R_{3,h}^n$ and $R_{4,h}^n$ given by equations (A.13) - (A.16). In addition, $R_{5,h}^n$ and $R_{6,h}^n$ are given by

$$\begin{aligned}
R_{5,h}^n =& \frac{k_n\Delta_n}{2}\sum_{i=0}^{[t/(k_n\Delta_n)]}\sum_{j,k,l,m=1}^{d}\left(\partial^2_{jk,lm}F_h(c_{ik_n\Delta_n})\left(c_{jl,ik_n\Delta_n}c_{km,ik_n\Delta_n} + c_{jm,ik_n\Delta_n}c_{kl,ik_n\Delta_n}\right)\right.\\
&\left.-\partial^2_{jk,lm}F_h(\widehat{c}'_{ik_n\Delta_n})\left(\widehat{c}'_{jl,ik_n\Delta_n}\widehat{c}'_{km,ik_n\Delta_n} + \widehat{c}'_{jm,ik_n\Delta_n}\widehat{c}'_{kl,ik_n\Delta_n}\right)\right).\\
R_{6,h}^n =& \frac{k_n\Delta_n}{2}\sum_{i=0}^{[t/(k_n\Delta_n)]}\sum_{j,k,l,m=1}^{d}\left(\partial^2_{jk,lm}F_h(\widehat{c}'_{ik_n\Delta_n})\left(\widehat{c}'_{jl,ik_n\Delta_n}\widehat{c}'_{km,ik_n\Delta_n} + \widehat{c}'_{jm,ik_n\Delta_n}\widehat{c}'_{kl,ik_n\Delta_n}\right)\right.\\
&\left.-\partial^2_{jk,lm}F_h(\widehat{c}_{ik_n\Delta_n})\left(\widehat{c}_{jl,ik_n\Delta_n}\widehat{c}_{km,ik_n\Delta_n} + \widehat{c}_{jm,ik_n\Delta_n}\widehat{c}_{kl,ik_n\Delta_n}\right)\right) + k_n\left(V(\Delta_n, X; F) - V'(\Delta_n, X; F)\right).
\end{aligned}$$

We have shown in the proof of Proposition 1 that $R_i^n = O_p(k_n^2\Delta_n)$, for $i = 1, 2, 3$. Therefore, these terms do not contribute to the asymptotic variance of $\widetilde{V}'(\Delta_n, X; F)$.

Next, we show that $R_{5,h}^n$ is also $o_p(k_n\sqrt{\Delta_n})$. By (A.9) and the mean-value theorem, we have

$$\mathrm{E}|R_{5,h}^n| \leq Kk_n\Delta_n\sum_{i=0}^{[t/(k_n\Delta_n)]}\mathrm{E}\left\|c_{ik_n\Delta_n} - \widehat{c}'_{ik_n\Delta_n}\right\| \leq K(k_n^{-1/2} + k_n\Delta_n) = o_p(k_n\sqrt{\Delta_n}).$$

As to $R_{6,h}^n$, by (A.12) and the mean-value theorem, we have

$$\mathrm{E}|R_{6,h}^n| \leq Kk_n\Delta_n\sum_{i=0}^{[t/(k_n\Delta_n)]}\mathrm{E}\left\|\widehat{c}_{ik_n\Delta_n} - \widehat{c}'_{ik_n\Delta_n}\right\| = O_p(a_n\Delta_n^{(2-r)\varpi}) = o_p(k_n\sqrt{\Delta_n}).$$

Hence, $\varpi \geq \frac{1-\varsigma}{2-\gamma} > \frac{1}{4-2\gamma}$ is sufficient to warrant the desired rate.

As a result, except for the term that is related to R_4^n, all the remainder terms on the right-hand side of (A.17) vanish. We write R_4^n as

$$R_4^n = k_n \sum_{i=1}^{[t/(k_n\Delta_n)]k_n} \sum_{l,m=1}^{d} \omega_i^{n,lm} \alpha_i^{n,lm}, \quad \text{where} \quad \omega_i^{n,lm} = \partial_{lm} F(c_{[(i-1)/k_n]k_n\Delta_n}).$$

where $\omega_i^{n,lm}$ is a vector measurable with respect to $\mathcal{F}_{(i-1)\Delta_n}$, and $\|\omega_i^n\| \leq K$. To prove the stable convergence result, we start with

$$\frac{1}{\sqrt{\Delta_n}} \mathrm{E} \left\| \sum_{i=0}^{[t/(k_n\Delta_n)]k_n} \omega_i^{n,lm} \mathrm{E}\left(\alpha_i^{n,lm} | \mathcal{F}_{i\Delta_n}\right) \right\| \leq \frac{1}{\sqrt{\Delta_n}} \sum_{i=0}^{[t/(k_n\Delta_n)]k_n} K\Delta_n^{3/2}(\sqrt{\Delta_n} + \mathrm{E}\left(\eta_i^n\right)) \to 0,$$

where we use (A.6) and (A.10). Moreover, by (A.5), we have

$$\frac{1}{\Delta_n^2} \mathrm{E} \left(\sum_{i=0}^{[t/(k_n\Delta_n)]k_n} \|\omega_i^n\|^4 \mathrm{E}\left(\|\alpha_i^n\|^4 | \mathcal{F}_{i\Delta_n} \right) \right) \leq K\Delta_n \to 0.$$

Also, similar to (4.18) in Jacod and Rosenbaum (2013), we have $\mathrm{E}\left(\alpha_i^{n,lm}\Delta_i^n N | \mathcal{F}_{i\Delta_n}\right) = 0$, for any N that is an arbitrary bounded martingale orthogonal to W, which readily implies

$$\frac{1}{\sqrt{\Delta_n}} \sum_{i=0}^{[t/(k_n\Delta_n)]k_n} \omega_i^{n,lm} \mathrm{E}\left(\alpha_i^{n,lm}\Delta_i^n N | \mathcal{F}_{i\Delta_n}\right) \xrightarrow{\mathrm{p}} 0.$$

Finally, note that for any $1 \leq p, q \leq r$, by (A.7),

$$\frac{1}{\Delta_n} \sum_{i=0}^{[t/(k_n\Delta_n)]k_n} |\omega_{i,p}^{n,jk}\omega_{i,q}^{n,lm}| \left\| \mathrm{E}\left(\alpha_i^{n,jk}\alpha_i^{n,lm} | \mathcal{F}_{i\Delta_n}\right) - \left(c_{i\Delta_n,jl}c_{i\Delta_n,km} + c_{i\Delta_n,jm}c_{i\Delta_n,kl}\right)\Delta_n^2 \right\| \leq K\Delta_n^{1/2},$$

which implies

$$\begin{aligned}
&\frac{1}{\Delta_n} \sum_{i=0}^{[t/(k_n\Delta_n)]k_n} \omega_{i,p}^{n,jk}\omega_{i,q}^{n,lm} \mathrm{E}\left(\alpha_i^{n,jk}\alpha_i^{n,lm} | \mathcal{F}_{i\Delta_n}\right) \\
=&\frac{1}{\Delta_n} \sum_{i=0}^{[t/(k_n\Delta_n)]k_n} \omega_{i,p}^{n,jk}\omega_{i,q}^{n,lm} \left(c_{i\Delta_n,jl}c_{i\Delta_n,km} + c_{i\Delta_n,jm}c_{i\Delta_n,kl}\right)\Delta_n^2 \\
\xrightarrow{\mathrm{p}} &\int_0^t \partial_{jk}F_p(c_s)\partial_{lm}F_q(c_s)\left(c_{s,jl}c_{s,km} + c_{s,jm}c_{s,kl}\right).
\end{aligned}$$

Finally, by Theorem IX.7.28 of Jacod and Shiryaev (2003), we establish

$$\frac{1}{k_n\sqrt{\Delta_n}} R_n^4 \xrightarrow{\mathcal{L}-\mathrm{s}} \mathcal{W}_t,$$

where $\mathcal{W}_t$ is conditional Gaussian on an extension of the probably space, with a covariance matrix

$$\mathrm{E}\left(\mathcal{W}_{p,t}\mathcal{W}_{q,t} | \mathcal{F}\right) = \sum_{j,k,l,m=1}^{d} \int_0^t \partial_{jk}F_p(c_s)\partial_{lm}F_q(c_s)\left(c_{s,jl}c_{s,km} + c_{s,jm}c_{s,kl}\right).$$

