Testing the Eigenvalue Structure of Spot and Integrated Covariance*

Prosper Dovonon† Abderrahim Taamouti‡ Julian Williams§

January 19, 2021

Abstract

For vector Itô semimartingale dynamics, we derive the asymptotic distributions of likelihood-ratio-type test statistics for the purpose of identifying the eigenvalue structure of both integrated and spot covariance matrices estimated using high-frequency data. Unlike the existing approaches where the cross-section dimension grows to infinity, our tests do not necessarily require large cross-section and thus allow for a wide range of applications. The tests, however, are based on non-standard asymptotic distributions with many nuisance parameters. Another contribution of this paper consists in proposing a bootstrap method to approximate these asymptotic distributions. While standard bootstrap methods focus on sampling point-wise returns, the proposed method replicates features of the asymptotic approximation of the statistics of interest that guarantee its validity. A Monte Carlo simulation study shows that the bootstrap-based test controls size and has power for even moderate size samples.

Keywords: Eigenvalue; Eigenvector; High frequency; Itô semimartingale; Principal components; Likelihood ratio test; bootstrap

JEL classification: C13; C14; C15; C58; G01

1 Introduction

Decomposing a q-dimensional covariance matrix into a set of eigenvalues and eigenvectors, referred to as eigen- or spectral- decomposition, is an important part of a number of statistical methods, notably Principal Component Analysis (PCA) and Factor Analysis (FA). PCA and FA summarize the variation of a high dimensional data set through a smaller number of factors. These factors are designed to capture the maximal feasible fraction of the cross sectional variation of the data set. Depending on the structural assumptions, the factors may be correlated or uncorrelated, and in the latter setting the factors are often referred to as principal components.

Empirical work often raises the question of deciding about the number, say Q, of factors or principal components to retain for data modelling. Such a decision is based on testing the clustering eigenvalues

*The authors thank two anonymous referees, an Associate Editor, and the Editor Torben G. Andersen for several useful comments.

†Department of Economics, Concordia University, 1455 de Maisonneuve Blvd. West, H 1155, Montreal, Quebec, Canada, H3G 1M8. Tel: +1(514)848 2424 ext. 3479. Email: prosper.dovonon@concordia.ca.

‡Department of Economics and Finance, Durham University Business School. Address: Mill Hill Lane, Durham, DH13LB, UK. Tel: +44(0)191 33 45423. Email: abderrahim.taamouti@durham.ac.uk.

§Department of Economics and Finance, Durham University Business School. Address: Mill Hill Lane, Durham, DH13LB, UK. Tel: +44(0)191 33 45301. Email: julian.williams@durham.ac.uk.

of the covariance matrix of interest. Specifically, if $1-\pi$ is a targeted proportion of data dispersion supported by the principal components (typically $\pi = 0.10$ or 0.05), Q can be found by testing whether the ratio of sum of the smallest $q-Q$ eigenvalues by the sum of all eigenvalues is at most as large as π. Moreover, particular structures of the eigenvalues can be used to infer properties of the underlying data generating processes. For instance, the equality of the $q-m$ smallest eigenvalues is a testable implication of a factor representation of the primitive vector process with m factors and uncorrelated idiosyncratic shocks of the same magnitude.

While the covariance matrix is a well-accepted measure of multivariate dispersion of data observed at low frequency, the so-called spot variance matrix at date t (denoted by c_t) and integrated covariance matrix (hereafter IV) play a similar role for financial data observed at high frequency and have established themselves as important components of financial portfolios' risk measures. The eigenvalue structure of c_t or IV is of similar importance for PCA and FA of the *continuous part* of high-frequency vectors of stock prices as does that of covariance for low frequency data. Specifically, the eigenvectors of c_t associated to its largest eigenvalues determine the principal components of the vector price process at that specific date t, and similarly, when volatility is locally constant, the eigenvectors of IV determine the principal components over the estimation interval. Our aim is to provide statistical inference for the eigenvalue structure of c_t and IV estimated using high frequency data.

In this paper, we consider a vector process of stock prices belonging to a class of continuous-time multivariate stochastic processes, namely the class of Itô semimartingale processes and propose a collection of tests for the eigenvalue structure of the associated spot and integrated covariance matrices – c_t and IV, respectively. Three tests are proposed herein. The first test provides statistical inference on the equality of adjacent eigenvalues without specifying their common value whereas the second applies to the case where a common value is set under the null hypothesis. The third test investigates whether a given number Q of the principal components of the continuous part of the price vector process captures at least a given proportion of the total dispersion as measured by c_t or IV.

We construct these tests by first considering the subclass of continuous Lévy processes. This class offers a parametric setting whereby the likelihood ratio can fully be derived for the first two tests while a test statistic for the third one is based on the behaviour under the null hypothesis of the maximum likelihood estimator of the eigenvalues. Under the null hypothesis, the likelihood ratio test statistics are asymptotically distributed as a standard chi-squared distribution with $\frac{1}{2}q_k(q_k+1)$ and $\frac{1}{2}(q_k-1)(q_k+2)$ degrees of freedom, respectively; where q_k is the number of equal eigenvalues of IV (or c_t) in the k-th cluster being tested. The third test has a test statistic that is asymptotically normally distributed. In this configuration of continuous Lévy process, the same tests can be applied to the spot covariance matrix, c_t, with the same asymptotic distribution.

Our results for these three tests are reminiscent to those of Anderson (1963) who proposes a likelihood ratio test statistic for the equality of adjacent eigenvalues of a population covariance matrix when the data is independent and identically distributed (i.i.d.) as Gaussian.[1] Our framework, however, differs from his as our inference results are based on so-called infill asymptotics where the

[1]Further extension of the work of Anderson (1963) to non Gaussian and to dependent data can be found in Waternaux (1976), Davis (1977), Tyler (1981), Muirhead (1982), Eaton and Tyler (1991), Cook and Setodji (2003), and Onatski (2010) to name a few.

process is supposed to be observed over a fixed time interval but more and more frequently. In this case, although observed log-returns from continuous Lévy processes are i.i.d. Gaussian, their common variance changes with sample size.

For the more general class of Itô semimartingales, we use the same test statistics and derive their asymptotic distributions (under the null) which are all now noticeably non-standard with many nuisance parameters, both for c_t and IV. Using consistent estimators of these nuisance parameters, we indicate how approximations of these distributions can be simulated. Nevertheless, the set of nuisance parameters can become quite large as it includes eigenvectors of c_t or IV, spot covariances as well as integrated multivariate quarticities. Clearly, the estimation of so many parameters is costly and these estimates may also distort the finite sample properties of the test by affecting the simulated distribution in a perverse manner. This observation motivates our exploration of the bootstrap as an alternative approach to estimate the asymptotic distributions of our tests.

A further contribution of this paper therefore consists in proposing asymptotically valid bootstrap methods that approximate these asymptotic distributions. The bootstrap procedures that we introduce are simple to implement and do not require the estimation of many nuisance parameters. The main difficulty associated to a bootstrap proposal in this context is that the eigenvalue structure of c_t or IV is notoriously different from that of their respective estimates. Since existing bootstrap methods aim at bootstrapping the sample estimates, it results that the eigenvalue structure that they point to would not reflect that of the population quantities and then leading to invalid tests. To circumvent this problem, we propose valid bootstrap methods that, instead of the test statistics themselves, replicate features of their asymptotic approximations that guarantee first-order asymptotic validity. This approach to bootstrapping is similar to that of Chen and Fang (2019). Even though the inference on the eigenvalue structure of IV and that of c_t are of our primary interest, we also derive results for the eigenvalue structure of the correlation matrix R, with $R_{i,j} = v_{i,j}/\sqrt{v_{i,i}v_{j,j}}$ and $v =$ IV or c_t. R being a smooth function of IV or c_t, the validity of our bootstrap methods extend easily to R.

The literature on the use of PCA and FA with high-frequency data is relatively recent. Aït-Sahalia and Xiu (2019) develop a methodology to conduct PCA with high-frequency data with emphasis on inference in a setup of fixed cross-section dimension. However, their results take as given the clustering structure of the so-called integrated eigenvalues which are then supposed known to the researcher. Our work complement theirs as the proposed tests aim at depicting such structures for c_t and IV.

Aït-Sahalia and Xiu (2017) and Pelger (2019) propose estimators of the number of factors in continuous-time factor models. In the spirit of Bai and Ng (2002), the estimator of Aït-Sahalia and Xiu (2017) minimizes a criterion that involves estimated eigenvalues of IV and a penalty term, whereas Pelger (2019) follows Onatski (2010) and Ahn and Horenstein (2013) by proposing an estimator that maximizes the ratio of perturbed adjacent estimated eigenvalues of IV. In line with the approximate factors theory of Chamberlain and Rothschild (1983), these estimators are all shown to be consistent when the cross-section dimension increases to infinity. Such estimators typically exploit the high dimension to extract the factor structure. Our work makes a direct contribution to this literature, and our proposed tests can be applied to large dimensions of high frequency return data to detect the number of factors needed to pick up a target proportion of the dispersion of the data as measured by

c_t or IV. In addition, factor structures reflected by the equality of the smallest eigenvalues of c_t or IV can be investigated through our testing framework. It is worthwhile to stress that our tests apply to settings with fixed cross-section dimension and are valid even if this is not large. Therefore, we allow for empirical applications with a small number of assets as opposed to the aforementioned methods, for example in extracting factors from the term structure of bond and futures prices.

Other papers in this literature include Chen et al. (2019) who document a number of issues when conducting PCA on high frequency data in the presence of time varying multivariate stochastic volatility and provide a battery of approaches to overcoming the identified difficulties. Todorov and Bollerslev (2010) formulate a two-factor model for financial asset prices that allows them to disentangle and estimate assets' exposure to the diffusive and the jump components of the market systematic risk. And more recently, with an application to market completeness, Jacod and Podolskij (2013) have outlined a test for maximal rank of the spot volatility matrix.

The literature on bootstrap methods for multivariate high-frequency volatility measures includes Dovonon, Gonçalves, and Meddahi (2013) who propose a non-parametric i.i.d. bootstrap to approximate the distribution of the so-called realized beta and realized correlation between assets. Generalizing the work of Hounyo, Gonçalves, and Meddahi (2017) to multivariate settings, Hounyo (2017) proposes a wild blocks of blocks bootstrap for estimating the distribution of various estimators of the integrated covariance matrix. Note that, as already mentioned, a naive application of the bootstrap method of Hounyo (2017) does not lead to a valid test of the eigenvalue structure of IV because of the mismatch with the eigenvalue structure of the estimated IV. Nevertheless, we show that the first-order asymptotic approximation of our test statistics can be validly bootstrapped. Variants of the blocks of blocks bootstrap method have been proposed that consistently estimate the asymptotic distribution of our test statistics related to IV. This result is further extended to the spot variance matrix. This extension, however, is not straightforward as it requires additional laws of large numbers for functions of successive local returns that we provide and that may be of additional independent interest.

The finite sample properties of the results obtained have been investigated by a Monte Carlo simulation study in which several data generating processes have been considered as well as different sampling frequencies and small and large cross-section dimensions. The results reveal that the bootstrap test has a very good size and power performance. We also report the rejection rates based on the standard asymptotic chi-squared test, which is valid only if the underlying process is continuous Lévy. It turns out that the latter systematically over rejects the null except, as expected, in the case of *non*-Lévy dynamics. We outline a case study using high frequency data for the cross section of the one hundred most actively traded constituents of the S&P 500 to build candidate factors and test them against the cross section of all available listed US equities in the CRSP stock universe.

The structure of this paper is as follows. Section 2 presents the theoretical framework and outlines the relevant existing results. In Section 3, the likelihood ratio test statistics are derived for the test of equality of eigenvalues in the case of continuous Lévy process and their asymptotic distributions provided under general dynamics. This section also features the test of proportion of volatility supported by the main principal components. We conclude the section with an analysis of eigenvalue structure of the correlation matrix. Section 4 presents our bootstrap methodologies and establishes their va-

lidity. Section 5 extends the results to the spot variance matrix. It develops both the asymptotic and bootstrap-based inference for the eigenvalue structure of spot variance matrix estimated using high-frequency data. The Monte Carlo experiments are reported in Section 6. In Section 7 we present an illustration of the analysis on five minute data for the 100 most actively traded stocks from the S&P 500. Concluding remarks on the scope of the tests are provided in Section 8.

2 Set-up and existing results

Let X be a q-dimensional Itô semimartingale defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, a filtered probability space, with Grigelionis decomposition:

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

where W is a q-dimensional Wiener process, μ is a Poisson measure with compensator $\nu(dt, dz) = dt \otimes \lambda(dz)$, with λ the Lebesgue measure on $\mathbb{R}^q$; δ is a real function on $\Omega \times \mathbb{R}_+ \times \mathbb{R}^q$, and σ_s is the volatility process. We let $c_s = \sigma_s \sigma_s^\top$ denote the spot variance matrix. We assume that X satisfies the following assumption for some $r \in [0, 2]$:

Assumption (H-r). *b_t is locally bounded and σ_t is càdlàg, and $\|\delta(\omega, t, z)\| \wedge 1 \leq \gamma_n(z)$ for all (ω, t, z) with $t \leq \tau_n(\omega)$, where (τ_n) is a localizing sequence of stopping times and each function γ_n satisfies $\int \gamma_n(z)^r \lambda(dz) < \infty$.*

The process X represents the vector of log-prices of q assets that we assume are observed at regular time interval Δ_n over a time period $[0, T]$. The main objects of interest in this paper are the spot variance matrix at date t, c_t and the integrated covariance matrix of X over the time interval $[0, T]$

$$IV_T = \int_0^T c_s ds,$$

which corresponds to the quadratic variation (QV) of the continuous part X^c of the process X at time T, that is $IV_T = [X^c, X^c]_T$, where $X_t^c = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s$. Let $\Delta_n^i X = X_{i\Delta_n} - X_{(i-1)\Delta_n}$ be the log-return process over $((i-1)\Delta_n, i\Delta_n]$ for $i = 1, \ldots, n \simeq \lfloor T/\Delta_n \rfloor$, where $\lfloor x \rfloor$ is the largest integer smaller or equal to x. The integrated covariance matrix IV_T is estimated by

$$\widehat{IV}^n = \sum_{i=1}^{\lfloor T/\Delta_n \rfloor} (\Delta_n^i X)(\Delta_n^i X)' 1_{\{\|\Delta_n^i X\| \leq \alpha \Delta_n^{\varpi}\}}, \tag{2}$$

for some $\alpha > 0$ and $\varpi \in (0, \frac{1}{2})$. Under mild conditions, it is shown that $\widehat{IV}^n$ is a consistent estimator of IV_T where the setting of asymptotic analysis is the so-called infill asymptotics in which prices are supposed to be sampled more and more often over the same time interval $[0, T]$, i.e. $\Delta_n \to 0$. The spot variance matrix at t is estimated by a local version of $\widehat{IV}^n$. Let $i \in \mathbb{N}$ be such that $t \in ((i-1)\Delta_n, i\Delta_n]$

and k_n a sequence of integers such that $k_n \to \infty$ and $k_n\Delta_n \to 0$ as $n \to \infty$. Define:

$$\hat{c}(t, k_n) = \frac{1}{k_n\Delta_n} \sum_{m=0}^{k_n-1} (\Delta^n_{i+1+m}X)(\Delta^n_{i+1+m}X)' 1_{\{\|\Delta^n_{i+1+m}X\| \leq \alpha\Delta_n^{\varpi}\}}. \tag{3}$$

Jacod and Protter (2012, Th. 9.3.2) establish that $\hat{c}(t, k_n)$ is consistent for c_t as $n \to \infty$.

Our interest resides in the asymptotic distribution of the characteristic roots and vectors of $\hat{c}(t, k_n)$ and $\widehat{IV}^n$ and tests for equality of all or some roots of their respective limits. These tests are useful to carry out inference on the importance of principal components of c_t as determined by its eigenvectors. They are also useful to test for some specific factor decomposition of c_t. Even though the principal component interpretation of the eigenvectors of IV_T is not meaningful when c_t is not constant over $t \in [0, T]$, we make these tests available for IV_T as they may be incidentally of interest for some statistical analysis. We next introduce some existing results that provide the groundwork for our main contributions that appear in the next section.

The asymptotic behaviour of $\widehat{IV}^n$ is well-known [(see e.g. Aït-Sahalia and Jacod, 2014, Th. A.16)]. If Assumption (H-r) holds for some $r \in [0, 1)$ and the truncation level $\varpi \in \left[\frac{1}{2(2-r)}, \frac{1}{2}\right)$, then

$$\frac{1}{\sqrt{\Delta_n}} \left(\widehat{IV}^n - IV_T\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}_T, \tag{4}$$

where $\mathcal{W}_T$ is a random vector defined on an extension of the original probability space and conditionally on $\mathcal{F}$, is Gaussian with conditional mean 0 and conditional variance covariance given by

$$\mathbb{E}\left(\mathcal{W}_T^{uv}\mathcal{W}_T^{kl}\middle|\mathcal{F}\right) = \int_0^T \left(c_s^{uk}c_s^{vl} + c_s^{ul}c_s^{vk}\right) ds, \tag{5}$$

with $u, v, k, l = 1, \ldots, q$. '$\mathcal{L}$-$s$' stands for convergence stable in distribution. We refer to Aït-Sahalia and Jacod (2014, Section 3.2) for further details on this mode of convergence.

The asymptotic normality of $\hat{c}(t, k_n)$ is established by Jacod and Protter (2012, Th. 13.3.3). Under A.1 in Appendix, we have:

$$\sqrt{k_n}(\hat{c}(t, k_n) - c_t) \xrightarrow{\mathcal{L}-s} Z_t, \tag{6}$$

where Z_t is a random vector defined on an extension of the original probability space and conditionally on $\mathcal{F}$, is Gaussian with conditional mean 0 and conditional variance covariance given by

$$E\left(Z_t^{uv}Z_t^{kl}\middle|\mathcal{F}\right) = c_t^{uk}c_t^{vl} + c_t^{ul}c_t^{vk},$$

with $u, v, k, l = 1, \ldots, q$.

Let $\mathcal{M}_q$ denote the Euclidean space of all $q \times q$ real-valued symmetric matrices, and $\mathcal{M}_q^+$ ($\mathcal{M}_q^{++}$) the subset of all positive semidefinite (definite) elements of $\mathcal{M}_q$. Most of the quantities of interest in this paper are continuously differentiable functions of the integrated variance matrix or the spot variance matrix. Thus, using the delta method, the large sample behaviour of their estimators can be based on (4) or (6). For this, let φ be a generic function defined on $\mathcal{M}_q^+$ with value in $\mathbb{R}^r$. Assuming

that φ is continuously differentiable on the support of $\theta_0 = IV_T$ (or c_t), we have:

$$\frac{1}{\sqrt{\Delta_n}}\left(\varphi\left(\hat{\theta}\right) - \varphi\left(\theta_0\right)\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}^{\varphi}, \tag{7}$$

with $\hat{\theta} = \widehat{IV}^n$ (or $\hat{c}(t, k_n)$) and where, similarly to $\mathcal{W}_T$ (or Z_t), $\mathcal{W}^{\varphi}$ is defined on an extension of the original probability space and, conditionally on $\mathcal{F}$ is centered Gaussian with conditional covariance matrix given by:

$$\mathbb{E}\left(\mathcal{W}^{\varphi}\mathcal{W}^{\varphi'}|\mathcal{F}\right) = \sum_{u,v,k,l=1}^{q}\left(\frac{\partial\varphi(M)}{\partial M_{uv}}\bigg|_{M=\theta_0}\right)V(u,v,k,l)\left(\frac{\partial\varphi(M)}{\partial M_{kl}}\bigg|_{M=\theta_0}\right)', \tag{8}$$

where $V(u,v,k,l) = \int_0^T \left(c_s^{uk}c_s^{vl} + c_s^{ul}c_s^{vk}\right) ds$ for $\theta_0 = IV_T$ and $V(u,v,k,l) = c_t^{uk}c_t^{vl} + c_t^{ul}c_t^{vk}$ for $\theta_0 = c_t$.

Let λ denote the eigenvalue function defined on $\mathcal{M}_q^+$, with nonincreasing elements and value in a suitable subset of $\mathbb{R}^q$. Let $A \in \mathcal{M}_q^+$ with r clusters $\mathscr{L}_k$ (for $k = 1, \ldots, r$) of q_k-repeated eigenvalues with common values λ_k, where $\mathscr{L}_k$ is the collection of the ranks of eigenvalues (sorted from largest to smallest) of A equal to λ_k. The components of $\lambda(A) \equiv (\delta_i)_{1\leq i\leq q}$ have the structure:

$$\begin{array}{ccccccccc}
\delta_1 & = & \delta_2 & = & \cdots & = & \delta_{q_1} & = & \lambda_1, \\
\delta_{q_1+1} & = & \delta_{q_1+2} & = & \cdots & = & \delta_{q_1+q_2} & = & \lambda_2, \\
\vdots & & & & & & & & \\
\delta_{q-q_r+1} & = & \delta_{q-q_r+2} & = & \cdots & = & \delta_q & = & \lambda_r,
\end{array} \tag{9}$$

with $\lambda_1 > \lambda_2 > \cdots > \lambda_r$. The eigenvalue function $\lambda(\cdot)$ is locally Lipschitz continuous on $\mathcal{M}_q^+$ and differentiable only at points A of $\mathcal{M}_q^+$ with no repeated eigenvalues, i.e. $r = q$ [see Tao, 2012]. Nevertheless, some relevant functions of $\lambda(\cdot)$ are differentiable. Consider again $A \in \mathcal{M}_q^+$ with eigenvalue structure given by (9). It is known - see e.g. Chu (1990) and Corollary 3.11 of Hiriart-Urruty and Ye (1995) - that there exists a neighborhood of A on which, for $k = 1, \ldots, r$, the functions:

$$\varphi_k : M \in \mathcal{M}_q^+ \mapsto \varphi_k(M) = \sum_{i=1}^{q_k} \lambda_{\iota_{k-1}+i}(M),$$

with $\iota_0 = 0$ and $\iota_k = \sum_{i=1}^k q_i$, are continuously differentiable. Note that φ_k is the sum of eigenvalues with ranks in the same cluster $\mathscr{L}_k$. Assuming that $\theta_0 = IV_T$ or c_t has an eigenvalue structure given by (9), Aït-Sahalia and Xiu (2019) consider the function:

$$\varphi^{\lambda}(M) = \left(\frac{1}{q_1}\varphi_1(M), \frac{1}{q_2}\varphi_2(M), \ldots, \frac{1}{q_r}\varphi_r(M)\right)' \tag{10}$$

and establish that:

$$\frac{1}{\sqrt{\Delta_n}}\left(\varphi^{\lambda}\left(\hat{\theta}\right) - \varphi^{\lambda}(\theta_0)\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}^{\varphi^{\lambda}}, \tag{11}$$

where $\mathcal{W}^{\varphi^\lambda}$ is defined similarly to $\mathcal{W}^{\varphi}$ with

$$\left.\frac{\partial \varphi_k^\lambda(M)}{\partial M}\right|_{M=\theta_0} = \frac{1}{q_k} U_{T,q_{k-1}+1:q_k} U'_{T,q_{k-1}+1:q_k},\ k=1,\ldots,r,$$

with U_T being any orthogonal matrix such that $U_T'\theta_0 U_T = diag(\lambda_1(\theta_0),\ldots,\lambda_q(\theta_0))$, where, for $u \in \mathbb{R}^q$, $diag(u)$ is the diagonal matrix of size q with u as main diagonal.

If for some $i \in \{1,\ldots,q\}$, $\lambda_i(\theta_0)$ is a simple eigenvalue, the function $\gamma_i(A)$ returning the i-th normalized eigenvector of A defines, up to the sign[2], a differentiable function in a neighborhood of θ_0 (see Magnus, 1985, Th. 1). Aït-Sahalia and Xiu (2019) show that:

$$\frac{1}{\sqrt{\Delta_n}}\left(\gamma_i\left(\hat{\theta}\right) - \gamma_i(\theta_0)\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}^{\gamma_i}, \tag{12}$$

where $\mathcal{W}^{\gamma_i}$ is defined similarly to $\mathcal{W}^{\varphi}$ with

$$\left.\frac{\partial \gamma_i(M)}{\partial (vec[M])'}\right|_{M=\theta_0} = \gamma_i(\theta_0)' \otimes \left[\lambda_i(\theta_0) I_q - \theta_0\right]^+,$$

where A^+ is the Moore-Penrose inverse of A and $vec[M]$ is the standard vectorizing operator that transforms the matrix M into a vector by stacking its columns.

3 Testing the eigenvalue structure of Integrated Covariance

The results in the previous section take the structure of the eigenvalues of $\theta_0 = IV_T$ or, c_t - such as the one in (9) - as known to the researcher. In particular, the rank and multiplicity of the eigenvalues. In this case, confidence intervals can be built for eigenvalues or their average along the framework of Aït-Sahalia and Xiu (2019) as recalled. These results can also be used to test some restriction on the true eigenvalues, and such a test would be asymptotically valid if the maintained eigenvalue structure is correct. However, this structure is not known in general. Our aim in this section is to introduce a test for this purpose. Our focus throughout is on IV_T while tests related to spot variance matrix c_t are shown in Section 5. It is a well-known fact that - unlike for the spot variance matrix - the eigenvectors of the integrated volatility (barring the case of constant volatility) have no clear connection to the principal components of the price process. Nevertheless, we first analyze IV_T since this allows us to set up the useful statistics that we subsequently apply to c_t. We further propose a test to investigate whether a given set of factors "explains" at least a certain proportion of integrated variance in the continuous part of the process X (vectors of log-prices). An extension to tests for eigenvalue structure of correlation matrices is also provided.

To build a test for eigenvalue structure, we first consider a simpler version of the stochastic process X in (1). Namely, we assume that X is a continuous Lévy process, that is $\delta \equiv 0$ and $b_s \equiv b$ and $\sigma_s \equiv \sigma$ are constant. This gives rise to a parametric model in which $\Delta_n^i X$ (for $i = 1,\ldots,n$) are independent and identically distributed $N(\Delta_n b, \Delta_n c)$, with $c = \sigma\sigma'$ and $IV_T = Tc = T\sigma\sigma'$.

[2]If the first nonzero element of $\gamma_i(\theta_0)$ is its hth entry, then restricting the hth entry of $\gamma_i(\cdot)$ to be nonnegative makes this a well-defined function in a neighborhood of θ_0.

By definition, the eigenvalues of IV_T are then T times those of c whereas the corresponding eigenvectors are the same. Even though the variance of $\Delta_n^i X$ tends to zero with the sample size, the fact that they are independent and normally distributed allows us to draw from the work of Anderson (1963) to build the aforementioned tests. In the case of more complex dynamics for X than continuous Lévy, the same test statistics will be used and their asymptotic distributions will be derived. These asymptotic distributions are typically untractable as we shall see, which motivates the bootstrap approximations that will be proposed in the next section.

Under the assumption of continuous Lévy dynamics, the likelihood function of the model in terms of b and c is given by

$$\mathcal{L}(b,c) = (2\pi\Delta_n)^{-\frac{qn}{2}} |c|^{-\frac{n}{2}} \exp\left(-\frac{1}{2\Delta_n} tr\left(c^{-1} \sum_{i=1}^{n} (\Delta_n^i X - \Delta_n b)(\Delta_n^i X - \Delta_n b)' \right) \right),$$

where $|c|$ is the determinant of c and tr is the usual trace operator. It is not hard to see that the maximum likelihood estimators of b, c and IV_T are:

$$\tilde{b} = \frac{1}{n\Delta_n} \sum_{i=1}^{n} \Delta_n^i X, \quad \tilde{c} = \frac{1}{n\Delta_n} \sum_{i=1}^{n} \left(\Delta_n^i X - \Delta_n \tilde{b} \right) \left(\Delta_n^i X - \Delta_n \tilde{b} \right)' \quad \text{and} \quad \widetilde{IV}^n = T\tilde{c}, \tag{13}$$

with $n = \lfloor T/\Delta_n \rfloor$. We will make throughout the standard simplifying assumption that T/Δ_n is an integer. Note that $\tilde{b}$ is an unbiased estimator of b while $\tilde{c}$ is a consistent estimator of c. The log-likelihood of this model can also be expressed in terms of the eigenvalues of Tc, i.e. IV_T with the restriction that the latter has the eigenvalue structure in (9). The log-likelihood maximized in the direction of b is

$$-\frac{qn}{2} \log(2\pi\Delta_n) - \frac{n}{2} \log|c| - \frac{1}{2\Delta_n} tr(n\Delta_n c^{-1} \tilde{c}).$$

Hence, up to a constant independent of the model parameters, the log-likelihood is equal to

$$-\frac{n}{2} \log|IV_T| - \frac{n}{2} tr\left((IV_T)^{-1} \widetilde{IV}^n \right).$$

This is a similar expression to that of Equation (3.2) of Anderson (1963) and by the same arguments as his leading to his Equation (3.5), we can claim that, up to a constant (in model parameters) term, the log-likelihood of the model in terms of eigenvalues of IV_T is given by

$$\log \mathcal{L}(\lambda_1, \ldots, \lambda_r) = cst - \frac{n}{2} \sum_{k=1}^{r} q_k \log \lambda_k - \frac{n}{2} \sum_{k=1}^{r} \sum_{i \in \mathscr{L}_k} \frac{\tilde{d}_i}{\lambda_k}, \tag{14}$$

where $\tilde{d} = \lambda\left(\widetilde{IV}^n \right)$ is the vector of eigenvalues of $\widetilde{IV}^n$. We have the following result:

Proposition 3.1. *Let X be a continuous Lévy process.*

(a) If the characteristic roots of IV_T are $\lambda_1 > \cdots > \lambda_r > 0$ with multiplicities $q_1, \ldots, q_r$, respectively, the maximum likelihood estimate of λ_k is: $\hat{\lambda}_k = \frac{1}{q_k} \sum_{i \in \mathscr{L}_k} \tilde{d}_i$; $(k = 1, \ldots, r)$, where $\mathscr{L}_k = \{q_1 + \cdots + q_{k-1} + 1, \ldots, q_1 + \cdots + q_k\}$.

(b) The likelihood ratio criterion for testing the roots of IV_T with rank indexes in $\mathscr{L}_k$:

$$H_0: \quad \delta_{q_1+\cdots+q_{k-1}+1} = \cdots = \delta_{q_1+\cdots+q_k} = \lambda_k,$$

where λ_k is unknown, i.e. the kth row in the eigenvalue structure (9), is given by

$$\tilde{\ell}_k = \left(\prod_{i \in \mathscr{L}_k} \tilde{d}_i \Big/ \left(q_k^{-1} \sum_{j \in \mathscr{L}_k} \tilde{d}_j \right)^{q_k} \right)^{\frac{n}{2}}. \tag{15}$$

(c) The likelihood ratio test statistic for H_0 is $\widetilde{LR}_k = -2 \log \tilde{\ell}_k$ and is asymptotically distributed as a $\chi^2_{\frac{1}{2}(q_k-1)(q_k+2)}$. If $\delta_{q_1+\cdots+q_{k-1}+1} \neq \delta_{q_1+\cdots+q_k}$, then $\widetilde{LR}_k \to \infty$, in probability.

(d) The likelihood ratio criterion for testing: $H_0(\lambda): \quad \delta_{q_1+\cdots+q_{k-1}+1} = \cdots = \delta_{q_1+\cdots+q_k} = \lambda$, with λ specified is given by

$$\tilde{\ell}_{k,\lambda} = \left(\frac{\prod_{i \in \mathscr{L}_k} \tilde{d}_i}{\lambda^{q_k}} \right)^{\frac{n}{2}} \exp\left(-\frac{n}{2} \left[\sum_{i \in \mathscr{L}_k} \frac{\tilde{d}_i}{\lambda} - q_k \right] \right) \tag{16}$$

and, under the null, the likelihood ratio test statistic $\widetilde{LR}_{k,\lambda} = -2 \log \tilde{\ell}_{k,\lambda}$ is asymptotically distributed as a $\chi^2_{\frac{1}{2}q_k(q_k+1)}$. Under the alternative, $\widetilde{LR}_{k,\lambda} \to \infty$, in probability.

The proof of Proposition 3.1 is relegated to the appendix and is an adaptation of the result of Anderson (1963) to the infill asymptotics framework that we consider with $\Delta_n \to 0$. Interestingly, the chi-squared asymptotic distributions obtained for the likelihood ratio test statistics in Parts (c) and (d) under the null hypothesis are the same as those derived by Anderson (1963) for the case where the sample is assumed to be independent and identically normally distributed with fixed variance. To give an intuition of the obtained result, let Γ be the matrix of normalized eigenvectors of IV_T:

$$\Gamma' IV_T \Gamma = \mathbf{\Delta} \quad \text{and} \quad \Gamma'\Gamma = I_q, \tag{17}$$

where $\mathbf{\Delta}$ is the diagonal matrix containing $\delta_1 \geq \cdots \geq \delta_q > 0$, the eigenvalues of IV_T satisfying the structure in (9). The asymptotic distribution of $\widetilde{LR}_k$ and $\widetilde{LR}_{k,\lambda}$ in Parts (c) and (d) of Proposition 3.1 are deduced from the asymptotic distribution, say U, of

$$\widetilde{U} = \frac{1}{\sqrt{\Delta_n}} \left(\Gamma' \widetilde{IV}^n \Gamma - \mathbf{\Delta} \right)$$

as we show that

$$\widetilde{LR}_k = \frac{T}{2\lambda_k^2} \left(2 \sum_{i<j; i,j \in \mathscr{L}_k} \tilde{u}_{ij}^2 + \sum_{i \in \mathscr{L}_k} \tilde{u}_{ii}^2 - \frac{1}{q_k} \left(\sum_{i \in \mathscr{L}_k} \tilde{u}_{ii} \right)^2 \right) + o_P(1),$$

and

$$\widetilde{LR}_{k,\lambda} = \frac{T}{2\lambda^2} \left(\sum_{i,j \in \mathscr{L}_k} \tilde{u}_{ij}^2 \right)^2 + o_P(1),$$

where $\tilde{u}_{ij}$ are entries of $\widetilde{U}$. We also show that the limit distribution U of $\tilde{U}$ has its entries u_{ij} that are such that $u_{ij} = u_{ji}$, and $\{u_{ij}, j \leq i\}$ are pairwise independent with $u_{ii} \sim N(0, \lambda_k^2/T)$ and $u_{ij} \sim N(0, 2\lambda_k^2/T)$, for $i < j$ which yields the claimed distributions. Hence, likelihood ratio test statistics obtained using eigenvalues of any estimator of IV_T that is asymptotically equivalent to $\widetilde{IV}^n$ would be asymptotically equivalent to $\widetilde{LR}_k$ and $\widetilde{LR}_{k,\lambda}$, respectively. In particular, as established in an appendix, we have:

$$\frac{1}{\sqrt{\Delta_n}}\left(\widetilde{IV}^n - \overline{IV}^n\right) = o_P(1) \quad \text{and} \quad \frac{1}{\sqrt{\Delta_n}}\left(\widetilde{IV}^n - \widehat{IV}^n\right) = o_P(1), \tag{18}$$

where $\overline{IV}^n = \sum_{i=1}^n (\Delta_n^i X)(\Delta_n^i X)'$ and $\widehat{IV}^n$ is given by (2). This ensures that $\frac{1}{\sqrt{\Delta_n}}\left(\Gamma'\overline{IV}^n\Gamma - \mathbf{\Delta}\right)$ and $\frac{1}{\sqrt{\Delta_n}}\left(\Gamma'\widetilde{IV}^n\Gamma - \mathbf{\Delta}\right)$ are asymptotically equivalent to $\widetilde{U}$ implying that likelihood ratio test statistics using related eigenvalues have the same asymptotic distribution as in (c) and (d). The last statements in (c) and (d) emphasize the consistency of the respective tests.

It is worth mentioning that several lines (eigenvalues clusters) of (9) can jointly be tested. The corresponding likelihood ratio criterion for such a joint hypothesis is simply the product of likelihood ratio criterions $\tilde{\ell}_k$'s over relevant values of k and the asymptotic distribution of the resulting likelihood ratio test statistic is chi-squared with degrees of freedom equal to the sum over the relevant k's of $\frac{1}{2}(q_k - 1)(q_k + 2)$. Several lines can also be tested likewise if λ_k's are specified for each of them.

Remark 1. *Note that the researcher may be interested in testing whether two adjacent eigenvalues (say the third and fourth largest) are equal. Proposition 3.1 provides asymptotically exact tests in this context only if, under the null, the second and fifth (if available) largest eigenvalues are different from the third and fourth. One does not expect the researcher to be aware of this specific structure which may not even hold.*

In general, if one is testing the equality of $p < q_k$ eigenvalues with rank indices in $\mathscr{L}_k^p \subsetneq \mathscr{L}_k$, letting $\widetilde{LR}_k(p)$ be the test statistic for this null hypothesis, the test would be asymptotically correct only if

$$\lim_n P\left(\widetilde{LR}_k(p) > \chi^2_{\frac{1}{2}(p-1)(p+2),1-\alpha}\right) \leq \alpha, \tag{19}$$

for any relevant nominal level α; where $\chi^2_{\frac{1}{2}(p-1)(p+2),1-\alpha}$ is the critical value of the test. Our Simulation results in Section 5 indicate that these tests are conservative, thereby pointing towards the validity of (19)*. We save a formal proof of this result for future research*[3] *since this may be involved.*

Remark 2. *One may want to investigate whether a certain cluster of eigenvalue is a singleton. That is $q_k = 1$ for a given $k = 1, \ldots, r$. This can be done by testing $H_0 : q_k \neq 1$ against $H_1 : q_k = 1$. If $1 < k < r$, $H_0 \equiv H_{01} \vee H_{02}$ with $H_{01} : \delta_{q_1+\cdots+q_k-1} = \delta_{q_1+\cdots+q_k}$ and $H_{02} : \delta_{q_1+\cdots+q_k} = \delta_{q_1+\cdots+q_k+1}$. Note that $H_0 \equiv H_{01}$ if $k = r$ and $H_0 \equiv H_{02}$ if $k = 1$. In case $1 < k < r$, to test H_0 at a level α, it suffices to test both H_{01} and H_{02} at the level $\alpha/2$. Rejection of H_{01} and H_{02} amounts to rejection of H_0.*

We now turn our attention to more general dynamics of the process X. We assume that X has

[3] In fact, we can establish that $\widetilde{LR}_k(p) < \widetilde{LR}_k$ except for $(\tilde{d}_{q_1+\cdots+q_{k-1}+1}, \ldots, \tilde{d}_{q_1+\cdots+q_k})$ lying in a $\mathbb{R}^{q_k}$-subset of Lebesgue measure 0. To claim (19), it would be sufficient to show that the gap $\widetilde{LR}_k - \widetilde{LR}_k(p)$ is at least as large as $\chi^2_{\frac{1}{2}(q_k-1)(q_k+2),1-\alpha} - \chi^2_{\frac{1}{2}(p-1)(p+2),1-\alpha}$ with large probability.

the Itô semimartingale representation in (1), and we aim to use the likelihood ratio test settings in Proposition 3.1 to carry out inference about the eigenvalue structure of the quadratic variation over the time interval $[0,T]$ of the continuous part of X, that is IV_T. As already mentioned, IV_T is consistently estimated by $\widehat{IV}^n$ in (2), which is the sum of outer product of returns after removing jumps by truncation.