∎

Appendix A.9 Proof of Proposition 2

Proof. As we have seen from the above proof, we have for any $c \in \mathcal{C}$,

$$\|\partial_{jk}F_p(c)\partial_{lm}F_q(c)\left(c_{jl}c_{km}+c_{jm}c_{kl}\right)\| \leq K(1+\|c\|^2),$$

which, combined with the same argument in the proof of Theorem 1, establishes the desired result. ∎

Appendix A.10 Proof of Corollary 1

Proof. The first statement on consistency follows immediately from Theorem 1, as Assumption 2 holds with $\zeta = 1$. Next, we prove the central limit result. For any $1 \leq p \leq d$, we define f_p^λ as,

$$f_p^\lambda(\bar{x}) = \frac{1}{g_p - g_{p-1}} \sum_{j=g_{p-1}+1}^{g_p} \bar{x}_j,$$

hence we have

$$\partial f_p^\lambda(\bar{x}) = \frac{1}{g_p - g_{p-1}} \sum_{v=g_{p-1}+1}^{g_p} e^v, \quad \partial^2 f_p^\lambda(\bar{x}) = 0,$$

and f^λ is C^∞ and Lipchitz. By Lemma 4 we can derive

$$\partial_{jk}F_p^\lambda(c_s) = \sum_{u=1}^{d} O_{uj}\partial_u f_p^\lambda(\lambda(c_s))O_{uk} = \frac{1}{g_p - g_{p-1}}\sum_{u=1}^{d}\sum_{v=g_{p-1}+1}^{g_p} O_{uj}e_u^v O_{uk} = \frac{1}{g_p - g_{p-1}}\sum_{v=g_{p-1}+1}^{g_p} O_{vj}O_{vk}.$$

Therefore, the asymptotic covariance matrix is given by

$$\begin{aligned}
&\int_0^t \sum_{j,k,l,m=1}^{d} \partial_{jk}F_p^\lambda(c_s)\partial_{lm}F_q^\lambda(c_s)\left(c_{jl,s}c_{km,s}+c_{jm,s}c_{kl,s}\right)ds \\
=&\frac{1}{g_p - g_{p-1}}\frac{1}{g_q - g_{q-1}}\int_0^t \sum_{j,k,l,m=1}^{d}\sum_{v=g_{p-1}+1}^{g_p}\sum_{u=g_{q-1}+1}^{g_q} O_{vj}O_{vk}O_{ul}O_{um}\left(c_{jl,s}c_{km,s}+c_{jm,s}c_{kl,s}\right)ds \\
=&\frac{2}{(g_p - g_{p-1})(g_q - g_{q-1})}\int_0^t \sum_{l,m=1}^{d}\sum_{v=g_{p-1}+1}^{g_p}\sum_{u=g_{q-1}+1}^{g_q} O_{vl}O_{vm}O_{ul}O_{um}\lambda_{v,s}^2 ds \\
=&\frac{2}{(g_p - g_{p-1})(g_q - g_{q-1})}\int_0^t \sum_{v=g_{p-1}+1}^{g_p}\sum_{u=g_{q-1}+1}^{g_q} \delta_{u,v}\lambda_{v,s}^2 ds \\
=&\frac{2\delta_{p,q}}{(g_p - g_{p-1})}\int_0^t \lambda_{g_p,s}^2 ds.
\end{aligned} \tag{A.18}$$

Next, we calculate the bias-correction term. Recall that the estimator is given by

$$F_p^\lambda(\widehat{c}_{ik_n\Delta_n}) = \frac{1}{g_p - g_{p-1}}\sum_{v=g_{p-1}+1}^{g_p}\widehat{\lambda}_{v,ik_n\Delta_n},$$

where $\widehat{\lambda}_{v,ik_n\Delta_n}$ is the corresponding eigenvalue of the sample covariance matrix $\widehat{c}_{ik_n\Delta_n}$. Although $\widehat{c}_{ik_n\Delta_n}$ and $c_{ik_n\Delta_n}$ may have different eigenstructure, it is easy to verify that the functional forms of the second order derivative of F_p^λ evaluated at both points turn out to be the same, so here we only provide the calculations based on $\widehat{c}_{ik_n\Delta_n}$. Since almost surely, sample eigenvalues are simple, it implies from Lemma 4 that

$$\partial^2_{jk,lm}F_p^\lambda(\widehat{c}_{ik_n\Delta_n}) = \sum_{u,v=1}^{d}\mathcal{A}_{uv}^{f_p^\lambda}(\lambda(\widehat{c}_{ik_n\Delta_n}))\widehat{O}_{ul}\widehat{O}_{uj}\widehat{O}_{vk}\widehat{O}_{vm}$$

$$= \frac{1}{g_p - g_{p-1}} \sum_{h=g_{p-1}+1}^{g_p} \sum_{u,v=1,u\neq v}^{d} \frac{e_u^h - e_v^h}{\widehat{\lambda}_{u,i\Delta_n} - \widehat{\lambda}_{v,i\Delta_n}} \widehat{O}_{ul}\widehat{O}_{uj}\widehat{O}_{vk}\widehat{O}_{vm}$$
$$= \frac{1}{g_p - g_{p-1}} \sum_{h=g_{p-1}+1}^{g_p} \sum_{u=1,u\neq h}^{d} \frac{1}{\widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{u,ik_n\Delta_n}} \left(\widehat{O}_{ul}\widehat{O}_{uj}\widehat{O}_{hk}\widehat{O}_{hm} + \widehat{O}_{hl}\widehat{O}_{hj}\widehat{O}_{uk}\widehat{O}_{um}\right),$$

where $\widehat{O}$ is the orthogonal matrix such that $\widehat{O}\widehat{c}_{ik_n\Delta_n}\widehat{O}^\intercal = \mathrm{Diag}(\lambda(\widehat{c}_{ik_n\Delta_n}))$. The dependence of $\widehat{O}$ on $ik_n\Delta_n$ is omitted for brevity.

To facilitate the implementation, we consider the matrix $\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}$. Note that

$$\mathrm{Diag}(\widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{1,ik_n\Delta_n}, \widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{2,ik_n\Delta_n}, \ldots, \widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{d,ik_n\Delta_n}) = \widehat{O}(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n})\widehat{O}^\intercal,$$

hence we have

$$\left(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}\right)^+ = \widehat{O}^\intercal \mathrm{Diag}\left(\frac{1}{\widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{1,ik_n\Delta_n}}, \frac{1}{\widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{2,ik_n\Delta_n}}, \ldots, 0, \ldots \frac{1}{\widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{d,ik_n\Delta_n}}\right)\widehat{O}.$$

As a result, we obtain

$$\left(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}\right)^+_{km} = \sum_{u=1,u\neq p}^{d} \frac{1}{\widehat{\lambda}_{h,ik_n\Delta_n} - \widehat{\lambda}_{u,ik_n\Delta_n}} \widehat{O}_{uk}\widehat{O}_{um},$$

Therefore, we have

$$\partial^2_{jk,lm}F_p^\lambda(\widehat{c}_{ik_n\Delta_n}) = \frac{1}{g_p - g_{p-1}} \sum_{h=g_{p-1}+1}^{g_p} \widehat{O}_{hk}\left(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}\right)^+_{jl}\widehat{O}_{hm} + \widehat{O}_{hj}\left(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}\right)^+_{km}\widehat{O}_{hl}.$$

Now we can calculate the following term, which is used for bias-correction:

$$\begin{aligned}
&\sum_{j,k,l,m=1}^{d} \partial^2_{jk,lm}F_p^\lambda(\widehat{c}_{i\Delta_n})\left(\widehat{c}_{jl,i\Delta_n}\widehat{c}_{km,i\Delta_n} + \widehat{c}_{jm,i\Delta_n}\widehat{c}_{kl,i\Delta_n}\right)\\
=&\frac{1}{g_p - g_{p-1}} \sum_{h=g_{p-1}+1}^{g_p} \sum_{j,k,l,m=1}^{d} \left(\widehat{O}_{hk}\left(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}\right)^+_{jl}\widehat{O}_{hm} + \widehat{O}_{hj}\left(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}\right)^+_{km}\widehat{O}_{hl}\right)\\
&\cdot\left(\widehat{c}_{jl,ik_n\Delta_n}\widehat{c}_{km,ik_n\Delta_n} + \widehat{c}_{jm,ik_n\Delta_n}\widehat{c}_{kl,ik_n\Delta_n}\right)\\
=&\frac{2}{g_p - g_{p-1}} \sum_{h=g_{p-1}+1}^{g_p} \widehat{\lambda}_{h,ik_n\Delta_n}\mathrm{Tr}\left(\left(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n}\right)^+\widehat{c}_{ik_n\Delta_n}\right). \qquad \text{(A.19)}
\end{aligned}$$