Let $d = \lambda\left(\widehat{IV}^n\right)$ be the estimator of $\delta = \lambda\left(IV_T\right)$, the eigenvalues of IV_T. Let ℓ_k and $\ell_{k,\lambda}$ be the same as $\tilde{\ell}_k$ and $\tilde{\ell}_{k,\lambda}$ in (15) and (16), respectively, with $\tilde{d}$ replaced by d, and let $LR_k = -2\log\ell_k$ and $LR_{k,\lambda} = -2\log\ell_{k,\lambda}$.

We next derive the asymptotic distributions of LR_k and $LR_{k,\lambda}$. Note that, the representation of X being only partially parametric implies that ℓ_k and $\ell_{k,\lambda}$ cannot in general enjoy the interpretation of likelihood ratio criteria. Nevertheless, these test statistics can be relied upon once we are able to characterize their asymptotic distributions under the null hypothesis.

To obtain the asymptotic distribution of LR_k and $LR_{k,\lambda}$, let Γ be the orthogonal matrix of normalized eigenvectors of IV_T and $\mathbf{\Delta}$ be the diagonal matrix of eigenvalues of IV_T, respectively [see Equation (17)]. From (4) and using the delta method, we have:

$$\frac{1}{\sqrt{\Delta_n}}\left(\Gamma'\widehat{IV}^n\Gamma - \mathbf{\Delta}\right) \xrightarrow{\mathcal{L}-s} \mathcal{U}_T, \tag{20}$$

where $\mathcal{U}_T = \Gamma'\mathcal{W}_T\Gamma$ and $\mathcal{W}_T$ is given by (4). We have the following result.

Theorem 3.1. *Let X be an Itô semimartingale represented by (1). If Assumption (H-r) holds for some $r \in [0,1)$ and the truncation level $\varpi \in \left[\frac{1}{2(2-r)}, \frac{1}{2}\right)$, then:*

(a) Under H_0 as in Proposition 3.1(b), $LR_k \xrightarrow{\mathcal{L}-s} \frac{T}{2\lambda_k^2}\left(tr(\mathcal{U}_{kk}^2) - \frac{1}{q_k}\left(tr(\mathcal{U}_{kk})\right)^2\right)$, where $\mathcal{U}_{kk}$ is the (q_k, q_k)-submatrix of $\mathcal{U}_T$ at the intersection of the $(q_1+\cdots+q_{k-1}+1)$-th through the $(q_1+\cdots+q_k)$-th rows and columns.

(b) Under the alternative (i.e. if H_0 does not hold), $LR_k \to \infty$, in probability.

(c) Under $H_0(\lambda)$ as in Proposition 3.1(d), $LR_{k,\lambda} \xrightarrow{\mathcal{L}-s} \frac{T}{2\lambda^2}tr(\mathcal{U}_{kk}^2)$ and under the alternative, $LR_{k,\lambda} \to \infty$, in probability.

Theorem 3.1 generalizes the results in Proposition 3.1(c,d) to the class of Itô semimartingales. The asymptotic distributions of LR_k and $LR_{k,\lambda}$ are no longer guaranteed to be pivotal as previously. Indeed, they depend on nuisance parameters such as the common value λ_k of the relevant cluster of eigenvalues of IV_T, the conditional variance-covariance matrix of $\mathcal{W}_T$ which is equal to $\int_0^T (c_s^{il}c_s^{jm} + c_s^{im}c_s^{jl})ds$, for $i,j,l,m = 1,\ldots,q$ and the matrix Γ of normalized eigenvectors of IV_T.

As already mentioned, λ_k is consistently estimated by $\frac{1}{q_k}\sum_{i\in\mathscr{L}_k} d_i$. Estimators of the conditional variance of $\mathcal{W}_T$ have been proposed by Barndorff-Nielsen and Shephard (2004); see also Jacod and Protter (2012) for jump robust estimators. If Γ can be consistently estimated, then this asymptotic

distribution can be simulated to generate critical values for inference. However, the presence of multiple roots makes it impossible to consistently estimate Γ even if the identifying restriction that the elements of its main diagonal are positive is maintained. Nevertheless, the fact that only the trace of $\mathcal{U}_{kk}^h$, for $h = 1, 2$, is useful for these asymptotic distributions offers some possibility of simulating these distributions as we describe below.

Write $\Gamma = (\Gamma_1 \cdots \Gamma_r)$, where Γ_k, for $k = 1, ..., r$, corresponds to the eigenvectors associated to the sorted eigenvalues with rank indexes in the cluster $\mathscr{L}_k$ so that $\mathcal{U}_{kk} = \Gamma_k' \mathcal{W}_T \Gamma_k$. Consider $A_n = \Gamma' \widehat{IV}^n \Gamma$ and $\widehat{E}$ the matrix of its normalized eigenvectors with main diagonal elements restricted to be nonnegative. A_n and $\widehat{IV}^n$ have the same set of eigenvalues and $\widehat{\Gamma} = \Gamma \widehat{E}$ is a matrix of normalized eigenvectors of $\widehat{IV}^n$.

Write $\widehat{E} = (\widehat{E}_{kl})_{1 \leq k,l \leq r}$, where, for $k, l = 1, \ldots, r$, $\widehat{E}_{kl}$ is a block (q_k, q_l)-submatrix of $\widehat{E}$ at the intersection of its rows and columns in $\mathscr{L}_k$ and $\mathscr{L}_l$, respectively. Proposition A.1 in Appendix A.3 shows that $\widehat{E}_{kl} = o_P(1)$ for $k \neq l$ and $\widehat{E}_{kk} \widehat{E}_{kk}' = I_{q_k} + o_P(1)$. Thus, writing $\widehat{\Gamma} = \left(\widehat{\Gamma}_1 \cdots \widehat{\Gamma}_r \right)$ with $\widehat{\Gamma}_k$ defined similarly to Γ_k $(k = 1, \ldots, r)$, we have

$$\widehat{\Gamma} = \Gamma \widehat{E}, \quad \text{and} \quad \widehat{\Gamma}_k = \Gamma_k \widehat{E}_{kk} + o_P(1). \tag{21}$$

Even though $\widehat{\Gamma}$ is not a consistent estimator of Γ (unless $q_k = 1$), this estimator is useful to consistently simulate the asymptotic distribution of LR_k and $LR_{k,\lambda}$. Indeed, (21) implies that

$$\left(\widehat{\Gamma}_k' \mathcal{W}_T \widehat{\Gamma}_k \right)^h = \widehat{E}_{kk}' \left(\Gamma_k' \mathcal{W}_T \Gamma_k \right)^h \widehat{E}_{kk} + o_P(1), \quad \text{for} \quad h = 1, 2,$$

which, in turn, implies that

$$tr \left[\left(\widehat{\Gamma}_k' \mathcal{W}_T \widehat{\Gamma}_k \right)^h \right] = tr \left(\mathcal{U}_{kk}^h \right) + o_P(1).$$

It follows that, if one can simulate from the distribution of $\mathcal{W}_T$ (or its approximate distribution if the variance $\mathcal{W}_T$ is estimated), $\widehat{\Gamma}_k$ can be used to obtain consistent simulations from the distribution of $tr(\mathcal{U}_{kk}^h)$, which in turn can be used to generate asymptotically valid critical values for the tests of interest.

Despite it's usefulness, this direct simulation approach is anticipated to be quite computationally tedious to implement and possibly inaccurate due to the presence of many nuisance parameters requiring estimation. This motivates the bootstrap approach that we introduce in the next section as an alternative. Next, we discuss an application of the tests presented in Theorem 3.1 beyond their usefulness for the characterization of the eigenvalue structure in (9).

Test for the ratio of 'unexplained' volatility. By analogy with the principal component analysis of variance matrices, it is of interest to quantify the proportion of volatility (or quadratic variation) captured by the principal components of X^c - the continuous part of X - associated to the largest eigenvalues of IV_T. Formally, given a ratio $\pi \in (0, 1)$, we would like to test whether the total amount of volatility not captured by the first Q-principal components does not exceed π. This can be stated

as:

$$H_{0\pi}: \quad \sum_{i=Q+1}^{q} \delta_i \leq \pi \sum_{i=1}^{q} \delta_i, \tag{22}$$

where $\delta = \lambda(IV_T)$ is the vector of eigenvalues of IV_T. Let us consider the test statistic Z_n defined as follows:

$$Z_n = \frac{1}{\sqrt{\Delta_n}} \left(\sum_{i=Q+1}^{q} d_i - \pi \sum_{i=1}^{q} d_i \right), \tag{23}$$

where $d = \lambda\left(\widehat{IV}^n\right)$. We have the following result.

Theorem 3.2. *Let X be an Itô semimartingale represented by (1). Assume that Assumption (H-r) holds for some $r \in [0,1)$ and the truncation level $\varpi \in \left[\frac{1}{2(2-r)}, \frac{1}{2}\right)$. Assume that $\delta_Q > \delta_{Q+1}$ and let $\pi \in (0,1)$.*

(a) If $H_{0\pi}$ holds with equality, then $Z_n \xrightarrow{\mathcal{L}-s} Z \equiv tr\left(\mathcal{U}_{Q+1:q,Q+1:q;T}\right) - \pi \cdot tr\left(\mathcal{U}_T\right)$, where $\mathcal{U}_T$ is defined in Equation (20) and $\mathcal{U}_{Q+1:q,Q+1:q;T}$ is the bottom right $(q-Q, q-Q)$-submatrix of $\mathcal{U}_T$ determined by its last $q-Q$ rows and columns.

(b) If $H_{0\pi}$ holds with strict inequality, then

$$\lim_n P(Z_n > c_{1-\alpha}) = 0,$$

with $\alpha \in (0,1)$ and $c_{1-\alpha}$ is the $(1-\alpha)$-quantile of Z.

(c) If $H_{0\pi}$ does not holds, then $Z_n \to \infty$, in probability.

Theorem 3.2 derives the asymptotic distribution of Z_n when the exact ratio of unexplained volatility is π. If X is a continuous Lévy process, this asymptotic distribution is a centered Gaussian with variance derived from the result in Equation (A.14) in Appendix A.3:

$$\frac{2}{T}(1-\pi)^2 \sum_{k=1}^{k_1-1} q_k \lambda_k^2 + \frac{2}{T}\pi^2 \sum_{k=k_1}^{r} q_k \lambda_k^2,$$

where k_1 is such that $Q+1 \in \mathscr{L}_{k_1}$. Even in the simple case of Lévy process, this asymptotic distribution is not pivotal in general and its simulation presents a similar challenge as that discussed at Theorem 3.1. Direct simulation can be performed from some approximation of the distribution of $tr(\mathcal{U}_T)$ and $tr(\mathcal{U}_{Q+1:q,Q+1:q;T})$ using the matrix $\widehat{\Gamma}$ of normalized eigenvectors of $\widehat{IV}^n$. If approximate copies of $\mathcal{W}_T$ can be generated, then the equalities

$$tr(\widehat{\Gamma}'\mathcal{W}_T\widehat{\Gamma}) = tr(\mathcal{U}_T) + o_P(1) \quad \text{and} \quad tr(\widehat{\Gamma}'_{Q+1:q,Q+1:q}\mathcal{W}_T\widehat{\Gamma}_{Q+1:q,Q+1:q}) = tr(\mathcal{U}_{Q+1:q,Q+1:q;T}) + o_P(1)$$

provide useful copies of $tr(\mathcal{U}_T)$ and $tr(\mathcal{U}_{Q+1:q,Q+1:q;T})$. However, as already mentioned, simulating from the distribution of $\mathcal{W}_T$ can be tedious and one shall rely on the bootstrap method that we propose

in the next section. The divergence to infinity of Z_n under the alternative guarantees that the test is consistent.

We close this section by noting that the principal component factors are estimated by linear combinations of the jump-filtered returns process using, as usual, eigenvectors associated to the largest eigenvalues.

Eigenvalue structure in correlation. Correlation matrices are insensitive to linear transformations of data and this makes them more appealing than variance matrices for principal component analysis in many applications. We derive similar results to those in Theorem 3.1 for testing the eigenvalue structure of correlation matrices. Let X be an Itô semimartingale described by (1) with integrated variance IV_T over $[0,T]$. We define the correlation matrix R_T of X over $[0,T]$ by

$$R_T = G(IV_T), \tag{24}$$

where $G : \mathcal{M}_q^{++} \to \mathcal{M}_q^{++}$ and $\forall A \in \mathcal{M}_q^{++},\ G(A) = S(A)AS(A)$, with $S(A)$ is the diagonal matrix with diagonal elements $1/\sqrt{A_{ii}}$, for $i = 1, \ldots, q$.

Since G is differentiable, $\widehat{R}^n \equiv G(\widehat{IV}^n)$ is a consistent estimator of R_T, and thanks to (7), we have

$$\frac{1}{\sqrt{\Delta_n}}\left(\widehat{R}^n - R_T\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}_T^G, \tag{25}$$

with $\mathcal{W}_T^G$ defined as in Equation (7) for $\varphi = G$. We propose a test of eigenvalue structure of the correlation matrix R_T using the same statistics as those used for IV_T. Assume that R_T has an eigenvalue structure as in (9), and let LR_k^ρ and $LR_{k,\lambda}^\rho$ be defined as LR_k and $LR_{k,\lambda}$, respectively, but using $d = \lambda\left(\widehat{R}^n\right)$ as estimator of the vector of eigenvalues δ of R_T.

Let Γ^ρ be the (q,q)-orthogonal matrix such that $\Gamma^{\rho'} R_T \Gamma^\rho = \mathbf{\Delta}^\rho$, where $\mathbf{\Delta}^\rho$ is the diagonal matrix with diagonal elements equal to the eigenvalues δ_i's of R_T. By the delta method and using (25), we can claim that

$$\frac{1}{\sqrt{\Delta_n}}\left(\Gamma^{\rho'}\widehat{R}^n\Gamma^\rho - \mathbf{\Delta}^\rho\right) \xrightarrow{\mathcal{L}-s} \mathcal{U}_T^\rho, \tag{26}$$

where $\mathcal{U}_T^\rho = \Gamma^{\rho'}\mathcal{W}_T^G\Gamma^\rho$. We can state the following result.

Theorem 3.3. *Assume that the conditions of Theorem 3.1 hold. Then the results (a), (b) and (c) of Theorem 3.1 also hold when LR_k, $LR_{k,\lambda}$ and $\mathcal{U}_{kk}$ are replaced by LR_k^ρ, $LR_{k,\lambda}^\rho$ and $\mathcal{U}_{kk}^\rho$, respectively, with H_0 and $H_0(\lambda)$ involving restrictions on the eigenvalues of R_T.*

A similar application to the *test for the ratio of unexplained volatility* also extend to the correlation matrix R_T with the same testing procedures as those described for IV_T. As in Theorem 3.1, the asymptotic distributions provided in Theorem 3.3 are non standard and difficult to simulate directly. The bootstrap methods that we introduce next provide a useful alternative.

4 The bootstrap

This section introduces bootstrap methods for testing eigenvalue structure of the integrated variance matrix IV_T and correlation matrix R_T. These methods are of a particular interest when the price vector process follows a general form of Itô semimartingale dynamics. As already pointed out, the asymptotic distribution of the test statistics presented by Theorems 3.1, 3.2 and 3.3 have nuisance parameters that are costly to estimate. Although we focus on IV_T and R_T in this section, the techniques developed are also useful to spot variance matrix c_t as we will see in the next section.

We observe that all the statistics of interest for IV_T and R_T are functions of the integrated covariance matrix estimator $\widehat{IV}^n$. Therefore, an important step towards bootstrapping these statistics consists in bootstrapping $\widehat{IV}^n$ itself. The asymptotic distribution of estimators of IV_T and some of its functions have been object of approximation by bootstrap in recent literature. Dovonon, Gonçalves, and Meddahi (2013) have applied the non-parametric i.i.d. bootstrap to approximate the distribution of the so-called realized beta and realized correlation between assets. However, as they point out, the non-parametric i.i.d. bootstrap is not capable, in general, of reproducing the exact asymptotic distribution of estimators of IV_T. Hounyo (2017) has generalized to the multivariate setting the idea of wild blocks of blocks bootstrap of Hounyo, Gonçalves, and Meddahi (2017), which is of interest to us. While standard bootstrap methods focus on sampling point-wise returns, the wild blocks of blocks bootstrap of Hounyo (2017) samples the summands of $\widehat{IV}^n$ and, thereby, reproduces its exact asymptotic distribution. Note that the bootstrap method of Hounyo (2017) is designed to approximate the asymptotic distribution of estimators of IV_T that are robust to market microstructure noise with asynchronous data. The complexity of his data structure and model justifies the fact that he resorts to block bootstrap schemes.

We utilize the method in Hounyo (2017) to bootstrap $\widehat{IV}^n$. Since we are concerned with synchronous price observations that depart from noisy environments, we rely on a version of the wild bootstrap that does not involve blocks. Even though - for simplicity of exposition - we do not account for noise and non-synchronicity in this paper, the test statistics that we introduce in the previous section can be based on noise robust estimators of IV_T. In this case, we shall rely on the full wild blocks of blocks bootstrap method of Hounyo (2017) to obtain an accurate estimation of the asymptotic distributions. To introduce the wild bootstrap for $\widehat{IV}^n$, we first introduce some notation. For $i = 1, \ldots, n$, let

$$y_i = \Delta_i^n X 1_{\{\|\Delta_i^n X\| \leq \alpha \Delta_n^{\varpi}\}}, \quad \mathcal{Z}_i = y_i y_i'$$

and η_i, for $i = 1, \ldots, n$ be a sequence of independent and identically distributed random variables that are all independent of y_i's and such that $E(\eta_i) = 1$ and $Var(\eta_i) = 1/2$. Consider the wild bootstrap sample $\mathcal{Z}_i^*$ $(i = 1, \ldots, n)$ of $\mathcal{Z}_i$ $(i = 1, \ldots, n)$ which is given by

$$\mathcal{Z}_i^* = \mathcal{Z}_{i+1} + (\mathcal{Z}_i - \mathcal{Z}_{i+1})\eta_i \quad \text{if} \quad i = 1, \ldots, n-1, \quad \text{and} \quad \mathcal{Z}_n^* = \mathcal{Z}_n. \tag{27}$$

Let $\widehat{IV}^{*^n} = \sum_{i=1}^n \mathcal{Z}_i^*$ be the bootstrap analogue of $\widehat{IV}^n$ and let

$$S_n^* = \frac{1}{\sqrt{\Delta_n}} \left(\widehat{IV}^{*^n} - \widehat{IV}^n \right) \quad \text{be the bootstrap analogue of} \quad S_n = \frac{1}{\sqrt{\Delta_n}} \left(\widehat{IV}^n - IV_T \right).$$

Under some regularity conditions, we can show that S_n^* has the same asymptotic distribution as S_n under the bootstrap measure making the wild bootstrap first-order asymptotically valid. Before stating this result formally in Proposition 4.1 below, we first recall the following standard notation related to the bootstrap theory. We let P^*, E^* and Var^* denote the probability measure, the expected value and the variance, respectively, induced by the bootstrap resampling conditional on the original sample. Let Y_n^* be a sequence of bootstrap statistics indexed by n. We say that $Y_n^* \xrightarrow{P^*} 0$ in prob-P (also denoted by $Y_n^* = o_{P^*}(1)$ in prob-P) if, for any $\varepsilon > 0$, $P^*(|Y_n^*| > \varepsilon) \to 0$ in probability as $n \to \infty$. Similarly, we say that $Y_n^* = O_{P^*}(1)$ in prob-P if $\sup_n P^*(|Y_n^*| > M) \to 0$ in probability as $M \to \infty$. Finally, we write $Y_n^* \xrightarrow{d^*} Y$ in prob-P if, conditionally on the sample, Y_n^* converges weakly to Y under the measure P^* and this for all sample contained in a set with probability P converging to one. *'vech'* denotes the half-vectorization operator that stacks the lower-triangle part of a matrix into a vector. We can claim the following result.

Proposition 4.1. *Let X be an Itô semimartingale represented by (1). If Assumption (H-r) holds for some $r \in (0,1)$, $\varpi \in [\frac{1}{2(2-r)}, \frac{1}{2})$, and $E|\eta_i|^{2+\epsilon} < \infty$ for some $\epsilon > 0$. Then* $\sup_{x \in \mathbb{R}^{q(q+1)/2}} \mid P^*\left(vech(S_n^*) \leq x\right) - P\left(vech(S_n) \leq x\right) \mid \xrightarrow{P^*} 0$, *in probability.*

Let $\varphi : \mathcal{M}_q^+ \to \mathbb{R}^k$ be a smooth function and let

$$T_n = \frac{1}{\sqrt{\Delta_n}}\left(\varphi\left(\widehat{IV}^n\right) - \varphi(IV_T)\right) \quad \text{and} \quad T_n^* = \frac{1}{\sqrt{\Delta_n}}\left(\varphi\left(\widehat{IV}^{*^n}\right) - \varphi\left(\widehat{IV}^n\right)\right).$$

T_n converges stably in law to $\mathcal{W}_T^{\varphi}$ [see Equation (7)] with

$$\mathcal{W}_T^{\varphi} = \sum_{u,v=1}^{q} \left.\frac{\partial \varphi(M)}{\partial M_{uv}}\right|_{M=IV_T} \mathcal{W}_{T,uv},$$

with $\mathcal{W}_T$ defined in (4) and (5). The next result derives from Proposition 4.1 by the application of the delta method. In particular, it states that T_n and T_n^* have the same asymptotic distribution.

Corollary 4.1. *Under the same conditions as Proposition 4.1, if $\mathcal{W}_T^{\varphi}$ has a continuous distribution on $\mathbb{R}^k$, then* $\sup_{x \in \mathbb{R}^k} |P^*(T_n^* \leq x) - P(T_n \leq x)| \xrightarrow{P^*} 0$, *in probability.*

Corollary 4.1 establishes the validity of the proposed bootstrap to approximate the asymptotic distribution of any smooth function of $\widehat{IV}^n$. The most practical benefit of this result is that there is no need to estimate any of the multiple nuisance parameters, including $\partial\varphi(IV_T)/\partial M_{uv}$, that the asymptotic distribution of T_n depends on to estimate its quantiles. Bootstrap quantiles obtained from bootstrap replications of T_n^* can serve as asymptotically valid quantiles for T_n.

Corollary 4.1 has some immediate application for the inference on eigenvalues and eigenvectors of IV_T. If the eigenvalue structure of IV_T is known to be that displayed in (9) and φ^{λ} is the eigenvalue function defined in (10), then, by smoothness of φ^{λ}, we can claim using (11) that:

$$T_n^{\lambda^*} \equiv \frac{1}{\sqrt{\Delta_n}}\left(\varphi^{\lambda}\left(\widehat{IV}^{n^*}\right) - \varphi^{\lambda}\left(\widehat{IV}^n\right)\right) \xrightarrow{d^*} \mathcal{W}_T^{\varphi^{\lambda}},$$

in probability. This means that $T_n^{\lambda^*}$ provides asymptotically valid approximation to the distribution of $\mathcal{W}_T^{\varphi^\lambda}$ that can be used to carry out inference about any component of $\varphi^\lambda(IV_T)$.

Similarly, if the ith largest eigenvalue of IV_T is simple, then up to some identifying sign restriction, the function $\gamma_i(A)$ equal to the eigenvector associated to the ith largest eigenvalue of A is smooth in a neighborhood of IV_T and once again, from (12), we can claim that:

$$\frac{1}{\sqrt{\Delta_n}}\left(\gamma_i\left(\widehat{IV}^{n^*}\right)-\gamma_i\left(\widehat{IV}^{n}\right)\right)\xrightarrow{d^*}\mathcal{W}_T^{\gamma_i},$$

in probability and, from (12), inference on $\gamma_i(IV_T)$ or on any of its components can be carried out using the bootstrap.

We now turn to the main contribution of this section which is bootstrapping the test statistics in Theorems 3.1 and 3.2 and their related applications. A natural way to bootstrap these test statistics would consist in using bootstrap analogue of (15), (16) and (23), with the vector of bootstrap eigenvalues $d^* = \lambda(\widehat{IV}^{n^*})$ as input. However, such bootstrap procedures would fail since they would intrinsically test for the eigenvalue structure in $\widehat{IV}^n$ which is different than that of IV_T. Indeed, considering the asymptotic distribution of $\widehat{IV}^n$ in (4), conditionally on $\mathcal{F}$, $\widehat{IV}^n$ is a random matrix with any pair of eigenvalues different with probability approaching one. To circumvent this issue, we focus on the leading term in the expansion of the test statistics of interest - instead of the test statistics themselves - that we bootstrap by using the bootstrap of $\widehat{IV}^n$ as a key input. We then establish the first-order asymptotic validity of the proposed method.

As previously defined, let the matrix Γ of normalized eigenvectors of IV_T be $\Gamma = (\Gamma_1\Gamma_2\cdots\Gamma_r)$ where Γ_k is associated to the q_k-multiple eigenvalue λ_k $(k = 1,\ldots,r)$. Let Γ_0 be equal to Γ_k or $(\Gamma_{k_1}\cdots\Gamma_r)$, or even Γ, and let Q be the integer defined such that $\bigcup_{j=k_1}^r \mathscr{L}_j = \{Q+1,\ldots,q\}$. We observe that the asymptotic distributions of interest in the previous theorems are functions of $tr(\mathcal{U}^h_{T,\Gamma_0\Gamma_0}) \equiv tr\left[(\Gamma_0'\mathcal{W}_T\Gamma_0)^h\right]$, $h = 1,2$, where $\mathcal{W}_T$ is the asymptotic distribution of $\widehat{IV}^n$, which is defined in Equation (4).

If Γ were known, using Corollary 4.1, the distribution of $\Gamma'\mathcal{W}_T\Gamma$ can be estimated by that of $\Gamma' S_n^*\Gamma$. However, as already mentioned, Γ is unknown and cannot be consistently estimated in general. We have also seen in Section 3 that the fact that the asymptotic distribution of $\mathcal{U}_T$ appears through the trace operator makes some estimator of Γ useful for direct simulation. We have namely used $\widehat{\Gamma}$, the matrix of normalized eigenvectors of $\widehat{IV}^n$ [see Equation (21)]. Let $\widehat{\Gamma}_0$ be defined from $\widehat{\Gamma}$ as Γ_0 is defined from Γ. Using Equation (21), we have

$$\widehat{\Gamma}_0 = \Gamma\widehat{E}_0 = \Gamma_0\check{E}_0 + o_P(1),$$

where $\widehat{E}_0$ is a matrix equal to the collection of columns of $\widehat{E}$ indexed by $\mathscr{L}_k$ or $\bigcup_{k=1}^r \mathscr{L}_k$, or is equal to $\widehat{E}$ depending of Γ_0. $\check{E}_0 = \widehat{E}_{kk}$ or a block-diagonal matrix with $\widehat{E}_{kk}$ $(k = k_1,\ldots,r)$ on the main diagonal or $\check{E}_0 = \widehat{E}$ also depending on Γ_0. The order of magnitude above is obtained from the properties of $\widehat{E}$ outlined in Equation (21); see Proposition A.1 in Appendix A.3. This proposition also ensures that

$\check{E}_0\check{E}_0' = I + o_P(1)$. Thus,

$$tr\left[\left(\widehat{\Gamma}_0' S_n^* \widehat{\Gamma}_0\right)^h\right] = tr\left[\widehat{E}_0' \left(\Gamma_0' S_n^* \Gamma_0\right)^h \widehat{E}_0\right] + o_{P^*}(1) = tr\left[\left(\Gamma_0' S_n^* \Gamma_0\right)^h\right] + o_{P^*}(1), \quad h = 1, 2.$$

Therefore, Corollary 4.1 ensures that

$$tr\left[\left(\widehat{\Gamma}_0' S_n^* \widehat{\Gamma}_0\right)^h\right] \xrightarrow{d^*} tr\left[\left(\Gamma_0' \mathcal{W}_T \Gamma_0\right)^h\right] \tag{28}$$

in probability; showing that $tr\left[\left(\widehat{\Gamma}_0' S_n^* \widehat{\Gamma}_0\right)^h\right]$ is an asymptotically valid estimator of the distribution $tr(\mathcal{U}^h_{T,\Gamma_0\Gamma_0})$. With this insight, we can now introduce the bootstrap statistics for the tests of interest. Let

$$U^* = \frac{1}{\sqrt{\Delta_n}}\left(\widehat{\Gamma}' \widehat{IV}^{n^*} \widehat{\Gamma} - \mathbf{D}\right) = \widehat{\Gamma}' S_n^* \widehat{\Gamma},$$

where $\mathbf{D}$ is the diagonal matrix with diagonal equals to $d = \lambda(\widehat{IV}^n)$, the vector of sorted eigenvalues of $\widehat{IV}^n$. Let $\hat{\lambda}_k = \frac{1}{q_k}\sum_{i\in\mathscr{L}_k} d_i$, and U^*_{kk} and $U^*_{Q+1:q,Q+1:q}$ be, respectively, the (q_k, q_k)-submatrix of U^* at the intersection of the $(q_1 + \cdots + q_{k-1} + 1)$-th through the $(q_1 + \cdots + q_k)$-th rows and columns and the lower-right $(q - Q, q - Q)$-submatrix of U^*. We consider the following bootstrap test statistics:

$$\begin{aligned} LR_k^* &= \frac{T}{2\hat{\lambda}_k^2}\left(tr(U_{kk}^{*2}) - \frac{1}{q_k}\left(tr(U_{kk}^*)\right)^2\right) \\ LR_{k,\lambda}^* &= \frac{T}{2\lambda^2} tr(U_{kk}^{*2}) \\ Z_n^* &= tr\left(U^*_{Q+1:q,Q+1:q}\right) - \pi \cdot tr\left(S_n^*\right), \quad \text{for some } \pi \in (0,1). \end{aligned} \tag{29}$$

Note that $tr\left(S_n^*\right) = tr\left(U^*\right)$. These bootstrap test statistics can be seen as the bootstrap analogues of the first-order asymptotic approximation of the original test statistics. We have the following result.

Theorem 4.1. *Under the same conditions as in Proposition 4.1 and letting LR_k, $LR_{k,\lambda}$ be defined as in Theorem 3.1 and Z_n as in Theorem 3.2, we have:*

(a) Under H_0, as in Proposition 3.1(b), $\sup_{x\in\mathbb{R}}\left|P^*\left(LR_k^* \le x\right) - P(LR_k \le x)\right| \xrightarrow{P^*} 0$, *in probability.*

(b) Under $H_0(\lambda)$, as in Proposition 3.1(d), $\sup_{x\in\mathbb{R}}\left|P^*\left(LR_{k,\lambda}^* \le x\right) - P(LR_{k,\lambda} \le x)\right| \xrightarrow{P^*} 0$, *in probability.*

(c) Under $H_{0\pi}$, as in (22) and if $\delta_Q > \delta_{Q+1}$, $\sup_{x\in\mathbb{R}}\left|P^*\left(Z_n^* \le x\right) - P(Z_n \le x)\right| \xrightarrow{P^*} 0$, *in probability.*

This theorem establishes the asymptotic validity of the bootstrap when the specified bootstrap statistics are used. Also, these bootstrap statistics are all bounded in probability even under the alternative so that the bootstrap tests are consistent.

Remark 3. *It is worth reiterating that by construction, the proposed bootstrap method targets the replication of the first-order asymptotic approximation of the test statistics of interest as opposed to mimicking the original test statistics - which, we know, leads to invalid bootstrap approximations. In that respect, we may not be able to obtain the standard higher-order refinement properties for these*

bootstrap tests as this essentially amounts to a match of the higher-order cumulants of original and bootstrap test statistics. Nevertheless, as illustrated by the simulation results in the next section, the bootstrap approximation displays a satisfactory level of accuracy even for sample sizes as small as 160, which corresponds roughly to 5-minute observations within 2 trading days.

Remark 4. *The above results carry over to our asymptotic analysis of the correlation matrix. The useful bootstrap test statistics are defined similarly to LR_k^*, $LR_{k,\lambda}^*$ and Z_n^* in* (29) *but using*

$$U^{\rho^*} = \frac{1}{\sqrt{\Delta_n}} \left(\widehat{\Gamma}^{\rho\prime} G \left(\widehat{IV}^{n^*} \right) \widehat{\Gamma}^{\rho} - \mathbf{D}^{\rho} \right)$$

(instead of U^*), where G is defined as in (24)*, $\mathbf{D}^{\rho}$ is the diagonal matrix of the sorted eigenvalues of $\widehat{R}^n \equiv G(\widehat{IV}^n)$, and $\widehat{\Gamma}^{\rho}$ is the orthogonal matrix of eigenvectors of $\widehat{R}^n$. These bootstrap test statistics also use $\hat{\lambda}_k$ obtained from $\widehat{R}^n$ instead of $\widehat{IV}^n$.*

Before ending this section, we provide detailed algorithms of the implementation of the bootstrap tests for the equality of eigenvalues and for the proportion of unexplained volatility as introduced in Section 3.

ALG 1. **Bootstrap algorithm for testing $H_0(k)$: 'Equality of eigenvalues of IV_T in the cluster $\mathscr{L}_k$'. That is: $\delta_{q_1+\cdots+q_{k-1}+1} = \cdots = \delta_{q_1+\cdots+q_k} = \lambda_k$ (unknown); see** (9)**.**

1. Compute $\widehat{IV}^n$ as given by (2).
2. Compute $d = \lambda(\widehat{IV}^n)$ and $\widehat{\Gamma}$ the vector of sorted eigenvalues and the associated orthogonal matrix of eigenvectors and let $\mathbf{D}$ be the diagonal matrix such that $\widehat{\Gamma}'\widehat{IV}^n\widehat{\Gamma} = \mathbf{D}$.
3. Compute the test statistic: $LR_k = -2\log \ell_k$, where ℓ_k is given as in (15) but using d_i ($i \in \mathscr{L}_k = \{q_1 + \cdots + q_{k-1} + 1, \ldots, q_1 + \cdots + q_k\}$).
4. Bootstrap approximation of the asymptotic distribution of LR_k:
   (a) Draw n independent copies of η_i such that $E(\eta_i) = 1$ and $Var(\eta_i) = 1/2$. One possibility is to take $\eta_i = (v_{1i} + v_{2i})/4$ with $v_i \sim$ i.i.d.$\chi^2(2)$; and another one is to take: $\eta_i \sim \text{NID}(1, 1/2)$.
   (b) Get the bootstrap sample by computing $\mathcal{Z}_i^*$ ($i = 1, \ldots, n$) using (27)
   (c) Get $\widehat{IV}^{n^*} = \sum_{i=1}^n \mathcal{Z}_i^*$ and $U^* = \frac{1}{\sqrt{\Delta_n}} \left(\widehat{\Gamma}'\widehat{IV}^{n^*}\widehat{\Gamma} - \mathbf{D} \right)$.
   (d) U_{kk}^* is the $\mathscr{L}_k \times \mathscr{L}_k$ block of U^*.
   (e) Get a bootstrap copy of LR_k as: $LR_k^* = \frac{T}{2\hat{\lambda}_k^2} \left[tr[(U_{kk}^*)^2] - \frac{1}{q_k}[tr(U_{kk}^*)]^2 \right]$, with $\hat{\lambda}_k = \frac{1}{q_k}\sum_{i\in\mathscr{L}_k} d_i$ is the average of the q_k eigenvalues of $\widehat{IV}^n$ in the range $\mathscr{L}_k$.
5. Repeat Step 4 B times (e.g. $B = 399$) to get as many bootstrap copies of LR_k: $LR_{k,b}^*$, $b = 1, \ldots, B$.
6. Use this bootstrap sample to obtain the $(1-\alpha)$-quantile, say $l_{k,1-\alpha}^*$, of LR_k.

7. Reject $H_0(k)$ at the level α if $LR_k > l^*_{k,1-\alpha}$.

ALG 2. **Bootstrap algorithm for testing for ratio of 'unexplained' quadratic variation; i.e., $H_{0\pi}$: 'The first Q principal components support a proportion at least $1-\pi$ of variations.'**

1. Choose Q and $\pi \in (0,1)$: the number of factors to be tested for carrying at least a proportion $1-\pi$ of variation in IV_T.
2. Perform Steps 1 and 2 of the previous algorithm.
3. Compute Z_n, the test statistic for $H_{0\pi}$ as given in (23).
4. Bootstrap approximation of the distribution of Z:
   (a), (b), (c): Same as 4.(a), (b), (c) in the previous algorithm.
   (d) Obtain U^*_{22}, the lower-right $(q-Q, q-Q)$-submatrix of U^*.
   (e) Get a bootstrap copy of Z_n: $Z^*_n = tr(U^*_{22}) - \pi \cdot tr(U^*)$.
5. Repeat Step 4 B times (e.g. $B = 399$) to get as many bootstrap copies of Z_n: $Z^*_{n,b}$, $b = 1, \ldots, B$.
6. Use this bootstrap sample to obtain the $(1-\alpha)$-quantile, say $c^*_{1-\alpha}$, of Z_n.
7. Reject $H_{0\pi}$ at the level α if $Z_n > c^*_{1-\alpha}$.

5 Testing the eigenvalue structure of spot covariance

In this section, we extend the previously proposed tests for eigenvalue structure to spot variance matrix c_t at date t. As presented in Sections 3 and 4, these tests have a natural application to PCA of the vector of stock price process. The main interest of basing PCA on the spot variance is that, regardless of its dynamics, PCA on c_t yields the instantaneous factor structure of the vector of stock prices. In particular, the eigenvector associated to the largest eigenvalue of c_t gives the direction of the largest spot variance of X^c.