The last equality uses the following observation:[6]

$$(\widehat{\lambda}_{h,ik_n\Delta_n}\mathbb{I} - \widehat{c}_{ik_n\Delta_n})^+\widehat{O}^\intercal_{h,\cdot} = 0,$$

which concludes the proof of (ii). The proof of (iii) uses the same calculations as above. Note that we can apply Theorem 2 with only Assumption 4, because the spectral function here only depends on λ_g. ∎

Appendix A.11 Proof of Proposition 3

Proof. By Lemma 5 and the uniform convergence of $\widehat{c}_{ik_n\Delta_n} - \bar{c}_{ik_n\Delta_n}$ to 0 established above, we can restrict the domain of $\gamma_g(\cdot)$ to the set $\mathcal{C}$, in which it is C^∞ with bounded derivatives. By Theorem 21 of Protter (2004),

[6] See page 160 of Magnus and Neudecker (1999).

we have

$$\sum_{i=1}^{[t/(k_n\Delta_n)]-1} \gamma^{\intercal}_{g,ik_n\Delta_n}(X_{(i+1)k_n\Delta_n} - X_{ik_n\Delta_n}) \overset{\text{u.c.p}}{\Longrightarrow} \int_0^t \gamma_{g,s-}dX_s.$$

Therefore, it remains to show that

$$\sum_{i=1}^{[t/(k_n\Delta_n)]-1} \left(\widehat{\gamma}^{\intercal}_{g,(i-1)k_n\Delta_n} - \gamma^{\intercal}_{g,ik_n\Delta_n}\right)(X_{(i+1)k_n\Delta_n} - X_{ik_n\Delta_n}) \overset{\text{u.c.p}}{\Longrightarrow} 0.$$

Define a $\mathcal{F}_{(i+1)k_n\Delta_n}$-measurable function:

$$\xi_{ik_n} = \left(\widehat{\gamma}^{\intercal}_{g,(i-1)k_n\Delta_n} - \gamma^{\intercal}_{g,ik_n\Delta_n}\right)(X_{(i+1)k_n\Delta_n} - X_{ik_n\Delta_n}).$$

By standard estimates in (A.1) with c replaced by X, (A.3), and (A.9),

$$\begin{aligned}
\mathrm{E}|\mathrm{E}\left(\xi_{ik_n}|\,\mathcal{F}_{ik_n\Delta_n}\right)| =& \mathrm{E}|\widehat{\gamma}_{g,(i-1)k_n\Delta_n} - \gamma_{g,ik_n\Delta_n}||\mathrm{E}\left((X_{(i+1)k_n\Delta_n} - X_{ik_n\Delta_n})|\mathcal{F}_{ik_n\Delta_n}\right)| \\
\leq& K\mathrm{E}|\widehat{c}_{(i-1)k_n\Delta_n} - c_{ik_n\Delta_n}|(k_n\Delta_n) \\
\leq& K\left((k_n\Delta_n)^{1/2} + a_n\Delta_n^{(2-\gamma)\varpi} + \sqrt{k_n^{-1} + k_n\Delta_n}\right)(k_n\Delta_n)
\end{aligned}$$

Moreover, we have by the same estimates above,

$$\mathrm{E}\left(|\xi_{ik_n}|^2|\mathcal{F}_{ik_n\Delta_n}\right) \leq (k_n\Delta_n + a_n\Delta_n^{(4-\gamma)\varpi-1} + k_n^{-1} + k_n\Delta_n)k_n\Delta_n.$$

Finally, using Doob's inequality, and measurability of ξ_{ik_n}, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\sup_{0\leq s\leq t}|\sum_{i=1}^{[s/(k_n\Delta_n)]-1}\xi_{ik_n}|\right) \\
\leq& \sum_{i=1}^{[t/(k_n\Delta_n)]-1}\mathrm{E}|\mathrm{E}\left(\xi_{ik_n}|\,\mathcal{F}_{ik_n\Delta_n}\right)| + \left(\sum_{i=1}^{[t/(k_n\Delta_n)]-1}\mathrm{E}\left(|\xi_{ik_n}|^2|\mathcal{F}_{ik_n\Delta_n}\right)\right)^{1/2} \\
\leq& K\left((k_n\Delta_n)^{1/2} + a_n\Delta_n^{(2-\gamma)\varpi} + \sqrt{k_n^{-1} + k_n\Delta_n}\right) + K(k_n\Delta_n + a_n\Delta_n^{(4-\gamma)\varpi-1} + k_n^{-1} + k_n\Delta_n)^{1/2} \\
\to& 0,
\end{aligned}$$

because $(4-\gamma)\varpi \geq 1$ under our assumptions on ϖ and ς, which establishes the proof. ∎

Appendix A.12 Proof of Corollary 2

Proof. The (p,q) entry of the asymptotic covariance matrix is given by

$$\begin{aligned}
&\int_0^t \sum_{j,k,l,m=1}^{d} \partial_{jk}\gamma_{gp,s}\partial_{lm}\gamma_{gq,s}\left(c_{jl,s}c_{km,s} + c_{jm,s}c_{kl,s}\right)ds \\
=& \int_0^t \sum_{j,k,l,m=1}^{d} (\lambda_{g,s}\mathbb{I} - c_s)^+_{pj}(\lambda_{g,s}\mathbb{I} - c_s)^+_{ql}\gamma_{gk,s}\gamma_{gm,s}\left(c_{jl,s}c_{km,s} + c_{jm,s}c_{kl,s}\right)ds \\
=& \int_0^t \sum_{j,l=1}^{d} (\lambda_{g,s}\mathbb{I} - c_s)^+_{pj}(\lambda_{g,s}\mathbb{I} - c_s)^+_{ql}\left(\lambda_{g,s}c_{jl} + \lambda^2_{g,s}\gamma_{gl,s}\gamma_{gj,s}\right)ds
\end{aligned}$$

$$= \int_0^t \lambda_{g,s} \left((\lambda_{g,s}\mathbb{I} - c_s)^+ c_s (\lambda_{g,s}\mathbb{I} - c_s)^+ \right)_{p,q} ds,$$

where we use $(\lambda_{g,s}\mathbb{I} - c_s)^+ \gamma_{g,s} = 0$, and $\sum_{k=1}^d \gamma_{gk,s} c_{km,s} = \lambda_{g,s}\gamma_{gm,s}$. To calculate the asymptotic bias, we note that the $\widehat{c}_s$ has only simple eigenvalues almost surely. Denote $\widehat{\lambda}_h$ and $\widehat{\gamma}_h$ as the corresponding eigenvalue and eigenvector. We omit the dependence on time s to simplify the notations. By Lemma 2, we obtain