Let's assume that c_t has the eigenvalue structure in (9). As previously, our goal is to test for a given row $\mathscr{L}_k (k = 1, \ldots, r)$ of this structure in both cases of λ_k unknown (H_0) or set to a known value λ ($H_0(\lambda)$). Our approach is to rely on the likelihood ratio test statistic defined by:

$$LR^c_k = -2 \log \ell^c_k, \quad \text{and} \quad LR^c_{k,\lambda} = -2 \log \ell^c_{k,\lambda},$$

where ℓ^c_k and $\ell^c_{k,\lambda}$ are obtained using (15) and (16), respectively but with $d = \lambda(\hat{c}(t, k_n))$ and $\hat{c}(t, k_n)$ the consistent estimator of c_t given by (3). Unless X is a continuous Lévy process, the resulting test statistics are not likelihood ratios. Nevertheless, the following proposition, analogue to Theorem 3.1, shows that these statistics are useful to test for the eigenvalue structure of c_t.

Let Γ_c be the orthogonal matrix of normalized eigenvectors of c_t and $\mathbf{\Delta}_c$ the diagonal matrix of eigenvalues of c_t. An application of the delta method to (6) yields:

$$\sqrt{k_n} \left(\Gamma_c' \hat{c}(t, k_n) \Gamma_c - \mathbf{\Delta}_c \right) \xrightarrow{\mathcal{L}-s} \mathcal{U}_c, \tag{30}$$

with $\mathcal{U}_c = \Gamma_c' Z_t \Gamma_c$ with Z_t given by (6). We have the following result.

Proposition 5.1. *Let X be an Itô semimartingale represented by (1). If Assumption A.1 in Appendix holds then the conclusions (a), (b) and (c) of Theorem 3.1 hold with LR_k, $\mathcal{U}_{kk}$, $LR_{k,\lambda}$ and T replaced by LR_k^c, $\mathcal{U}_{c,kk}$, $LR_{k,\lambda}^c$ and 1, respectively.*

The proof of this theorem follows the same lines as that of Theorem 3.1 and therefore is omitted. Several clusters of eigenvalues can be jointly tested. As mentioned elsewhere, the test statistic to consider then is the sum of cluster specific statistics over the concerned clusters. The limit distribution corresponds to the sum of limits since convergence is joint across clusters.

The test for the ratio of 'unexplained' spot volatility can also be deployed for c_t. Given $\pi \in (0,1)$, the statement "that the total amount of volatility not captured by the first Q-principal components does not exceed π" amounts to the null hypothesis $H_{0\pi}$ in (22) with $\delta = \lambda(c_t)$. The useful test statistic for $H_{0\pi}$ for c_t is:

$$Z_n^c = \sqrt{k_n}\left(\sum_{i=Q+1}^{q} d_i - \pi \sum_{i=1}^{q} d_i\right), \quad \text{with} \quad d = \lambda(\hat{c}(t,k_n)).$$

We have the following result.

Proposition 5.2. *Let X be an Itô semimartingale represented by (1). If Assumption A.1 in Appendix holds and $\delta_Q > \delta_{Q+1}$ then the conclusions (a) and (b) of Theorem 3.2 hold with Z_n and $\mathcal{U}_T$ replaced by Z_n^c and $\mathcal{U}_c$, respectively.*

The proof is also similar to that of Theorem 3.2 and is omitted. The asymptotic distributions presented in Propositions 5.1 and 5.2 for the test statistics are not standard and have many nuisance parameters. We shall rely once again on the bootstrap for their approximation. Interestingly, thanks to the similarity between $\hat{c}(t,k_n)$ and $\widehat{IV}^n$ the bootstrap approximation of the asymptotic distribution in (6) is obtained the same way as in (27). Only returns local to t used in the expression of $\hat{c}(t,k_n)$ are bootstrapped. The bootstrap sample is given by:

$$\mathcal{Z}_m^* = \mathcal{Z}_{m+1} + (\mathcal{Z}_m - \mathcal{Z}_{m+1})\eta_m; \quad m = 0,\ldots,k_n-1, \quad \text{and} \quad \mathcal{Z}_{k_n-1}^* = \mathcal{Z}_{k_n-1};$$

we refer to (27) for more details. Let the bootstrap spot volatility estimate be given by $\hat{c}(t,k_n)^* = \frac{1}{k_n\Delta_n}\sum_{m=0}^{k_n-1}\mathcal{Z}_m^*$. Let

$$S_{nt} = \sqrt{k_n}(\hat{c}(t,k_n) - c_t), \quad \text{and its bootstrap analogue} \quad S_{nt}^* = \sqrt{k_n}(\hat{c}^*(t,k_n) - \hat{c}(t,k_n)).$$

Theorem 5.1. *Assume that Assumption A.1 holds for some $r \in [0,2)$ and $\tau \in (0,1/2)$; $\varpi \in (\frac{\tau}{2(2-r)} \vee \frac{1}{4-r}, \frac{1}{2})$; and $E|\eta_m|^{2+\epsilon} < \infty$ for some $\epsilon > 0$. Let t be fixed and assume that c_t nonsingular almost surely. Then*

$$\sup_{x\in\mathbb{R}^{q(q+1)/2}} \left|P^*\left(vech(S_{nt}^*) \le x\right) - P\left(vech(S_{nt}) \le x\right)\right| \xrightarrow{P^*} 0, \quad \textit{in probability.}$$

This result establishes the validity of the bootstrap for the estimation of the asymptotic distribution of S_{nt}. The proof of Theorem 5.1 requires a law of large numbers for functions of successive

local returns and with polynomial growth. This law of large numbers is presented in Appendix by Lemmas A.1 and A.2. The condition on ϖ is to ensure that the law of large numbers (see Lemma A.2) applies to local sample mean of function of log-returns that have polynomial growth of order $4+\epsilon$ for some $\epsilon > 0$. This is useful to verify the Lyapunov condition establishing bootstrap validity. The nonsingularity assumption for c_t ensure that stripping out the functionally related quantities, the asymptotic distribution of $\hat{c}(t, k_n)$ has a nonsingular variance matrix. This mild assumption is only made to simplify the proof.

Theorem 5.1 ensures that the proposed bootstrap method is valid for the approximation of the distributions of quantities analogues to those discussed for $\widehat{IV}^n$ in the previous section. Namely, $\varphi^\lambda(\hat{c}(t, k_n))$ and $\gamma_j(\hat{c}(t, k_n))$, where γ_j is the eigenvector function associated to the jth eigenvalue of c_t assuming it is simple. For both functions, the original and the bootstrap statistics are:

$$\sqrt{k_n}\left(\varphi(\hat{c}(t, k_n)) - \varphi(c_t)\right), \quad \text{and} \quad \sqrt{k_n}\left(\varphi(\hat{c}(t, k_n)^*) - \varphi(\hat{c}(t, k_n))\right),$$

respectively, with $\varphi = \varphi^\lambda, \gamma_j$ or G for the spot correlation (see Equation (24)). Likewise, with $\widehat{\Gamma}_c$ denoting the orthogonal matrix of normalized eigenvectors of $\hat{c}(t, k_n)$ and defining

$$U_c^* = \widehat{\Gamma}_c' S_{nt}^* \widehat{\Gamma}_c,$$

the statistics: LR_k^c, $LR_{k,\lambda}^c$ and Z_n^c have their bootstrap analogue, $LR_k^{c^*}$, $LR_{k,\lambda}^{c^*}$ and $Z_n^{c^*}$, defined as in (29) but with T and U^* replaced by 1 and U_c^*, respectively. We can verify along the same lines as in the proof of Theorem 4.1 that the bootstrap of these statistics is asymptotically valid. As a result, the algorithms ALG 1 and ALG 2 are also useful for inference on the eigenvalue structure of c_t. The main changes are: $\widehat{IV}^n$, ℓ_k, $\widehat{IV}^{n^*}$, LR_k, LR_k^*, U^*, Z_n Z_n^* are replaced by $\hat{c}(t, k_n)$, ℓ_k^c, $\hat{c}(t, k_n)^*$, LR_k^c, $LR_k^{c^*}$, U_c^*, Z_n^c and $Z_n^{c^*}$, respectively and the bootstrap sample $\mathcal{Z}_i^*$ $(i = 1, \ldots, n)$ replaced by $\mathcal{Z}_m^*$ $(m = 0, \ldots, k_n - 1)$.

6 Monte Carlo simulations

We conduct a Monte Carlo simulation study to investigate the finite sample performance of the tests proposed in Sections 3 and 4 for the equality of eigenvalues and ratio of unexplained volatility. Our primary focus is on assessing the empirical size and power of the asymptotic tests in Proposition 3.1 and the bootstrap-based test in Theorem 4.1 (see also Theorems 3.1 3.2) under a variety of data generating processes (DGPs).

For our simulation settings, we consider a continuous time process $X_t \in \mathbb{R}^q$ that represents a vector of q assets' prices, whose components are generated by the following m-factor model:

$$dX_{j,t} = \sum_{k=1}^{m^*} \beta_{k,j} df_{k,t} + de_{j,t}, \qquad \text{for } j = 1, ..., q,$$

where $X_{j,t}$ is the price of asset j at time t, $f_{k,t}$, for $k = 1, \ldots, m^*$, is the k-th factor at time t, $\beta_{k,j}$ is the

Table 1: Summary of Simulation Parameters

Stochastic Volatility and Jump Diffusion Parameters							
	κ_z	ξ_z	σ_z	ρ_z	μ_z^J	σ_z^J	ζ_z
SV	5	0.15^2	0.05	-0.5	0	0	-
SVJD	5	0.15^2	0.05	-0.5	0	0.01	$1/\Delta_n$

Parameters for the Lévy factors, Lévy noise process and factor loadings					
Lévy (h^f)	0.15^2	Noise (h^e)	$0.15^2 \times \bar{\pi}^*$	Loadings	$\beta_{i,j} \sim U(0,1)$

Note: The term $\bar{\pi}^*$ represents the target noise to signal ratio used in the simulations. Note that because $\beta_{i,j}$ is uniform then the total expected quadratic variation of the continuous factor component can be approximated by $\frac{1}{12} m^* \xi_z$ for the SV and SVJD factor models and $\frac{1}{12} m^* h^f$ for the Lévy model.

factor loading that captures the exposure of asset's j price to factor k, and $\{e_{j,t}\}$ is an uncorrelated noise process.

Following the literature, we next assume that the factors and noise are pairwise independent semi-martingale processes; see Pelger (2019). Generically, factors and noise are driven by the following Stochastic Volatility Jump Diffusion model with a constant jump intensity:

$$dz_{i,t} = \sqrt{h_{i,t}^z} dW_{i,t}^z + J_{i,t}^z d\bar{N}_{i,t}^z, \qquad \text{with} \qquad J_{i,t}^z \overset{i.i.d.}{\sim} N(\mu_z^J, \sigma_z^{J^2}), \tag{31}$$

$$dh_{i,t}^z = \kappa_z(\xi_z - h_{i,t}^z)dt + \sigma_z \sqrt{h_{i,t}^z} dW_{i,t}^{z,h}, \qquad \text{with} \qquad [W_i^z, W_i^{z,h}]_t = \rho_z t, \tag{32}$$

where $z \in \{f, e\}$ represents either the factor or the noise process, with their components indexed by i which belongs to either $\{1, \ldots, m^*\}$ or $\{1, \ldots, q\}$, respectively; $\bar{N}_{i,t}^z$ is a Poisson point process with arrival rate ζ_z, and $[\cdot, \cdot]$ denotes the cross variation of the arguments.

We consider three Monte Carlo designs. Common to all, we restrict ourselves to the case where the noise is a continuous Lévy process[4], with cross sectionally homogenous variance and no jumps; i.e., $h_{i,t}^e = h^e =$ Constant and $J_{i,t}^e = 0$. The factors are simulated using the following DGPs: (i) the continuous Lévy process with common variance $h_{i,t}^f = h^f$ and no jumps $J_{i,t}^e = 0$ [hereafter Lévy]; (ii) the Stochastic Volatility model without jumps $J_{i,t}^e = 0$ [hereafter SV]; and (iii) the full Stochastic Volatility Jump Diffusion model [hereafter SVJD]. The factor loadings $\beta_{i,j}$ are independent random draws from a standard Uniform distribution. A summary of the values of the parameters used in our simulations can be found in Table 1. We generate data over a range of sampling interval Δ_n and time horizons T with the number of assets set to $q = 20$ or 100 to replicate a number of different scenarios such as futures and bond curves (circa 20 variables) and the cross section of assets (circa 100 variables). In each case we set the number of factors in the data generating process to $m^* = 6$.

Our simulation experiments have three different parts. In the first one, we investigate the size of bootstrap tests for testing the number of elements in the cluster of smallest eigenvalues of IV and the number of factors supporting at least a given ratio of volatility (quadratic variation). The second

[4]We consider more general dynamics for the noise process but we do not report the results as they are similar to those included.

experiment highlights the power curves of the test for equality of eigenvalues. Finally, in our third simulation experiment, we study the power properties of the bootstrap test for the ratio of quadratic variation under different eigenvalue structures. Throughout, rejection rates are based on 10,000 Monte Carlo replications and the bootstrap critical values on 399 bootstrap samples.

6.1 Equality of eigenvalues and ratio of quadratic variation (QV)

Using the various DGPs that we defined in the previous section, in Tables 2 and 3 we report the rejection rates of our bootstrap tests for testing that: (i) the smallest $q - m^* = q - 6$ eigenvalues are equal [hereafter Equality Test, ALG 1] and (ii) the $q - m^* = q - 6$ components associated to the smallest eigenvalues explain at most $\pi = 5\%$ fraction of quadratic variation [hereafter Proportion QV Test, ALG 2], i.e. $\bar{\pi}^* = \pi = 5\%$, with $\bar{\pi}^*$ the noise to signal ratio. The null hypotheses under test are correct in all the DGPs. The rejection rates were calculated for different sampling frequency Δ_n but fixed time horizon T [see the upper panels of Tables 2 and 3] and for different T but fixed Δ_n [see the lower panels of Tables 2 and 3]. The columns of the tables report the results obtained for each of the DGP under consideration: Lévy, SV, and SVJD.

From these tables, we see that the size of the two tests are very close to 5% for almost all cases, both for $q = 20$ and $q = 100$. It is worth noticing that the size slightly increases when we increase the sampling frequency, with an inflection point at one minute. It is also interesting to note that when the asymptotic chi-squared test is used instead for testing the equality of trailing eigenvalues, its size (results are not reported, but available upon request) can reach 100% for small and moderate samples and for almost all cases, including the continuous Lévy process. Indeed, the bootstrap tests perform very well compared to the asymptotic tests.

6.2 Power curves for the test for equality of eigenvalues

Having examined the size of the bootstrap tests under the null, we now explore the rejection rates when we test for either too few or too many equal eigenvalues in the cluster of smallest eigenvalues. Testing for the equality of too many should lead to large rejection rates (as we are under the alternative). Testing for too few is consistent with the null hypothesis but the asymptotic distribution of the test statistic is not available in this case. See Remark 1. Size-correctness of the test requires that rejection rates do not exceed nominal level.

In this experiment, we use $5,000$ observations to allow for a fair comparison between asymptotic and bootstrap tests. We again set the number of factors in the DGPs to $m^* = 6$. This means that the correct number of equal trailing (smallest) eigenvalues Q is $Q^* = 14$ (for $q = 20$) and $Q^* = 94$ (for $q = 100$). Figure 1 presents the power curves for both asymptotic (chi-square) and bootstrap tests for $Q = Q^* - 6, \ldots, Q^* + 6$ ($q = 20$) and $Q = Q^* - 9, \ldots, Q^* + 6$ ($q = 100$).

The upper plots of Figure 1 illustrate the rejection rates for the asymptotic and bootstrap tests for the case when both the factors and noise are Lévy (Lévy-Lévy) processes. From this, the chi-squared test is correctly sized at 5% when testing for $Q = Q^*$ equal trailing eigenvalues. The bootstrap is slightly conservative, but only by a small degree. The bootstrap for both $q = 20$ and $q = 100$ has

Table 2: Rejection rates (in percentage) for the Bootstrap Tests when $q = 20$

Test	Equality Test			Ratio QV Test		
DGP	Levy	SV	SVJD	Levy	SV	SVJD
Δ_n	Fixed T Varying Δ_n					
5 sec	5.18	5.27	5.30	5.12	5.23	5.25
30 sec	4.93	5.08	5.11	4.94	4.98	4.97
1 min	4.71	4.95	4.88	4.75	4.87	4.79
5 min	4.87	4.88	4.93	4.83	4.88	4.94
T/Δ_n	Fixed Δ_n Varying T					
160	5.26	5.21	5.33	5.16	5.28	5.32
500	4.98	5.07	5.14	4.94	4.97	5.05
1200	4.81	4.73	4.92	4.72	4.81	4.84
2000	4.78	4.94	4.99	4.82	4.88	5.02

Table 3: Rejection rates (in percentage) for the Bootstrap Tests when $q = 100$

Test	Equality Test			Ratio QV Test		
DGP	Levy	SV	SVJD	Levy	SV	SVJD
Δ_n	Fixed T Varying Δ_n					
5 sec	5.02	5.40	5.37	5.31	5.19	5.66
30 sec	5.09	5.22	5.59	5.20	5.10	5.46
1 min	4.80	5.03	5.32	5.01	5.05	5.32
5 min	4.85	5.11	5.45	5.06	5.11	5.74
T/Δ_n	Fixed Δ_n Varying T					
160	5.23	5.48	5.80	5.39	5.92	6.81
500	5.08	5.31	5.61	5.20	5.29	5.52
1200	4.81	4.99	5.30	5.04	5.12	5.34
2000	4.89	5.02	5.15	5.02	5.05	5.21

Note: The tables report the rejection rates for the test of equality of the 14 and 94 smallest eigenvalues for $q = 20$ and $q = 100$, resp. (Equality Test) and that of the test for 6 principal factors supporting at least 95% of quadratic variation (Ratio QV Test). In the upper panels of the tables, the time horizon $T = 1$ month. In the lower panels, the time interval $\Delta_n = 5$ min and the number of observations T/Δ_n varies. The nominal level $\alpha = 0.05$.

a steeper curve than the chi-squared test, indicating more power when we test values at the right of the correct null, and is more conservative when failing to reject values at the left of the correct null. When the factors and noise are SVJD and Lévy (SVJD-Lévy) processes, respectively, the lower plots of Figure 1 show that - for both $q = 20$ and $q = 100$ - the asymptotic test does not control anymore its size, whereas the bootstrap test does and has a very good power.

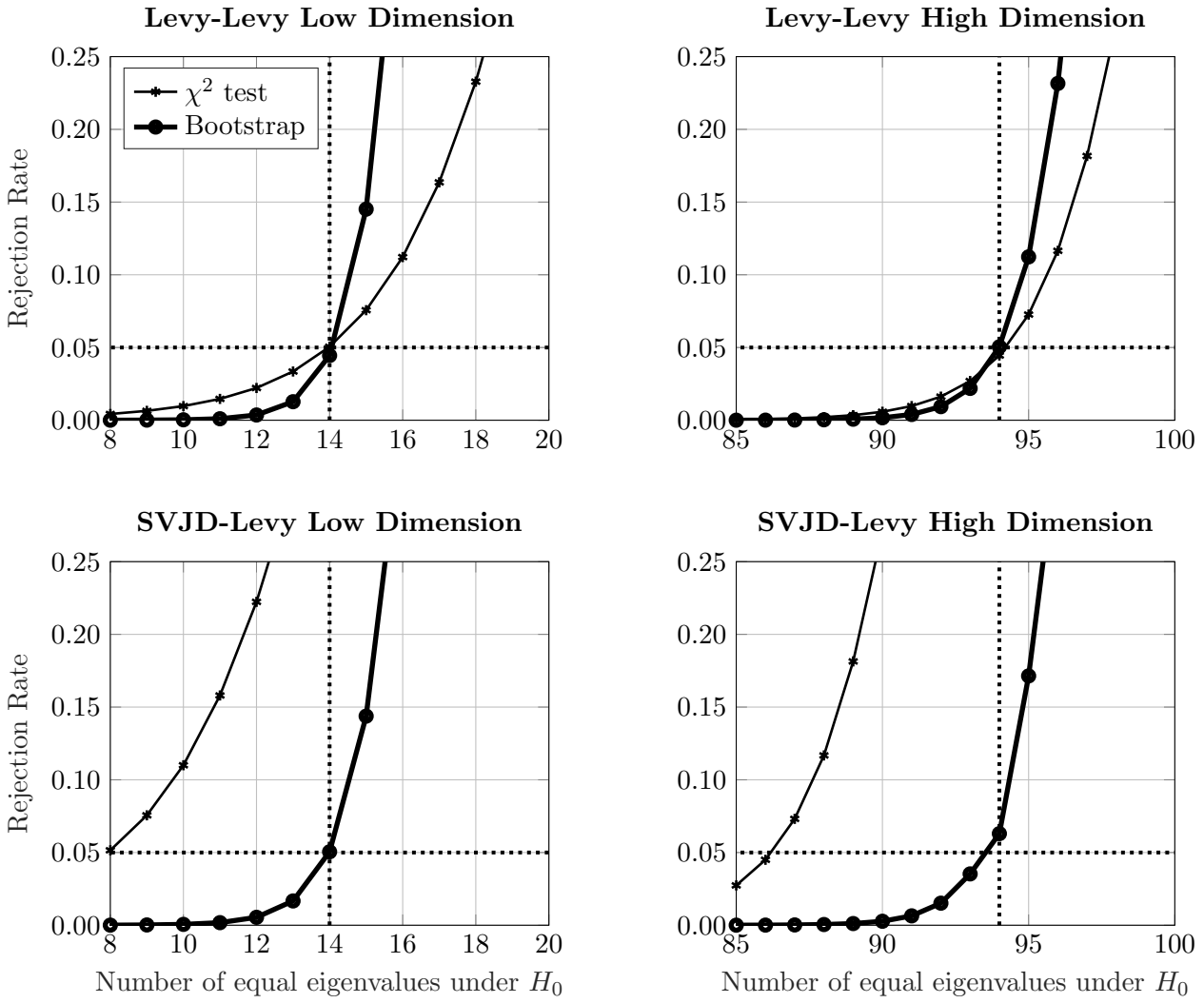


Figure 1: This figure illustrates the Size/Power comparison between asymptotic and bootstrap tests for testing the equality of the $Q = q - m$ smallest eigenvalues: $H_0 : \delta_{q-m+1} = \cdots = \delta_q$. In each DGP, the true value is $Q^* = q - m^*$ with $m^* = 6$ corresponding to the number of simulated factors (marked by the vertical dotted line). The nominal level is 5% marked by a horizontal dotted line. The data corresponds to $5,000$ observations at a five minute frequency.

6.3 Power curves for testing the ratio of quadratic variation

In this experiment, we assess the size and power of the bootstrap test for testing the ratio of quadratic variation explained by factors associated to the trailing (smallest) eigenvalues of the integrated covariance matrix. In Section 6.1, we set the noise to signal ratio $\bar{\pi}^*$ in the DGPs to be equal to the proportion π of quadratic variation explained by the trailing factors, for which we were testing the value $\pi = 5\%$. In that framework, the factor component accounted for $1 - \bar{\pi}^* = 95\%$ of the variation in the simulated prices and noise for $\bar{\pi}^* = 5\%$, and we tested for the trailing eigenvalues explaining at most $\pi = 5\%$ of the quadratic variation. That simulation framework, however, was chosen to check the size of the test under a straightforward eigenvalue structure.

In this section, we look at a series of cases that illustrate a more complicated eigenvalue structures

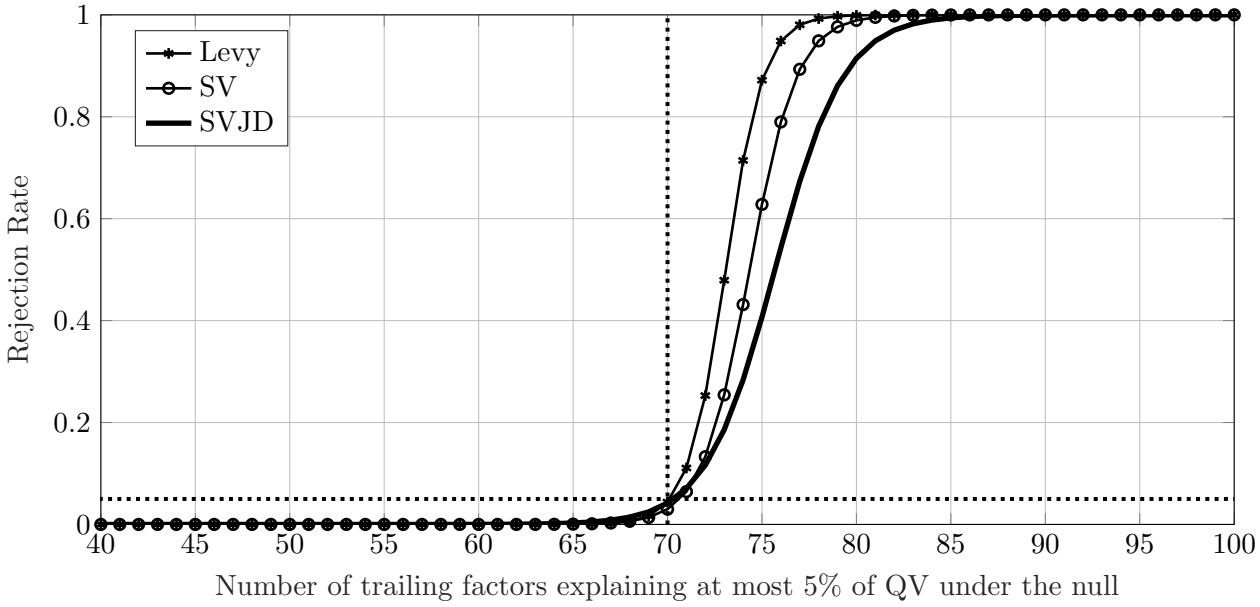


Figure 2: This figure illustrates the Size/Power for Case 1; i.e., rejection rates when using bootstrap to test for the number of trailing eigenvalues on the abscissa axis that explain at most $\pi = 0.05$ fraction of the quadratic variation. The frequency of sampling is $\Delta_n = 5$ min and T/Δ_n is set to $5,000$ observations. The DGP is driven by 30 factors, with 6 dominant factors (explaining 85% of the variation) and 24 minor factors explaining another 10%, with the noise explaining 5%. The null hypothesis under the DGP is correctly specified when the number of trailing eigenvalues is equal to 70. The three plot lines refer to the three DGP types for the factor model Lévy, stochastic volatility (SV) and stochastic volatility jump diffusion (SVJD). For each case we use a Lévy noise process.

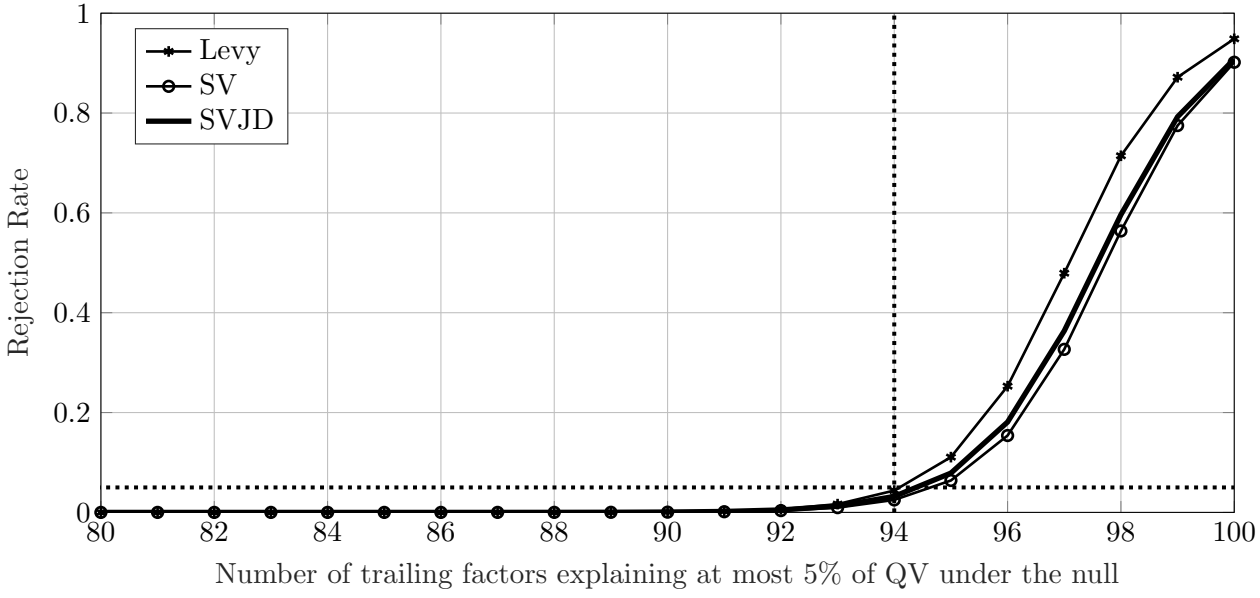


Figure 3: This figure illustrates the Size/Power for Case 2. The simulation conditions are the same as in Figure 2, except that here the first six dominant factors now explain 95% of the quadratic variation, the remaining 24 factors explain 4%, and the noise accounts for 1%. As such the null hypothesis is correctly specified when the number of trailing eigenvalues in the DGP under consideration is 94.

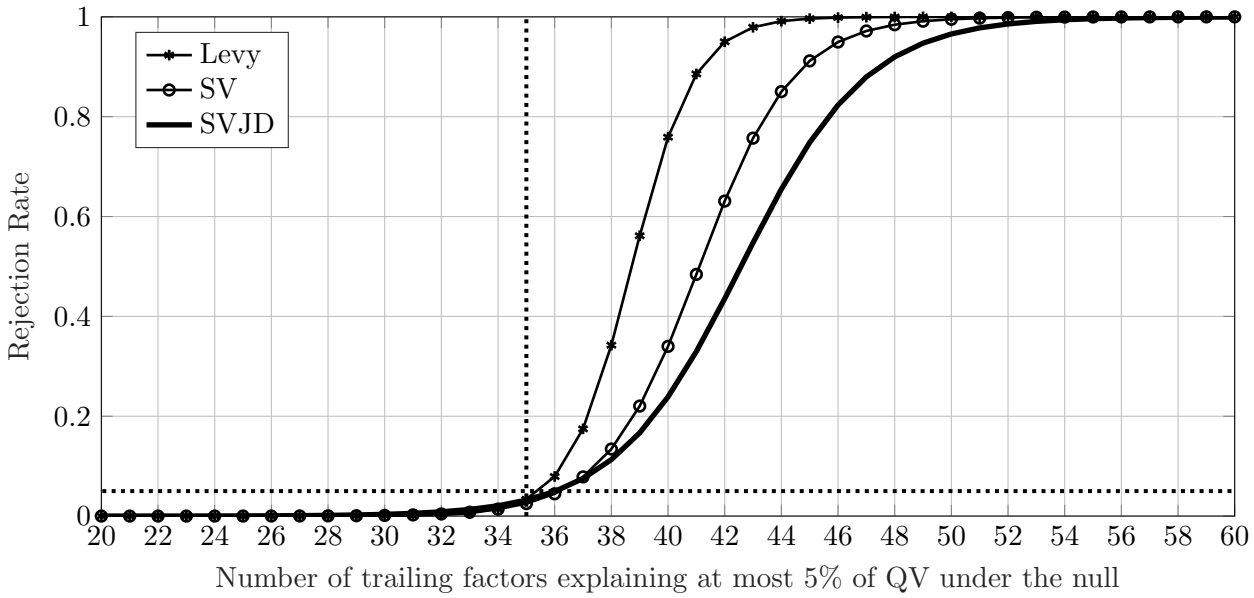


Figure 4: This figure illustrates the Size/Power for Case 3. The simulation conditions are the same as in Figure 2 and 3, except that the first six factors in the simulation explain 85% of the quadratic variation, the remaining 24 factors explain 5% of the quadratic variation and the noise 10% of the quadratic variation. As such, the null hypothesis is correctly specified when the number of trailing factors explaining at most 5% of the quadratic variation is equal to 35.

that we build into our DGPs. For this, we focus on $q = 100$ and presume that the DGPs are representing a broad cross section of assets, such as the S&P 100. We follow, in part, the simulation settings in Aït-Sahalia and Xiu (2017) to build three blocks into the simulated eigenvalue structure. Overall, in each of the three cases described below we have 6 'dominant' factors explaining a large fraction of the quadratic variation, then 24 further 'weak' factors that have cross sectionally homogenous variance and share only a fraction of the quadratic variation to make a total of 30 factors. We then have 70 trailing eigenvalues with associated factors explaining a small, but nonzero fraction of the quadratic variation. These structures were built by adding a constant to each element of the factor loading $\beta_{i,j}$, such that the resulting fractions of quadratic variation deliver the proportions of quadratic variation explained by each block in the following cases:

Case 1: The first six factors in the simulation explain 85% of the quadratic variation, the remaining 24 factors explain 10% of the quadratic variation and the noise 5% of the quadratic variation. Thus, test should reject at 5% when the number of trailing factors explaining at most 5% of the variation is equal to 70.

Case 2: The first six factors in the simulation explain 95% of the quadratic variation, the remaining 24 factors explain 4% of the quadratic variation and the noise 1% of the quadratic variation. Thus, test should reject at 5% when the number of trailing factors explaining at most 5% of the variation is equal to 94.

Case 3: The first six factors in the simulation explain 85% of the quadratic variation, the remaining 24 factors explain 5% of the quadratic variation and the noise 10% of the quadratic variation. Thus,

the test should reject when there are 35 trailing eigenvalues.

We simulate the factors described above using the 3 DGPs Lévy, SV and SVJD with Lévy noise in each case. We set $\Delta_n = 5$ min and $T/\Delta_n = 5,000$ observations. The rejection rates (size and power) for testing the proportion of unexplained variance under these different eigenvalue structures are reported in Figures 2 to 4, which correspond to each of the above 3 cases, respectively.

Figure 2 illustrates the rejection rates to the left and right of the null given the actual DGP. For Case 1, the number of trailing factors that explain at most $\pi = 5\%$ of the quadratic variation is 70. From this figure, we see that the bootstrap test is correctly sized for all DGPs and has a good power. To the right of the null, we see that the bootstrap test has much more power when the factors are generated by Lévy process, followed by SV and SVJD processes.

Figure 3 shows the empirical rejection rates when the eigenvalue structure corresponds to 6 dominant factors that explain 95% of the quadratic variation within the sample and 24 remaining factors explain only 4% of the quadratic variation (Case 2). Similarly to the results in Case 1, the bootstrap test is correctly sized when the number of trailing factors explaining at most 5% is equal to 94. Furthermore, the test has a reasonable power, with some difference in the power depending on the DGP under consideration: the test reaches a higher power for Lévy process, followed by SVJD and SV processes.

Finally, Figure 4 illustrates some of the difficulties in identifying the eigenvalue structure when quadratic variation is spread most evenly. In Case 3, 85% of the variation in the data is explained by 6 dominant factors with the remaining 15% split between 24 "weak" factors (5%) and the noise (10%), characterized by 70 equal eigenvalues. For this case, the test is under the correct null when the number of trailing eigenvalues is equal to 35. The results in Figure 4 show that the bootstrap test is doing well in terms of size control, but is slightly conservative for the SVJD and SV processes when testing for 35 trailing eigenvalues.

7 Empirical study

We now conduct a two part empirical exercise to illustrate the usefulness of one of our bootstrap tests as a guide to modelling financial time series data. The objective is to generate a series of candidate factors from a subset of actively traded stocks. We then use these factors to extract the systematic returns from a large cross section of returns from US equity market securities.

Data and methodology: We make use of two data sets for this empirical study: (i) the historical record of tick-by-tick best-bid and best-offer data for 597 members of the S&P 500 traded between January 1, 1996 to the end of the week of April 6, 2020 from the SIRCA-Thomson Reuters data files and (ii) the daily and monthly CRSP returns data file for US stocks with 21,965 firms reporting returns between January 1, 1996 to December 31, 2019 (the latest available sample at the time of writing this section).

For the tick-by-tick data set, we collect the data into trading weeks (Monday to Friday) and trading months (first to the last day of the month). Trading weeks are checked against historical holidays and shutdowns (such as September 11, 2001). For all of the available stocks for a given trading week or

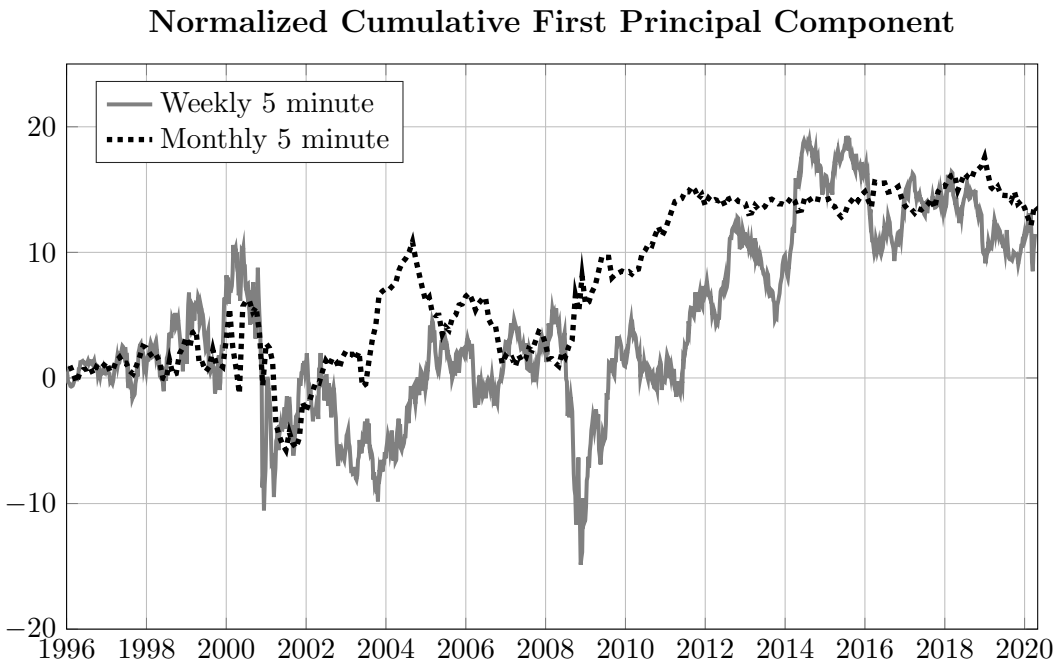


Figure 5: Comparison of the weekly and monthly cumulative first principal component normalized by their standard deviations for plotting comparison.

month, the mid price is computed from the best-bid and best-offer after June 2005. The stocks are next sorted by the total number of price changes over the week or month. The 100 most actively traded are then selected for the covariance estimation and PCA analysis.

The tick-by-tick prices, observed at specific times stamps, are sampled to an equidistant five minute grid over the trading day that goes from 09:30 to 16:00 Eastern Time, and then used to calculate returns.