$$\begin{aligned}
&\sum_{j,k,l,m=1}^d \partial^2_{jk,lm} \widehat{\gamma}_{gh} (\widehat{c}_{jl}\widehat{c}_{km} + \widehat{c}_{jm}\widehat{c}_{kl}) \\
=& -2 \sum_{p\neq g} \frac{\widehat{\lambda}_p \widehat{\lambda}_g}{(\widehat{\lambda}_g - \widehat{\lambda}_p)^2} \sum_{l,m=1}^d \widehat{\gamma}^2_{gl} \widehat{\gamma}_{pm} \widehat{\gamma}_{ph} \widehat{\gamma}_{gm} + \sum_{p\neq g} \frac{\widehat{\lambda}_p \widehat{\lambda}_g}{(\widehat{\lambda}_g - \widehat{\lambda}_p)^2} \left(\sum_{l,m=1}^d \widehat{\gamma}^2_{pl} \widehat{\gamma}_{gm} \widehat{\gamma}_{ph} \widehat{\gamma}_{pm} + \sum_{l,m=1}^d \widehat{\gamma}^2_{pm} \widehat{\gamma}_{gl} \widehat{\gamma}_{pl} \widehat{\gamma}_{ph} \right) \\
&+ \sum_{p\neq g} \sum_{q\neq p} \frac{\widehat{\lambda}_p \widehat{\lambda}_g}{(\widehat{\lambda}_g - \widehat{\lambda}_p)(\widehat{\lambda}_p - \widehat{\lambda}_q)} \left(\sum_{l,m=1}^d \widehat{\gamma}_{ql} \widehat{\gamma}_{pl} \widehat{\gamma}_{pm} \widehat{\gamma}_{gm} \widehat{\gamma}_{qh} + \sum_{l,m=1}^d \widehat{\gamma}^2_{pm} \widehat{\gamma}_{gl} \widehat{\gamma}_{ql} \widehat{\gamma}_{qh} \right) \\
&+ \sum_{p\neq g} \sum_{q\neq p} \frac{\widehat{\lambda}_q \widehat{\lambda}_g}{(\widehat{\lambda}_g - \widehat{\lambda}_p)(\widehat{\lambda}_p - \widehat{\lambda}_q)} \left(\sum_{l,m=1}^d \widehat{\gamma}^2_{ql} \widehat{\gamma}_{gm} \widehat{\gamma}_{ph} \widehat{\gamma}_{pm} + \sum_{l,m=1}^d \widehat{\gamma}_{qm} \widehat{\gamma}_{gl} \widehat{\gamma}_{ql} \widehat{\gamma}_{pm} \widehat{\gamma}_{ph} \right) \\
&+ \sum_{p\neq g} \sum_{q\neq g} \frac{\widehat{\lambda}_p \widehat{\lambda}_q}{(\widehat{\lambda}_g - \widehat{\lambda}_p)(\widehat{\lambda}_g - \widehat{\lambda}_q)} \left(\sum_{l,m=1}^d \widehat{\gamma}_{pm} \widehat{\gamma}^2_{ql} \widehat{\gamma}_{gm} \widehat{\gamma}_{ph} + \sum_{l,m=1}^d \widehat{\gamma}_{qm} \widehat{\gamma}_{pl} \widehat{\gamma}_{ql} \widehat{\gamma}_{gm} \widehat{\gamma}_{ph} \right) \\
=& -\sum_{p\neq g} \frac{\widehat{\lambda}_p \widehat{\lambda}_g}{(\widehat{\lambda}_g - \widehat{\lambda}_p)^2} \widehat{\gamma}_{gh}.
\end{aligned}$$

Since $\gamma_g(\cdot)$ is a C^∞ function, it is straightforward using the proof of Theorem 2 that the desired CLT holds, even though $\gamma_g(\cdot)$ is not a spectral function. ∎

Appendix A.13 Proof of Propositions 4 and 5

Proof. This follows by applying the "delta" method, using Lemma 2 and Theorem 13.2.4 in Jacod and Protter (2011). ∎

Appendix B Figures and Tables

	κ_j	θ_j	η_j	ρ_j	μ_j	$\tilde{\kappa}_j$	$\tilde{\theta}_{i,j}$	$\tilde{\xi}_j$
$j=1$	3	0.05	0.3	-0.6	0.05	1	$\mathcal{U}[0.25, 1.75]$	0.5
$j=2$	4	0.04	0.4	-0.4	0.03	2	$\mathcal{N}(0, 0.5^2)$	0.6
$j=3$	5	0.03	0.3	-0.25	0.02	3	$\mathcal{N}(0, 0.5^2)$	0.7
	λ^F	$\mu^F_{+/-}$	λ^Z	$\mu^Z_{+/-}$	μ^{σ^2}	ρ^F_{12}	ρ^F_{13}	ρ^F_{23}
	$1/t$	$4\sqrt{\Delta}$	$2/t$	$6\sqrt{\Delta}$	$\sqrt{\Delta}$	0.05	0.1	0.15
						κ	θ	η
						4	0.3	0.06

Table 1: Parameters in Monte Carlo Simulations

Note: In this table, we report the parameter values used in the simulations. The constant matrix $\tilde{\theta}_{i,j}$ is generated randomly from the described distribution, and is fixed throughout all replications. The dimension of X_t is 100, whereas the dimension of F_t is 3. Δ is the sampling frequency, and t is the length of the time window. The number of Monte Carlo replications is 1,000.

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.4686	-0.0001	0.0075	0.4703	0.0007	0.0152
10	0.6489	0.0001	0.0110	0.6552	-0.0014	0.0223
15	0.8927	-0.0004	0.0149	0.8975	-0.0011	0.0296
20	1.3044	0.0003	0.0225	1.3148	-0.0018	0.0424
30	2.1003	-0.0002	0.0356	2.1134	-0.0033	0.0688
50	2.9863	0.0002	0.0514	3.0104	-0.0054	0.1002
100	6.6270	-0.0004	0.1141	6.6732	-0.0127	0.2179
	1 Week, 5 Minutes			1 Month, 5 Minutes		
5	0.4879	0.0107	0.0452	0.5642	0.0006	0.0247
10	0.6839	0.0084	0.0574	0.7143	-0.0024	0.0258
15	0.9397	0.0128	0.0724	1.0167	-0.0024	0.0382
20	1.3765	0.0130	0.1076	1.3882	-0.0041	0.0503
30	2.2157	0.0210	0.1670	2.2383	-0.0073	0.0806
50	3.1554	0.0267	0.2410	3.1518	-0.0125	0.1155
100	7.0000	0.0552	0.5314	6.9632	-0.0270	0.2451

Table 2: Simulation Results: First Eigenvalue Estimation

Note: In this table, we report the summary statistics of 1,000 Monte Carlo simulations for estimating the first integrated eigenvalue. Column "True" corresponds to the average of the true integrated eigenvalue; Column "Bias" corresponds to the mean of the estimation error; Column "Stdev" is the standard deviation of the estimation error.

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.3145	0.00004	0.0051	0.3173	0.0003	0.0110
10	0.4268	-0.0001	0.0071	0.4289	0.0006	0.0146
15	0.5531	-0.0003	0.0086	0.5572	0.0004	0.0188
20	0.6517	-0.0008	0.0103	0.6556	-0.0006	0.0215
30	0.9186	-0.0015	0.0144	0.9251	-0.0017	0.0305
50	1.2993	-0.0017	0.0205	1.3080	-0.0026	0.0458
100	2.3273	-0.0041	0.0361	2.3441	-0.0065	0.0766
	1 Week, 5 Minutes			1 Month, 5 Minutes		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.3255	0.0020	0.0269	0.3639	-0.0003	0.0189
10	0.4384	0.0028	0.0379	0.4469	-0.0003	0.0186
15	0.5751	0.0014	0.0427	0.6524	-0.0017	0.0257
20	0.6758	0.0046	0.0535	0.7779	-0.0021	0.0297
30	0.9586	0.0017	0.0725	1.1365	-0.0036	0.0438
50	1.3508	-0.0022	0.1046	1.5630	-0.0064	0.0683
100	2.4312	-0.0084	0.1787	2.9148	-0.0136	0.1113

Table 3: Simulation Results: Second Eigenvalue Estimation

Note: In this table, we report the summary statistics of 1000 Monte Carlo simulations for estimating the second integrated eigenvalue. Column "True" corresponds to the average of the true integrated eigenvalue; Column "Bias" corresponds to the mean of the estimation error; Column "Stdev" is the standard deviation of the estimation error.

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.1338	0.0001	0.0022	0.1345	0.0003	0.0044
10	0.2132	0.0001	0.0035	0.2149	0.0002	0.0070
15	0.2941	0.0002	0.0046	0.2954	0.0002	0.0093
20	0.3212	0.0001	0.0052	0.3242	0.0000	0.0105
30	0.4825	0.0005	0.0078	0.4853	0.0001	0.0153
50	0.6943	0.0001	0.0113	0.7016	-0.0005	0.0226
100	1.3808	0.0004	0.0221	1.3935	-0.0013	0.0438
	1 Week, 5 Minutes			1 Month, 5 Minutes		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.1364	0.0041	0.0124	0.1404	0.0006	0.0054
10	0.2199	0.0033	0.0183	0.2367	-0.0001	0.0091
15	0.3018	0.0056	0.0264	0.3396	0.0009	0.0131
20	0.3324	0.0037	0.0294	0.3546	0.0003	0.0132
30	0.4971	0.0055	0.0409	0.5686	0.0006	0.0211
50	0.7216	0.0028	0.0577	0.8051	-0.0010	0.0306
100	1.4302	-0.0016	0.1119	1.6278	-0.0028	0.0612

Table 4: Simulation Results: Third Eigenvalue Estimation

Note: In this table, we report the summary statistics of 1,000 Monte Carlo simulations for estimating the third integrated eigenvalue. Column "True" corresponds to the average of the true integrated eigenvalue; Column "Bias" corresponds to the mean of the estimation error; Column "Stdev" is the standard deviation of the estimation error.