Our analysis is then performed as follows. For the tick-by-tick data, we first remove: (i) all zero prices and instant reversions from the bid and ask series; (ii) any records where the standing bid price is higher than the standing offer price; and (iii) any records where the bid/ask price is more than 500% different from the daily median bid-ask. We next use the bid-ask tick series to compute the tick-by-tick mid price/return and record the tick times. Weekly and monthly covariance matrices are then estimated after we removed jumps using the threshold $3(\mathrm{BV}/T)^{0.5}\Delta_n^{0.47}$, where BV is the estimated bipower variation.

Thereafter, we apply our bootstrap-based test to the weekly and monthly returns to determine the number of components that explains at least the $1-\pi$ proportion of quadratic variation for the data set, with π set to 5%. We use 399 bootstrap replications and sequentially increase the number of components from the largest eigenvalue downwards and stop when the test fails to reject null at the level $\alpha = 5\%$.

For each week and month, we collect the Q eigenvalues indicated by the quadratic variation test and their corresponding eigenvectors. The eigenvectors are then used to construct factor components as in Chen et al. (2019). Figure 5 presents the cumulative returns for the first principal components for weekly and monthly returns. From this, on the one hand, we see that the pattern of the weekly

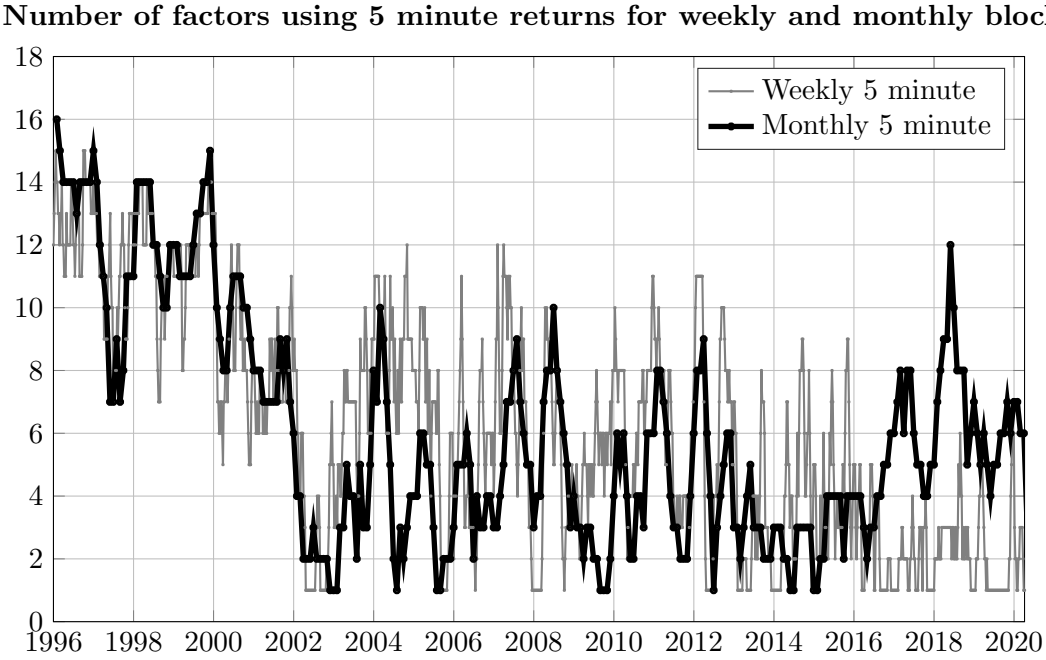
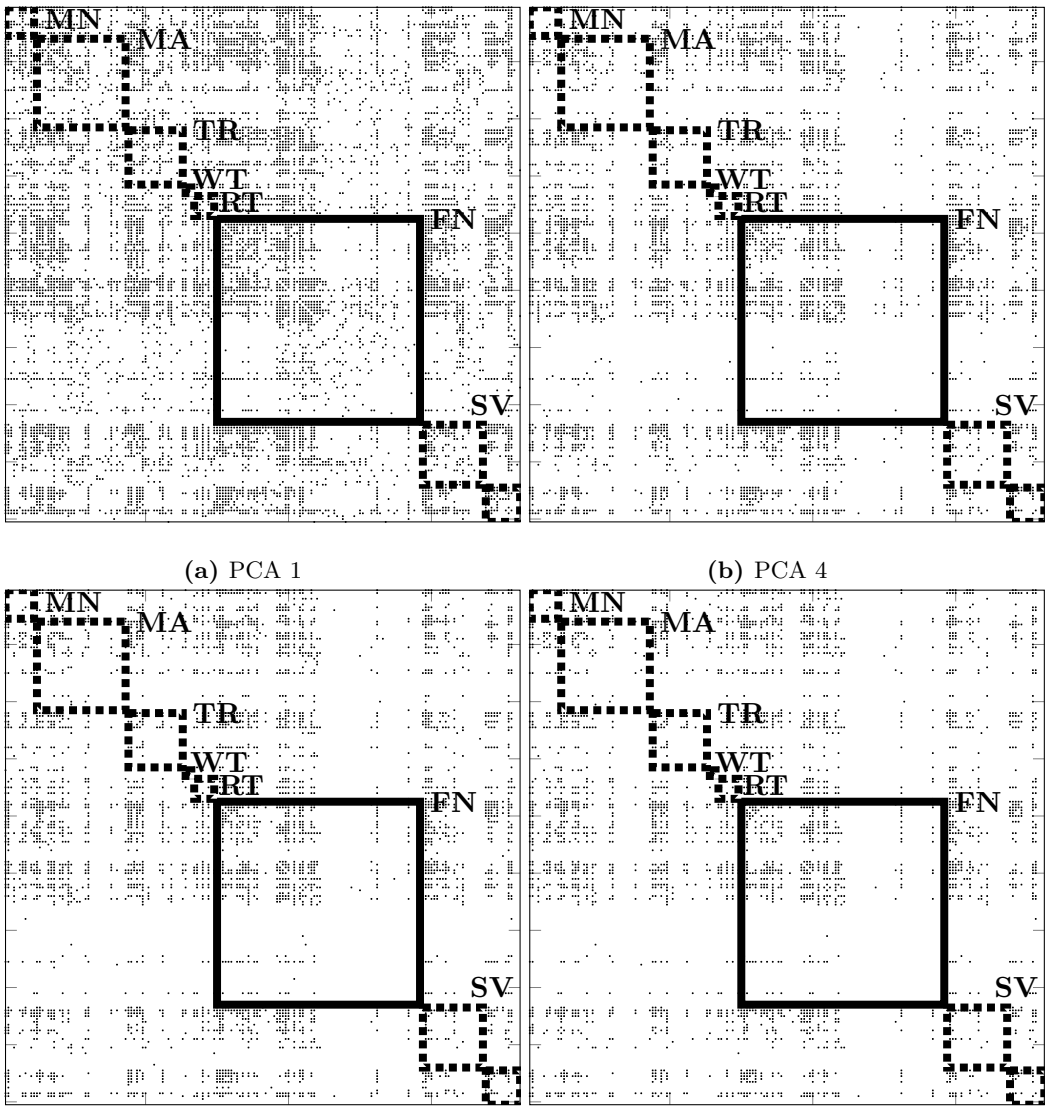


Figure 6: This figure illustrates the number of factors selected each week (month) using our ratio test, with a target proportion of quadratic variation explained within the cross section of high frequency returns being 95%.

series is quite consistent with the results found in Chen et al. (2019) (over the period from 2004 to 2014) and in Aït-Sahalia and Xiu (2019) (over the period from 2003 to 2013) who argue that the first principal component shares the time series features of the market return. On the other hand, the monthly pattern is substantially smooth and, interestingly, the short positions over the 2008-2010 period appear sufficient to smooth over the evident market dip for the first component cumulative returns we obtained using weekly data.

Results: Using our bootstrap-based test, Figure 6 presents the weekly and monthly time series of the number of principal components needed to explain at least 95% ($\pi = 0.05$) of quadratic variation over one-week and one-month-long periods, respectively. There are several interesting points that are worth commenting on. First, the weekly pattern of the number of principal components over the 2003-2013 period is substantially similar to the one found in Aït-Sahalia and Xiu (2017) as we notice a rise in the number of components prior to 2009 and a fall after the commencement of the financial crisis as a single factor dominates. However, the longer time frame of our study provides additional insights. During the time-period of 1996–2002, we detect a larger number of components for both weekly and monthly data. Intriguingly, this is prior to the national market service (regulation NMS) implemented in 2005. Indeed, the number of price changes recorded for the best-bid and best-ask time series is substantially lower prior to 2002. Furthermore, as noted in Chen et al. (2019), asynchronous trading due to stale prices in the national best-bid and best-ask (which did not formally exist prior to 2005) can add significant levels of idiosyncratic noise to the covariance estimation. Similar effects are mentioned in Aït-Sahalia and Xiu (2017) who argue that subsampling does not fully mitigate the effect of microstructure noise.



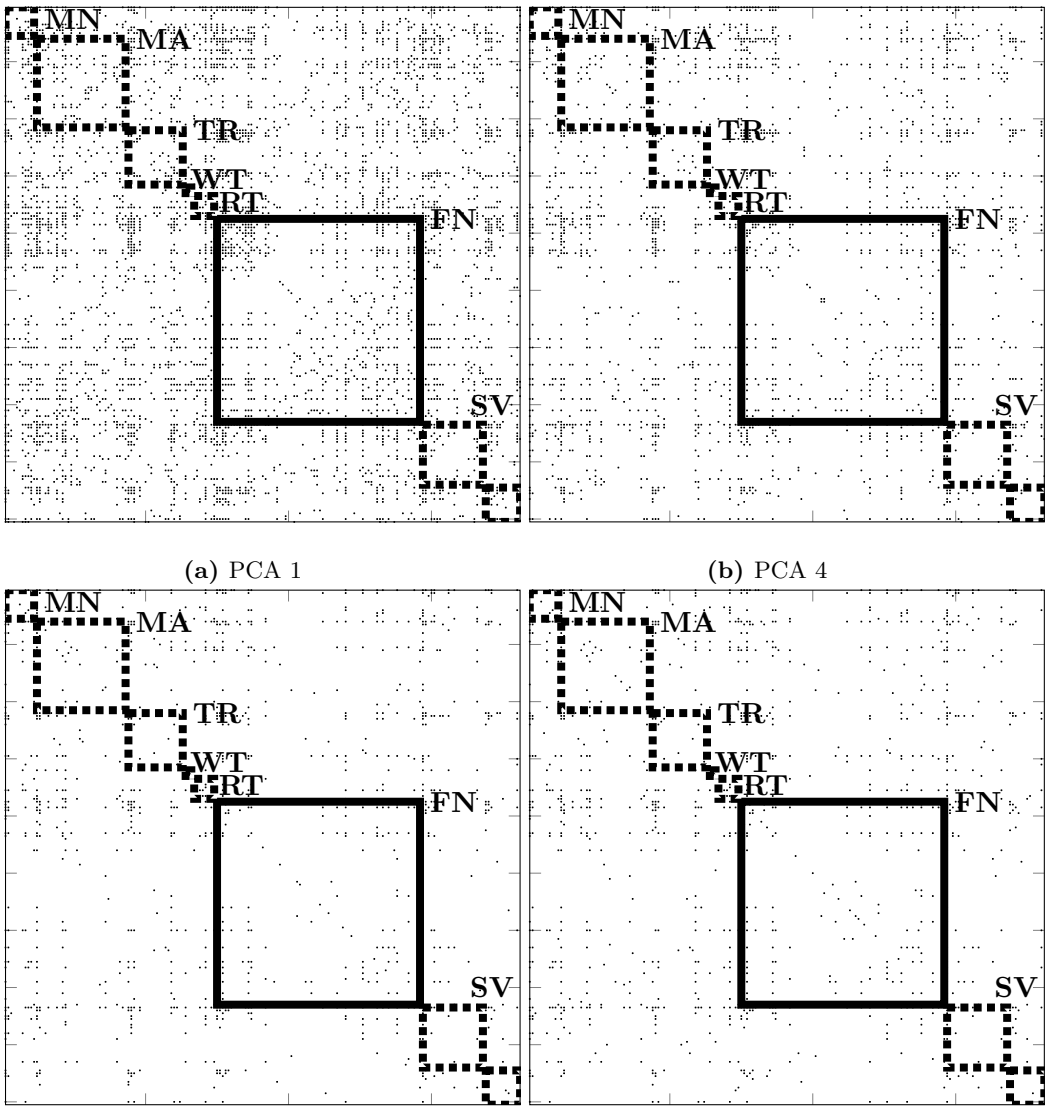
(a) PCA 1 **(b)** PCA 4

(c) PCA 11 **(d)** PCA 14

Figure 7: This figure illustrates the matrix of the tests for significant pairwise correlations between pricing errors from the regression of weekly excess returns regressed on to the principal components determined from five minute return covariance matrices as pricing factors, that is $\boldsymbol{\varepsilon}_i = \boldsymbol{r}_{i,weekly} - \boldsymbol{\beta}_i' \boldsymbol{F}_{weekly}$. We exclude stocks with less than 75 available weekly returns. The dots in this figure indicate the rejection of the null hypothesis of no pairwise-correlation. In each subfigure, the total number of tested pairwise correlations is 42,822,885 i.e. the number of off-diagonal correlations ($(N^2 - N)/2$, with $N = 9255$). We set the p-value to 0.05/42,822,885, which yields a critical value of around 6.22 for the t-test. The squares highlight several sectors, notably Mining (MN), Manufacturing (MA), Transport (TR), Retail (RT), Finance (FN) and Services (SV).

Having built a set of weekly and monthly factors from the principal component analysis and identified the candidate number of factors based on our test, we now use the latter factors to explaining the cross sectional variation of a large number of stocks from the CRSP return dataset.

For comparison with the weekly factors, we aggregated the daily CRSP stock returns to weekly returns for the available firms in the US stock market. CRSP monthly returns are collected in the



(a) PCA 1 **(b)** PCA 4

(c) PCA 11 **(d)** PCA 14

Figure 8: Significant pairwise correlations for our Principal Component Factors derived from covariance matrices computed monthly using five minute returns. For the definition of the points in the plot see the caption in Figure 7.

standard way. For the second stage of the analysis, a stock must have at least six years of daily holding returns (approximately one quarter of the targeted sample size) to be kept, which leaves us with $9,255$ stocks over the sample period January 1, 1996 to December 31, 2019. For each stock i, we run the following time series regression:

$$r_{i,t}^m = \boldsymbol{\beta}_i' \boldsymbol{f}_t^m + e_{i,t}, \quad \text{for } i = 1, \ldots, 9255, \tag{33}$$

where $r_{i,t}^m$ is the weekly or monthly CRSP stock return and $\boldsymbol{f}_t^m$ is the cumulative return over a week or month for the vector of factors, respectively. We next compute the pairwise covariances between residuals and record this in a matrix $\boldsymbol{S} = [\text{cov}(e_{i,t}, e_{j,t})]_{i,j}$. Notice that because the elements of $\boldsymbol{S}$ are

constructed pairwise, the matrix $\boldsymbol{S}$ is not guaranteed to be positive semi-definite. It is inherently a reduced rank matrix as the number of variables is larger than the number of observations. We next calculate the pairwise correlations between residuals $\rho_{ij} = \text{cov}(e_{i,t}, e_{j,t})/\sqrt{var(e_{i,t})var(e_{j,t})}$, for $i \neq j$ and $i, j = 1, \ldots, 9255$, and record this in a matrix S^{ρ}. An adjusted standard t-test is then used to test $H_0 : \rho_{ij} = 0$ against $H_1 : \rho_{ij} \neq 0$, for all i, j, and in a sized matrix M we record 1 if the null is rejected at 5% significance level, and 0 otherwise. Following Aït-Sahalia and Xiu (2017), we apply pair-wise sphericity tests and adjust the individual p-values to match the joint power of these tests; see Onatski et al. (2013) for an example in a similar context. In this instance, 9,255 stocks yield 42,822,885 pairwise correlations.

Thereafter, we construct matrix plots with points representing the significance of the pairwise correlation. The stocks are sorted as follows. First, they are placed into sector bins following the example in Aït-Sahalia and Xiu (2017) and second, they are sorted (largest to smallest) by the absolute value of the weighting of the stock in the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{S}$. This approach follows in the tradition of Cochrane (1996); Gomes, Kogan, and Zhang (2003); Jagannathan and Wang (2007); Cooper, Gulen, and Schill (2008). However, we use individual stock returns in the spirit of Ang, Liu, and Schwarz (2020), with the objective of assessing to what extent do the factors from the high frequency PCA mimic the well established factors in orthogonalizing a very large cross section of lower frequency returns.

Subplots (a) to (d) of Figures 7 and 8 graphically report our 42,822,885 pair-wise correlation tests for 1, 4, 11 and 14 factors, respectively. These plots are slightly different from Aït-Sahalia and Xiu (2017) who use a fixed threshold to illustrate the degree of pairwise residual correlation for the assets within the sample used to compute the factors. Similarly to Aït-Sahalia and Xiu (2017), the blocks represent the top level industrial sectors and these are marked by a mnemonic. When residual return correlation is significant, the location is represented by a marker.

The first point to note here is that, consistent with Aït-Sahalia and Xiu (2017), increasing the number of weekly or monthly factors does reduce the density of residual correlation. However, this reduction is not the same for the weekly and monthly returns data sets. In Figure 7, we observe that even for 14 weekly factors there is a fair amount of residual correlation and the latter does have some structure with certain groups of firms clearly having bands outside their industry boxes.

However, for factors extracted from monthly five minute data, the visible density of residual correlation drops markedly and the pattern does not have any substantive banding or clustering. Indeed, with 14 monthly factors there is no discernible pattern to the location of the significant correlations. The sparse covariance structure left in the residuals within given industries might, however, be explained by the fact that our factors are global factors and do not capture all industry specific characteristics. We conclude that generating 14 approximate factors, obtained by PCA, from high frequency return data for a subset of 100 most actively traded stocks will effectively capture the systematic variation of the weekly and monthly holding returns for the available cross section of CRSP stocks.

8 Conclusion

This paper introduces a testing framework for the eigenvalue structure of the integrated variance matrix (IV) and the spot variance matrix (c_t) of stock price vector processes represented by an Itô semimartingale dynamics. Likelihood ratio-type tests are proposed for the equality of clusters of eigenvalues of these matrices as well as their related correlation matrices. Unlike the existing approaches that are valid only in settings with large cross-section dimension, our tests do not require large cross-section and thus they are useful in a wide variety of applications. Our tests are shown to be useful to principal component analysis of the price vector process based on c_t or IV. More specifically, a test is proposed that detects the number of principal components or factors sufficient to capture at least a certain prespecified proportion of variation in the data using c_t or IV as dispersion measure. Further, our tests are also useful to test some special factor structures of the price process, especially those factor structures that translate into the equality of the smallest eigenvalues of c_t or IV.

We derive the asymptotic distributions (under the null) of our test statistics and find that they are, in general, non-standard with many nuisance parameters. Another main contribution of this paper consists in proposing some variant of the blocks of blocks bootstrap to approximate these asymptotic distributions. The proposed bootstrap procedures do not require the estimation of many nuisance parameters, they provide a better test than the asymptotic approximation, and are simple to implement. Their first-order asymptotic validity is established.

The finite sample properties of the asymptotic and bootstrap tests have been investigated by an extensive Monte Carlo simulation study where several data generating processes have been considered as well as different sampling frequencies and small and large cross-section dimensions. The results reveal that the bootstrap tests are correctly sized and has power in determining the eigenvalue structure.