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.0596	0.0002	0.0007	0.0597	0.0005	0.0015
10	0.0596	0.0001	0.0004	0.0597	0.0003	0.0008
15	0.0596	0.0001	0.0003	0.0597	0.0004	0.0006
20	0.0596	0.0001	0.0002	0.0597	0.0004	0.0005
30	0.0596	0.0001	0.0002	0.0597	0.0004	0.0004
50	0.0596	0.0001	0.0001	0.0597	0.0004	0.0003
100	0.0596	0.0001	0.0001	0.0597	0.0004	0.0002
	1 Week, 5 Minutes			1 Month, 5 Minutes		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
5	0.0595	0.0026	0.0041	0.0595	0.0006	0.0017
10	0.0595	0.0029	0.0029	0.0595	0.0006	0.0009
15	0.0595	0.0027	0.0024	0.0595	0.0006	0.0007
20	0.0595	0.0028	0.0021	0.0595	0.0006	0.0006
30	0.0595	0.0030	0.0020	0.0595	0.0007	0.0005
50	0.0595	0.0029	0.0018	0.0595	0.0006	0.0004
100	0.0595	0.0031	0.0016	0.0595	0.0006	0.0003

Table 5: Simulation Results: Repeated Eigenvalue Estimation, Fourth and Beyond

Note: In this table, we report the summary statistics of 1,000 Monte Carlo simulations for estimating the repeated integrated eigenvalues. Column "True" corresponds to the average of the true integrated eigenvalue; Column "Bias" corresponds to the mean of the estimation error; Column "Stdev" is the standard deviation of the estimation error.

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
	0.2457	-0.0002	0.0173	0.2475	0.0074	0.1559
	0.8987	0.0028	0.0261	0.8963	0.0174	0.2135
5	0.0566	0.0013	0.0235	0.0536	-0.0018	0.1281
	0.0633	0.0001	0.0119	0.0639	0.0037	0.0786
	0.3110	0.0017	0.0202	0.3086	0.0042	0.0645
	0.0506	0.0004	0.0049	0.0495	0.0029	0.0266
	0.3444	0.0006	0.0087	0.3452	0.0080	0.0691
	0.4632	0.0011	0.0129	0.4619	0.0121	0.0966
	0.1422	0.0008	0.0137	0.1422	0.0094	0.0990
10	0.4166	0.0004	0.0220	0.4164	0.0045	0.1562
	0.0864	0.0008	0.0158	0.0863	0.0089	0.1117
	0.3460	0.0008	0.0088	0.3440	0.0101	0.0600
	0.1268	0.0005	0.0076	0.1259	0.0066	0.0451
	0.3409	0.0005	0.0174	0.3415	0.0046	0.1268
	0.4262	0.0007	0.0112	0.4271	0.0099	0.0942

Table 6: Simulation Results: Eigenvector Estimation

Note: In this table, we report the summary statistics of 1,000 Monte Carlo simulations for estimating the integrated eigenvector associated with the first eigenvalue. Column "True" corresponds to the average of the true integrated vector; Column "Bias" corresponds to the mean of the estimation error; Column "Stdev" is the standard deviation of the estimation error. The setup for these simulations is identical to that of Tables 2-5.

	1 Week, 5 Seconds			1 Week, 1 Minute		
# Stocks	True	Bias	Stdev	True	Bias	Stdev
	0.2587	-4×10^{-6}	0.0026	0.2581	0.00004	0.0090
	0.1557	3×10^{-7}	0.0015	0.1558	-0.0001	0.0054
	0.2364	-0.00004	0.0015	0.2359	-0.00008	0.0053
	0.0608	1×10^{-6}	0.0028	0.0608	0.00009	0.0096
	0.1398	-0.00001	0.0019	0.1394	0.00002	0.0069
	0.2471	-3×10^{-6}	0.0017	0.2476	-0.00001	0.0059
	0.2250	0.00001	0.0014	0.2244	0.0001	0.0043
	0.2642	0.00002	0.0027	0.2647	0.00008	0.0094
	0.2004	0.00005	0.0015	0.2006	0.0002	0.0047
	0.0446	-0.00004	0.0022	0.0447	0.0002	0.0075
	0.2311	0.00005	0.0023	0.2312	0.00001	0.0080
	0.2560	0.00002	0.0021	0.2565	-0.00002	0.0075
	0.1979	-0.00005	0.0013	0.1979	0.00008	0.0043
	0.2255	-0.00004	0.0019	0.2247	-0.0001	0.0066
	0.2185	6×10^{-7}	0.0014	0.2186	-0.00007	0.0042
30	0.1262	-0.00006	0.0015	0.1263	0.0002	0.0051
	0.1916	0.00002	0.0013	0.1917	-0.0001	0.0045
	0.0661	-0.00004	0.0025	0.0661	0.0002	0.0088
	0.2118	-0.00004	0.0020	0.2117	-0.00003	0.0070
	0.0414	0.00006	0.0013	0.0414	0.00008	0.0045
	0.1007	-0.00004	0.0025	0.1008	-0.00006	0.0086
	0.0518	-0.00006	0.0017	0.0517	-0.00003	0.0058
	0.0550	-0.00001	0.0022	0.0548	0.0002	0.0077
	0.2345	-0.00005	0.0013	0.2349	-0.0002	0.0043
	0.2003	0.00003	0.0014	0.2006	-0.0001	0.0047
	0.1276	-0.00003	0.0025	0.1275	0.00008	0.0085
	0.2813	0.00002	0.0033	0.2813	0.0003	0.0116
	0.0319	-0.00005	0.0030	0.0319	-0.00001	0.0106
	0.1417	0.00006	0.0022	0.1414	0.0002	0.0075
	0.1233	-0.00001	0.0024	0.1235	0.0002	0.0085

Table 7: Simulation Results: Eigenvector Estimation

Note: In this table, we report the summary statistics of 1,000 Monte Carlo simulations for estimating the integrated eigenvectors associated with the first eigenvalues. The column "True" corresponds to the average of the true integrated vector; "Bias" corresponds to the mean of the estimation error; "Stdev" is the standard deviation of the estimation error. The setup for these simulations is identical to that of Tables 2-5. To save space, we do not report the eigenvectors in dimensions larger than 30.

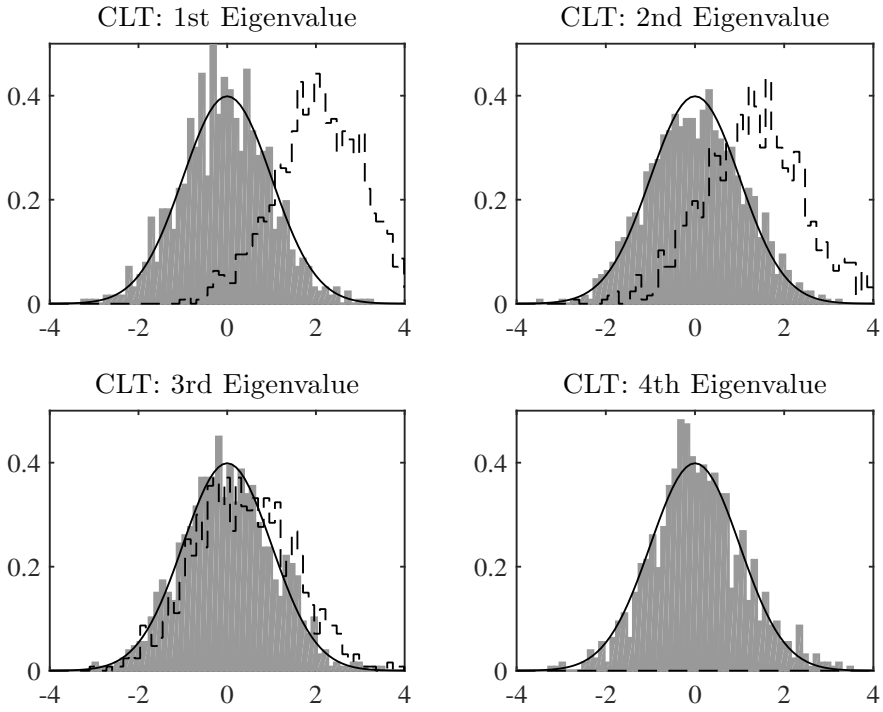

Figure 1: Finite Sample Distribution of the Standardized Statistics.

Note: In this figure, we report the histograms of the 1,000 simulation results for estimating the first four integrated eigenvalues using 5-second returns for 30 stocks over one week. The purpose of this figure is to validate the asymptotic theory, hence the use of a short 5-second sampling interval. The smallest 27 eigenvalues are identical so that the fourth eigenvalue is repeated. The solid lines plot the standard normal density; the dashed histograms report the distribution of the estimates before bias correction; the solid histograms report the distribution of the estimates after bias correction is applied and is to be compared to the asymptotic standard normal. Because the fourth eigenvalue is small, the dashed histogram on the fourth subplot is out of the x-axis range, to the right.

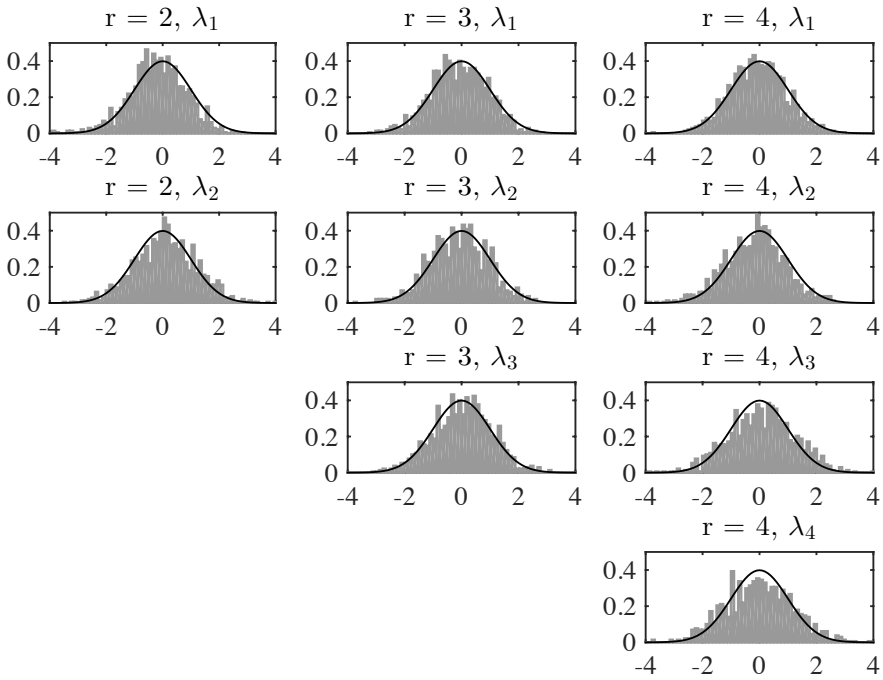

Figure 2: Finite Sample Distribution of the Standardized Statistics.

Note: In this figure, we report the histograms of the integrated simple eigenvalues using weekly one-minute returns of 100 stocks, a setting that matches that of our empirical analysis below. Columns 1, 2, and 3 report the results with $r = 2$, 3, and 4 common factors in the data generating process, respectively. The number of Monte Carlo replications is 1,000.

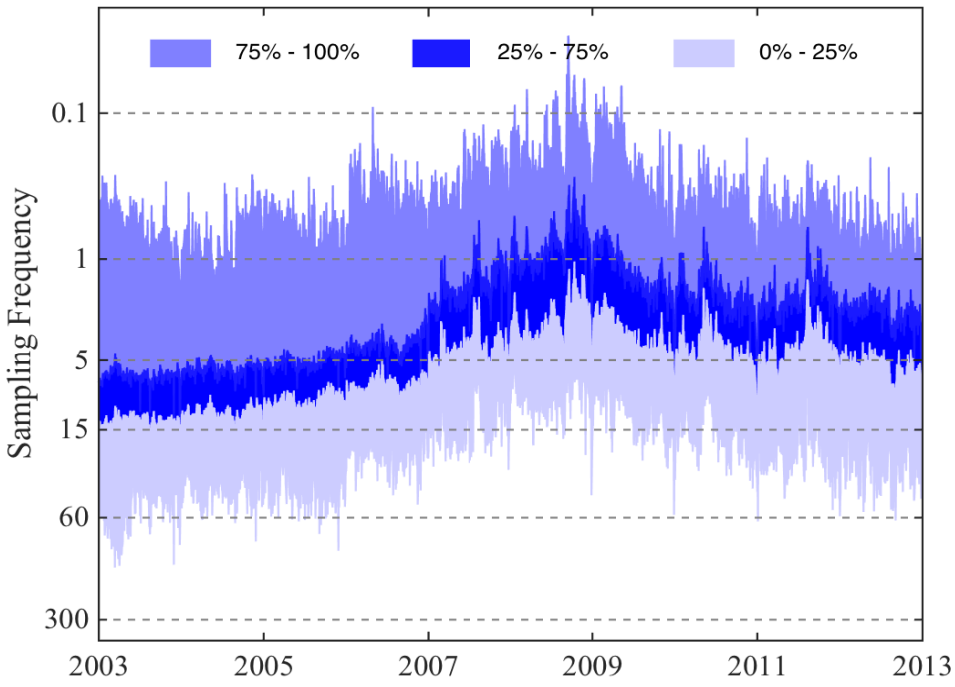

Figure 3: The Liquidity of S&P 100 Index Components

Note: In this figure, we provide quartiles of the average time between successive transactions, or sampling frequencies, for the 100 stocks S&P 100 Index constituents between 2003 and 2012 . We exclude the first and the last 5 minutes from the 6.5 regular trading hours in the calculation, during which trading intensities are unusually high. The y-axis reports the corresponding average sampling frequencies in seconds, computed from the high frequency transactions in the sample. From these transactions, we construct a synchronous one-minute sample using the latest tick recorded within that minute. The 10 least liquid stocks in the index are included in this figure but excluded from the empirical analysis that follows.

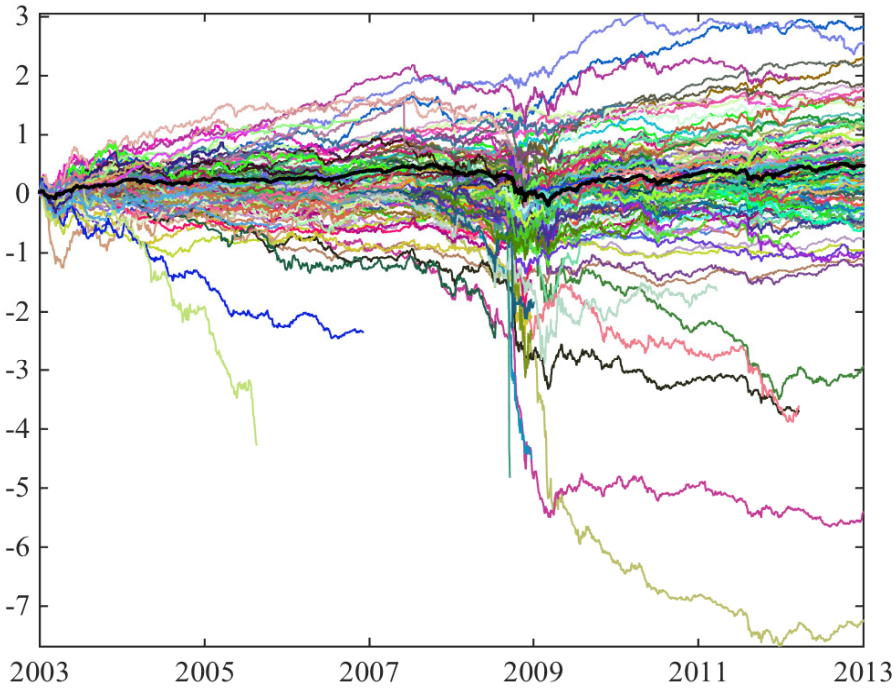

Figure 4: Time Series of the Cumulative Returns of S&P 100 Index Components

Note: In this figure, we plot the time series of cumulative daily open-to-close returns of 158 S&P 100 Index Components from 2003 to 2012. The thick black solid line plots the cumulative daily open-to-close S&P 100 Index returns. All overnight returns are excluded. Companies that exited the index during the sample period, including bankruptcies, are represented by a time series truncated at the time of delisting.

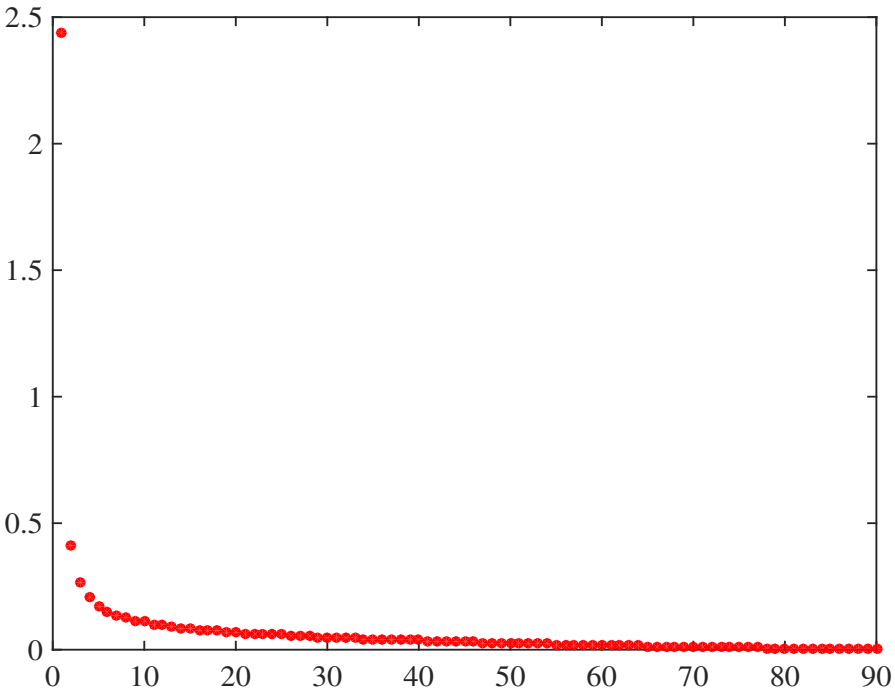

Figure 5: The Scree Graph

Note: In this figure, we report the time series average over the entire sample of the estimated eigenvalues against their order, or "scree graph". These integrated eigenvalues are estimated assuming that all of them are simple.

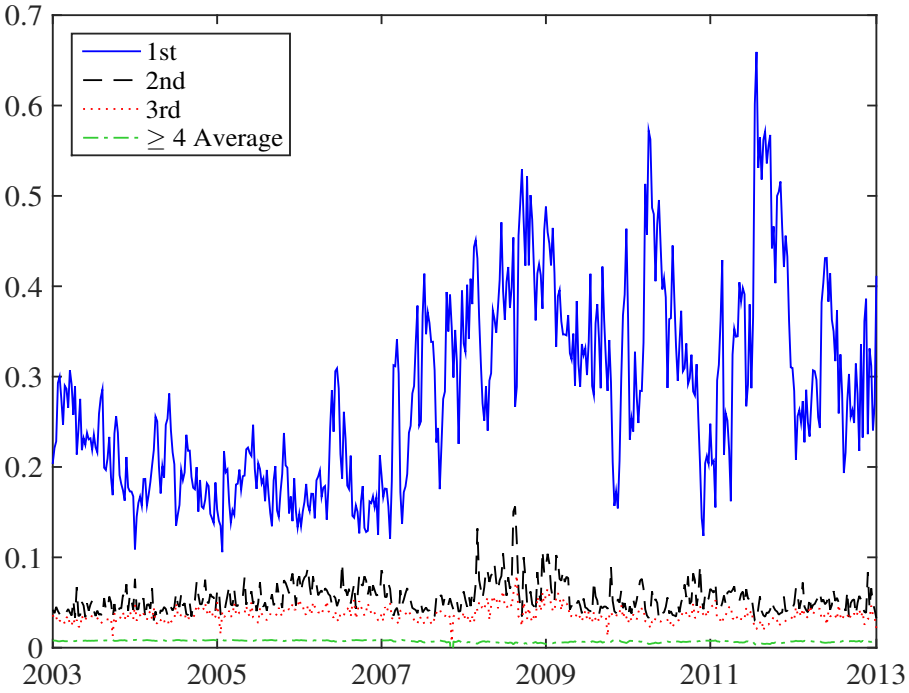

Figure 6: Time Series of Realized Eigenvalues

Note: In this figure, we plot the time series of the three largest realized eigenvalues, representing the percentage variations explained by the corresponding principal components, using weekly 1-min returns of 90 most liquid S&P 100 Index constituents from 2003 to 2012. For comparison, we also plot the average of the remaining 87 eigenvalues.

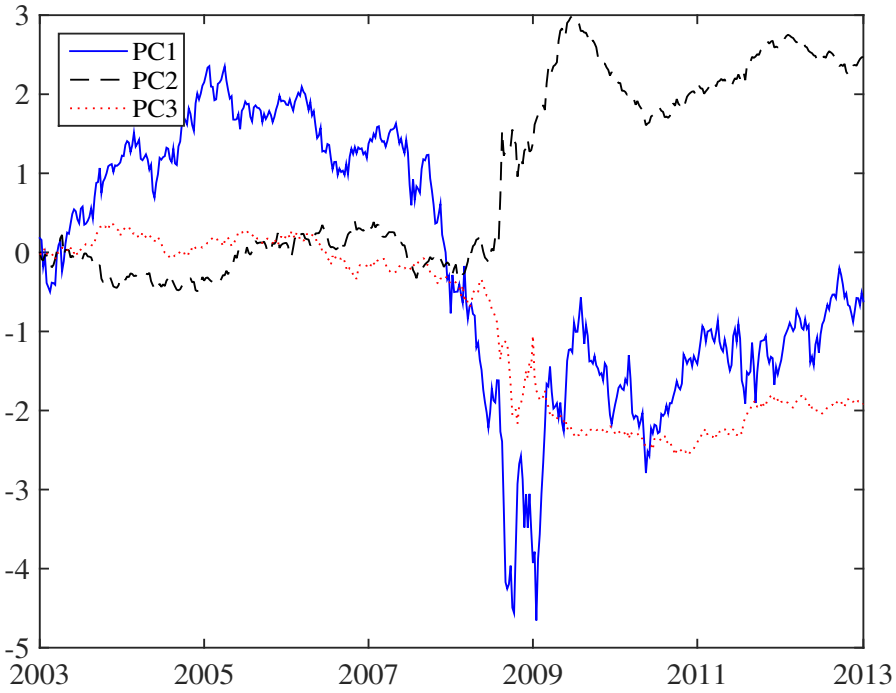

Figure 7: Time Series of Cumulative Realized Principal Components

Note: In this figure, we compare the cumulative returns of the first three principal components, using weekly 1-minute returns of the S&P 100 Index constituents from 2003 to 2012.

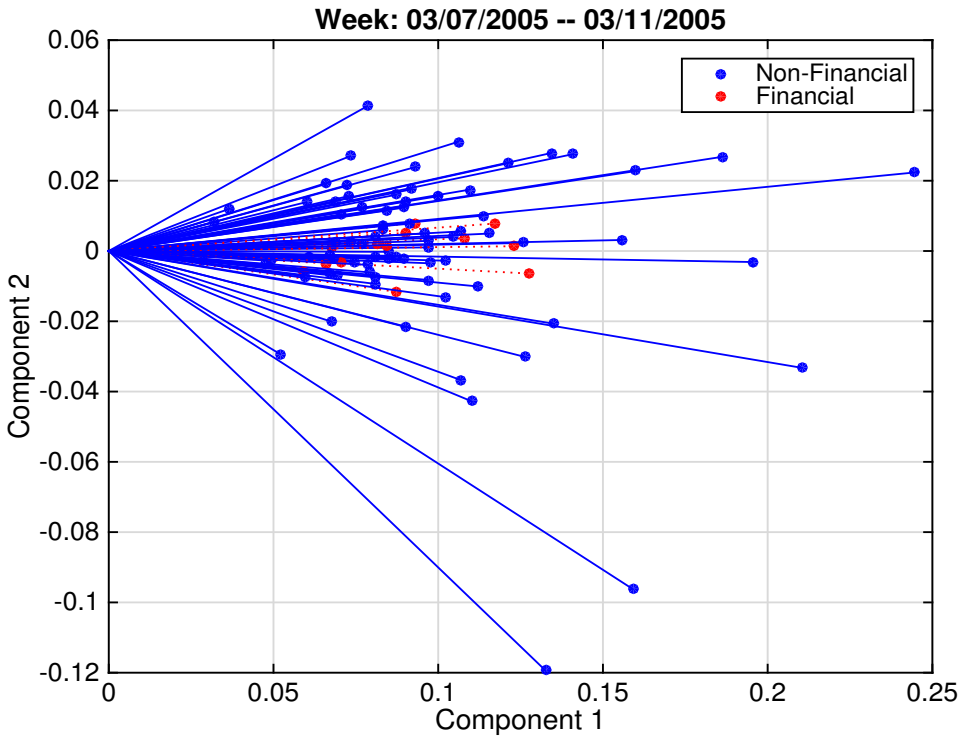

Figure 8: Biplot of the Integrated Eigenvector: Pre-Crisis Week

Note: This biplot shows the integrated eigenvector estimated during the week March 7-11, 2005. Each dot and the vector joining it and the origin represents a firm. The x and y coordinates of a firm denote its average loading on the first and second principal components, respectively. The dots and lines associated with firms in the financial sector are colored in red and dashed, while non-financial firms are colored in blue and drawn as solid lines. The figures shows little difference between the two groups during that week.

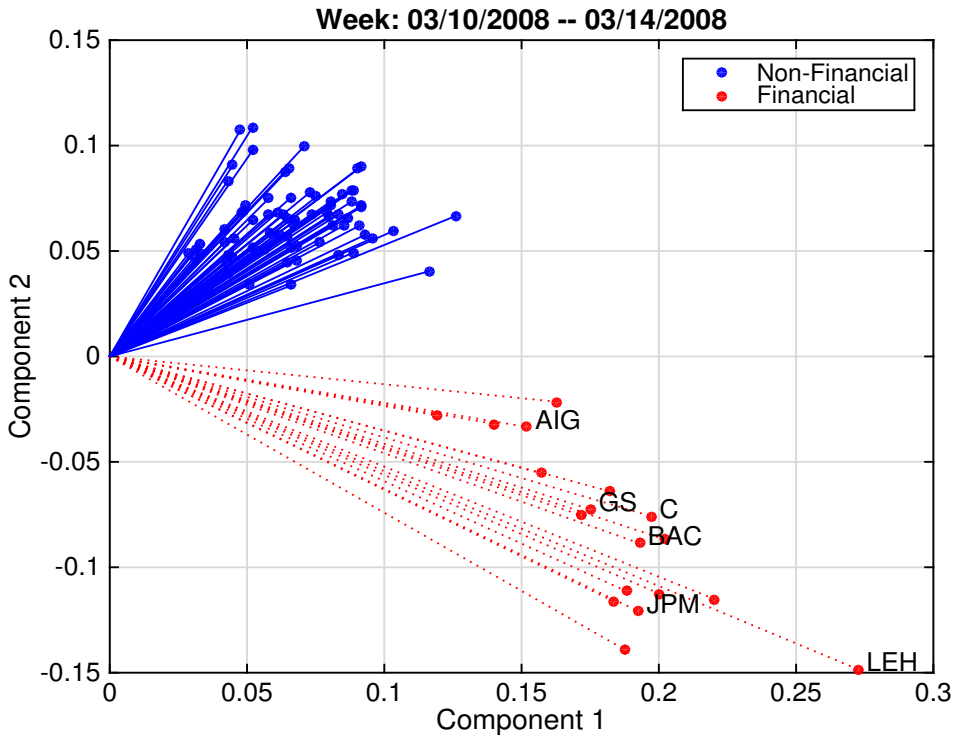

Figure 9: Biplot of the Integrated Eigenvector: Crisis Week

Note: This biplot shows the integrated eigenvector estimated during the week March 10-14, 2008. Each dot and the vector joining it and the origin represents a firm. The x and y coordinates of a firm denote its average loading on the first and second principal components, respectively. The dots and lines associated with firms in the financial sector are colored in red and dashed, with the following singled out: American International Group (AIG), Bank of America (BOA), Citigroup (C), Goldman Sachs (GS), J.P. Morgan Chase (JPM), and Lehman Brothers (LEH). Non-financial firms are colored in blue and drawn as solid lines. The figure shows a sharp distinction between the two groups' dependence on the second principal component during that week.

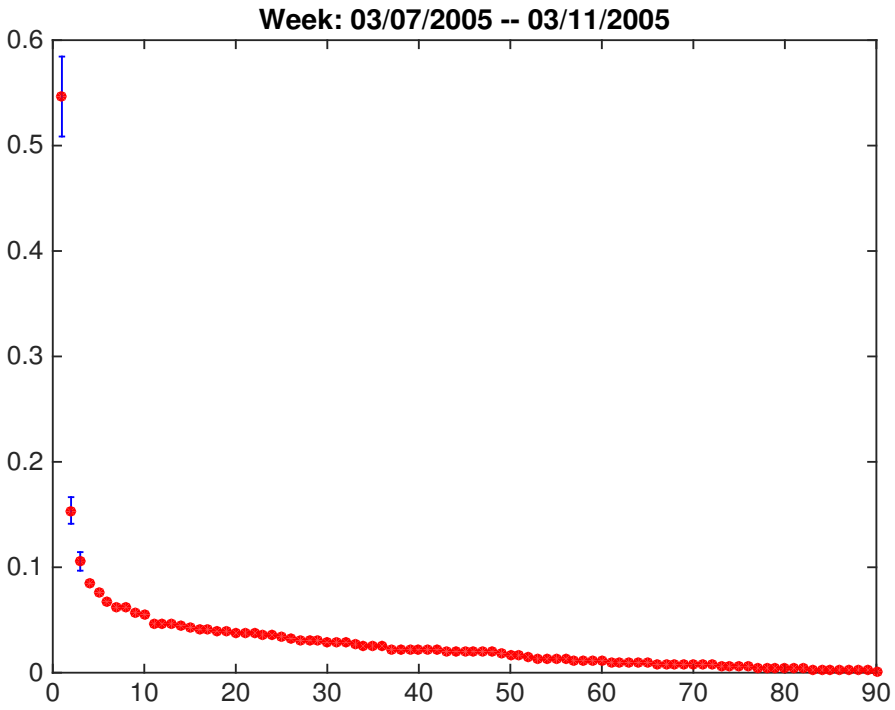

Figure 10: Scree Plot of the Integrated Eigenvalues: Pre-Crisis Week

Note: This figure contains a scree plot of the integrated eigenvalues estimated during the week March 7-11, 2005. The blue solid lines provide 95% confidence interval for the first three eigenvalues computed based on the results of Corollary 1.

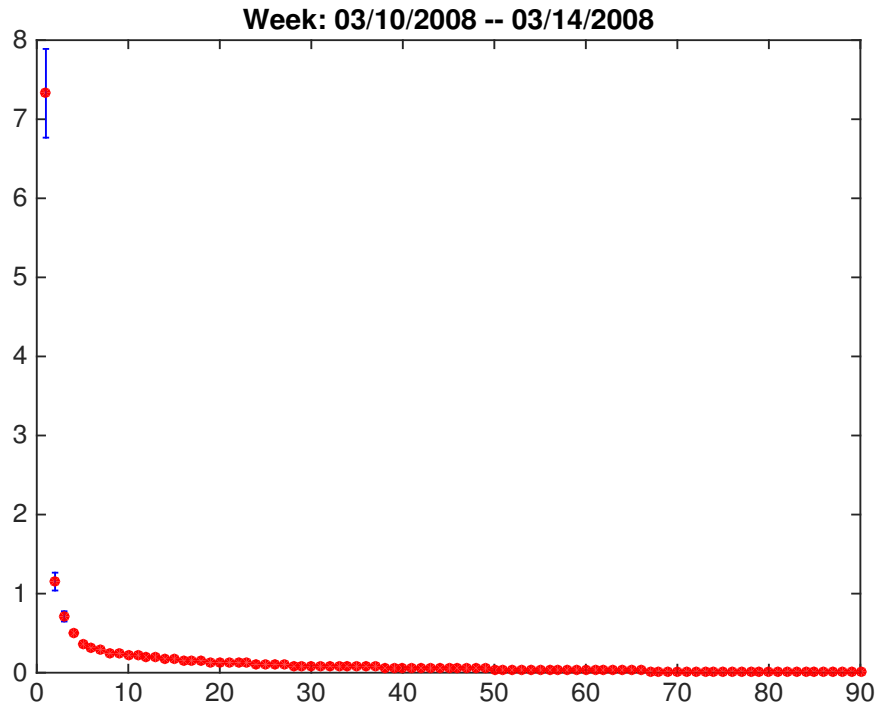

Figure 11: Scree Plot of the Integrated Eigenvalues: Crisis Week

Note: This figure contains a scree plot of the integrated eigenvalues estimated during the week March 10-14, 2008. The blue solid lines provide 95% confidence interval for the first three eigenvalues computed based on the results of Corollary 1.