We illustrate an application of our tests for factor construction from the 100 most actively traded constituents of the S&P 500 index. We then use the number of principal components selected by the test in a cross sectional model for all traded stocks in the CRSP datafile. Analysis of the pairwise correlation of the resulting residuals suggests that these principal components are viable pricing factors.

A Assumptions, Lemmas and Proofs

A.1 Assumptions

Assumption A.1. *X satisfies Assumption (H-r) for some $r \in [0,2)$ and the process σ satisfies (H-2); $k_n\sqrt{\Delta_n} \to 0$ and $k_n\Delta_n^\tau \to \beta \in (0,\infty)$ for some $\tau \in (0,1/2)$; $r < 2/(1+\tau)$ and $\varpi > \tau/[2(2-r)]$.*

A.2 Lemmas: Law of large numbers

Let $\Delta_i^n X = X_{i\Delta_n} - X_{(i-1)\Delta_n}$, and (k_n) and (v_n) two sequences of integers. Let f be a real-valued function defined on $(\mathbb{R}^d)^\ell$ and let

$$\hat{f}_i(k_n, v_n) = \frac{1}{k_n} \sum_{m=0}^{k_n-\ell} f\left(\frac{\Delta_{i+m}^n X}{\sqrt{\Delta_n}}, \frac{\Delta_{i+m+1}^n X}{\sqrt{\Delta_n}}, \ldots, \frac{\Delta_{i+m+\ell-1}^n X}{\sqrt{\Delta_n}} \right) \prod_{j=0}^{\ell-1} 1_{\{\|\Delta_{i+m+j}^n X\| \le v_n\}}. \tag{A.1}$$

For $t \in ((i-1)\Delta_n, i\Delta_n]$, define $\hat{f}(k_n, v_n, t) = \hat{f}_{i+1}(k_n, v_n)$.

Lemma A.1. *Assume (i) X satisfies assumption (H-r) for some $r \in [0,2]$, (ii) k_n satisfies: $k_n \to \infty$ and $k_n\Delta_n \to 0$ as $n \to \infty$, (iii) $v_n = \alpha\Delta_n^\varpi$ for some $\alpha > 0$ and $\varpi \in (0, 1/2)$, and (iv) f is a continuous real-valued function on $(\mathbb{R}^d)^\ell$ which satisfies*

$$|f(x_1, \ldots, x_\ell)| \le \prod_{j=1}^{\ell} \Psi(\|x_j\|)(1 + \|x_j\|^p) \tag{A.2}$$

for $p \ge 0$ and Ψ is a continuous function on $[0,\infty)$ which goes to 0 at infinity.

Then, when X is continuous, or when X jumps and either $p \le 2$ or

$$p > 2, \quad 0 < r < 2, \quad \varpi \ge \frac{p-2}{2(p-r)}, \tag{A.3}$$

we have, for any fixed time $t \ge 0$,

$$\hat{f}(k_n, v_n, t) \xrightarrow{P} \rho_{c_t}^{\otimes\ell} f,$$

where $\rho_{c_t}^{\otimes\ell} f = Ef(Z_1, \ldots, Z_\ell)$, with $Z_i \sim i.i.d. N(0, c_t)$.

This lemma covers in particular cases where the function f is continuous and bounded or has a polynomial growth, that is:

$$|f(x_1, \ldots, x_\ell)| \le K \prod_{j=1}^{\ell} (1 + \|x_j\|^p), \tag{A.4}$$

for $p \ge 0$ and a constant $K > 0$. Note that a function satisfying (A.4) also satisfies (A.2) for a value $p' = p + \varepsilon$ for any $\varepsilon > 0$. The cost of the application of this lemma to such a function is that when X jumps and $p > 2$, the range of ϖ warranting the stated convergence is $1/2 > \varpi \ge (p+\varepsilon-2)/(2(p+\varepsilon-r))$, for some $\varepsilon > 0$. This can be made sharper as shown by the following lemma.

Lemma A.2. *The statement in Lemma A.1 holds under the same conditions but with* (A.2) *replaced by:* (A.4).

Proof of Lemma A.1: Our proof follows the lines of that of Jacod and Protter (2012, Th. 9.3.2). By the standard localization procedure (see Jacod and Protter, 2012, p.114), the strengthened assumption (SH-r) below will be relied upon instead of (H-r):

Assumption (*SH-r*). *We have (H-r), and the processes b and σ are bounded, and $\|\delta(\omega, t, z)\| \wedge 1 \le \Gamma(z)$ with Γ bounded and $\int \Gamma(z)^r \lambda(dz) < \infty$.*

Let $i_n = i+1$, with i such that $t \in I(n,i) = (i\Delta_n, (i-1)\Delta_n]$. Let $t_n = (i_n - 1)\Delta_n$, $Y_t^n = \sigma_{t_n}(W_t - W_{t_n})1_{\{t_n \le t\}}$ and $Y_t'^n = \int_{t_n \wedge t}^t (\sigma_s - \sigma_{t_n}) dW_s$. By construction and thanks to the càdlàg property of c_s, $c_{t_n} \to c_t$ as $n \to \infty$.

Step 1: We show that

$$\hat{f}_{i+1,Y^n} \equiv \frac{1}{k_n} \sum_{m=0}^{k_n-l} f\left(\frac{\Delta_{i+1+m}^n Y^n}{\sqrt{\Delta_n}}, \frac{\Delta_{i+2+m}^n Y^n}{\sqrt{\Delta_n}}, \ldots, \frac{\Delta_{i+\ell+m}^n Y^n}{\sqrt{\Delta_n}} \right) \xrightarrow{P} \rho_{c_t}^{\otimes\ell} f.$$

By definition, we have $\hat{f}_{i+1,Y^n} = \frac{1}{k_n}\sum_{m=0}^{k_n-\ell} f(y_{i+1+m}, y_{i+1+m}, \ldots, y_{i+\ell+m})$, with $y_k \sim i.i.d.N(0, c_{t_n})$. It is not hard to see that

$$E\left(\hat{f}_{i+1,Y^n}\right) = \frac{k_n-\ell+1}{k_n}\rho_{c_{t_n}}^{\otimes\ell} f \to \rho_{c_t}^{\otimes\ell} f.$$

Also, $Var(\hat{f}_{i+1,Y^n}) = O(1/k_n) \to 0$ as $n \to \infty$. We conclude that $\hat{f}_{i+1,Y^n}$ converges to $\rho_{c_t}^{\otimes\ell} f$ in quadratic mean and therefore in probability.

Step 2: We now show that:

$$Z_n \equiv \hat{f}(k_n, v_n, t) - \hat{f}_{i+1,Y^n} = \hat{f}_{i+1}(k_n, v_n) - \hat{f}_{i+1,Y^n} \xrightarrow{P} 0. \tag{A.5}$$

For this, we will use the following inequality established in Step 3 bellow. For all $v \geq 1$, $\varepsilon \in (0,1]$ and $A > 0$ large enough, with $a_j \equiv x_j + y_j + z_j + w_j$; $j = 1, \ldots, \ell$, we have:

$$\begin{aligned}
&\left|f(a_1,\ldots,a_\ell)\prod_{j=1}^{\ell} 1_{\{|a_j|\leq v\}} - f(a_1,\ldots,a_\ell)\right| \\
\leq\ & \theta_A(\varepsilon) + K\Psi(A)\prod_{j=1}^{\ell}\left(1 + \|x_j\|^p + \|y_j\|^p \wedge v^p + \|z_j\|^p + \|w_j\|^p\right) \\
&+ KA^{2\ell}\sum_{j=1}^{\ell}\varepsilon^{-p}\left(\|y_j\|^p \wedge v^p + \|z_j\|^p + \|w_j\|^p\right) \\
&+ \sum_{k=1}^{\ell}\left[\prod_{j=1,j\neq k}^{\ell}\left(1+\|x_j\|^p\right)\right]\left[\frac{\|x_k\|^{p+2}}{v^2} + \|y_k\|^p \wedge v^p + \|z_k\|^p + \|w_k\|^p\right],
\end{aligned} \tag{A.6}$$

where $\theta_A(\varepsilon)$ is a positive valued function converging to 0 as $\varepsilon \to 0$; $\Psi(A) \to 0$ as $A \to \infty$ and $K > 0$ a generic constant.
Now we consider the decomposition of the process X given by Eq. (9.2.7) of Jacod and Protter (2012), that is: $X = X' + X''$ with $X' = X_0 + \int_0^t b''_s ds + \int_0^t \sigma_s dW_s$ where $b''_t = b_t + \int_0^t \delta(t,z)1_{\{\|\delta(t,z)\|>1\}}\lambda(dz)$.

Write $x_i = \Delta_{i_n+i}^n Y^n/\sqrt{\Delta_n}$, $y_i = \Delta_{i_n+i}^n X''/\sqrt{\Delta_n}$, $z_i = \Delta_{i_n+i}^n Y'^n/\sqrt{\Delta_n}$ and $w_i = \Delta_{i_n+i}^n B''/\sqrt{\Delta_n}$, with $B''_t = \int_0^t b''_s ds$.

Using (A.6), we obtain for some constant $K > 0$,

$$|Z_n| \leq \theta_A(\varepsilon) + \tfrac{K}{k_n}\sum_{m=0}^{k_n-\ell}\left(\Psi(A)Z^1_{n,m} + A^{2\ell}\varepsilon^{-p}Z^2_{n,m} + Z^3_{n,m}\right), \tag{A.7}$$

with:

$$\begin{aligned}
Z^1_{n,m} &= \prod_{j=1}^{\ell}\left(\|x_{m+j-1}\|^p + \|y_{m+j-1}\|^p \wedge \Delta_n^{p(\varpi-1/2)} + \|z_{m+j-1}\|^p + \|w_{m+j-1}\|^p\right), \\
Z^2_{n,m} &= \sum_{j=1}^{\ell}\left(\|y_{m+j-1}\|^p \wedge \Delta_n^{p(\varpi-1/2)} + \|z_{m+j-1}\|^p + \|w_{m+j-1}\|^p\right) \\
Z^3_{n,m} &= \sum_{s=1}^{\ell}\left[\prod_{j=1,j\neq s}^{\ell}\left(1+\|x_{m+j-1}\|^p\right)\right]\times \\
&\qquad \left[\tfrac{\|x_{m+s-1}\|^{p+2}}{\Delta_n^{2\varpi-1}} + \|y_{m+s-1}\|^p \wedge \Delta_n^{p(\varpi-1/2)} + \|z_{m+s-1}\|^p + \|w_{m+s-1}\|^p\right].
\end{aligned} \tag{A.8}$$

We have $\|y_{m+j-1}\|^p \wedge \Delta_n^{p(\varpi-1/2)} = \Delta_n^{p(\varpi-1/2)}\left[\left\|\frac{\Delta^n_{i_n+m+j-1}X''}{\Delta_n^{\varpi}}\right\|^p \wedge 1\right]$. Similar to the arguments of Jacod and Protter (2012, p.259), we can claim using their Eqs. (2.1.33), (2.1.34) and (2.1.45) that: for all i: $0 \leq i \leq k_n - 1$,

$$\|w_i\|^p = \left\|\frac{\Delta^n_{i_n+i}B''}{\sqrt{\Delta_n}}\right\|^p \leq K\Delta_n^{p/2}, \quad E\left(\|\Delta_{i_n+i}Y^n\|^p|\mathcal{F}_{(i_n+i-1)\Delta_n}\right) \leq K\Delta_n^{p/2},$$

$$E\left(\|\Delta_{i_n+i}Y'^n\|^p|\mathcal{F}_{(i_n+i-1)\Delta_n}\right) \leq K\Delta_n^{p/2}E(\gamma_n|\mathcal{F}_{(i_n+i-1)\Delta_n}) \leq K\Delta_n^{p/2},$$

where γ_n is a bounded random sequence converging almost surely to 0 and finally,

$$E\left(\left\|\frac{\Delta^n_{i_n+i}X''}{\Delta_n^{\varpi}}\right\|^p \wedge 1\middle| \mathcal{F}_{(i_n+i-1)\Delta_n}\right) \leq K\Delta_n^{p(1/2-\varpi)}\phi_n, \text{ (if } p\leq 2), \text{ and } \leq K\Delta_n^{1-r\varpi}\phi_n, \text{ (if } p>2),$$

where $\phi_n \to 0$ as $n\to\infty$. Hence:

$$E\left(\|y_i\|^p \wedge \Delta_n^{p(\varpi-1/2)}|\mathcal{F}_{(i_n+i-1)\Delta_n}\right) \leq K\phi_n, \text{ (if } p\leq 2), \text{ and } \leq K\Delta_n^{1-r\varpi+p(\varpi-1/2)}\phi_n, \text{ (if } p>2).$$

Then, under (A.3), $E\left(\|y_i\|^p \wedge \Delta_n^{p(\varpi-1/2)}|\mathcal{F}_{(i_n+i-1)\Delta_n}\right) \leq K\phi_n$ for all $p>0$.

By successive applications of the law of iterated expectations, we can see that:

$$\begin{aligned} E(Z^1_{n,m}) &\leq K, \\ E(Z^2_{n,m}) &\leq K\left(\phi_n + \Delta_n^{p/2} + E(\gamma_n)\right) \\ E(Z^3_{n,m}) &\leq K\left(\Delta_n^{1-2\varpi} + \phi_n + \Delta_n^{p/2} + E(\gamma_n)\right). \end{aligned}$$

This shows that $E|Z_n| \to 0$ by first letting $n\to\infty$ and then $A\to\infty$ and $\varepsilon\to 0$. As a result, $Z_n \to 0$ in probability.

Step 3: It remains to show (A.6). We have the following inequality which is a slight extension of Jacod and Protter (2012, Eq. (8.4.21)). For all $\varepsilon\in(0,1]$, A large enough and $x_i, y_i \in \mathbb{R}^d$,

$$\begin{aligned} & |f(x_1+y_1,\ldots,x_\ell+y_\ell) - f(x_1,\ldots,x_\ell)| \\ \leq\ & \theta_A(\varepsilon) + K\Psi(A)\textstyle\prod_{j=1}^\ell(1+\|x_j\|^p+\|y_j\|^p) + KA^{2\ell}\sum_{j=1}^\ell\left(\frac{\|y_j\|^p\wedge 1}{\varepsilon^p}\right), \end{aligned} \tag{A.9}$$

with $\theta_A(\varepsilon)\to 0$ as $\varepsilon\to 0$; $\Psi(A)\to 0$ as $A\to\infty$ and some constant $K>0$. By elementary calculations, we deduce for any $v\geq 1$:

$$\begin{aligned} & \left|f(x_1+y_1,\ldots,x_\ell+y_\ell)\textstyle\prod_{j=1}^\ell 1_{\{\|x_j+y_j\|\leq v\}} - f(x_1,\ldots,x_\ell)\right| \\ \leq\ & \theta_A(\varepsilon) + K\Psi(A)\textstyle\prod_{j=1}^\ell(1+\|x_j\|^p+\|y_j\|^p\wedge v^p) + KA^{2\ell}\sum_{j=1}^\ell\left(\frac{\|y_j\|^p\wedge v^p}{\varepsilon^p}\right) \\ & + |f(x_1,\ldots,x_\ell)|\left(1-\textstyle\prod_{j=1}^\ell 1_{\{\|x_j+y_j\|\leq v\}}\right) \\ \leq\ & \theta_A(\varepsilon) + K\Psi(A)\textstyle\prod_{j=1}^\ell(1+\|x_j\|^p+\|y_j\|^p\wedge v^p) + KA^{2\ell}\sum_{j=1}^\ell\left(\frac{\|y_j\|^p\wedge v^p}{\varepsilon^p}\right) \\ & + \textstyle\sum_{k=1}^\ell\prod_{j=1,j\neq k}^\ell(1+\|x_j\|^p)\left[\frac{\|x_k\|^{p+2}}{v^2} + \|y_k\|^p\wedge v^p\right]. \end{aligned} \tag{A.10}$$

In the last inequality, we use the fact that $(1+\|x\|^p))1_{\{\|x+y\|\geq v\}} \leq \delta\left(\frac{\|x\|^{p+2}}{v^2} + \|y\|^p\wedge v^p\right)$ with $\delta = 4(2^p+1)$. Then, (A.6) follows easily. □

Proof of Lemma A.2: As stated in the proof of Lemma A.1, by the localisation procedure, it is sufficient to establish this result under the stronger assumption (SH-r).

Let ψ be C^∞ on $\mathbb{R}$ with $1_{[1,\infty)} \leq \psi \leq 1_{[1/2,\infty)}$ and define $\psi_\varepsilon(x) = \psi(\|x\|/\varepsilon)$ and $\psi'_\varepsilon = 1-\psi_\varepsilon$. Write

$f_m(x_1,\ldots,x_\ell) = f(x_1,\ldots,x_\ell)\prod_{j=1}^{\ell}\psi'_m(x_j)$ and define $\hat{f}_{m,i}(k_n,v_n)$ as $\hat{f}_i(k_n,v_n)$ in (A.1) but with f replaced by f_m.

Step 1: We show that the result holds for f_m. By definition, each f_m is continuous and bounded therefore, from Lemma A.1, we have

$$\hat{f}_{m,i+1} \xrightarrow{P} \rho_{c_t}^{\otimes\ell}(f_m), \quad \text{as} \quad n\to\infty.$$

Also, by the Lebesgue dominated convergence theorem $\rho_{c_t}^{\otimes\ell}(f_m)\to\rho_{c_t}^{\otimes\ell}(f)$ as $m\to\infty$ (note that $|f_m|\le|f|$ for all m, $f_m\to f$, pointwise and $\rho_{c_t}^{\otimes\ell}(|f|)<\infty$).

Step 2: It remains to show that $Z_n \equiv \hat{f}_{m,i+1}-\hat{f}_{i+1} = o_p(1)$ for m,n large. For m fixed, for n large enough, $m\le u_n\equiv v_n/\sqrt{\Delta_n}$ and thus $\psi'_m(x)\le 1_{\{\|x\|\le u_n\}}$. Hence, we have:

$$\begin{aligned}
&|f(x_1,\ldots,x_\ell)\textstyle\prod_{j=1}^{\ell}1_{\{\|x_j\|\le u_n\}} - f_m(x_1,\ldots,x_\ell)| \\
= \;& |f(x_1,\ldots,x_\ell)|\cdot\left|\textstyle\prod_{j=1}^{\ell}1_{\{\|x_j\|\le u_n\}} - \prod_{j=1}^{\ell}\psi'_m(x_j)\right| \le |f(x)|\textstyle\prod_{j=1}^{\ell}1_{\{\|x_j\|\le u_n\}}\sum_{s=1}^{\ell}1_{\{\|x_s\|\ge m/2\}} \\
\le \;& K\textstyle\sum_{s=1}^{\ell}\prod_{j=1}^{\ell}(1+\|x_j\|^p 1_{\{\|x_j\|\le u_n\}})1_{\{\|x_s\|\ge m/2\}}.
\end{aligned}$$

It is not hard to show that $1+\|x+y\|^p 1_{\{\|x+y\|\le u_n\}} \le K\left(1+\|x\|^p+\|y\|^p\wedge u_n^p\right)$ and $(1+\|x+y\|^p 1_{\{\|x+y\|\le u_n\}})1_{\{\|x+y\|>m/2\}} \le K\left(\|x\|^{p+1}/m+\|y\|^p\wedge u_n^p\right)$; with K from now on a generic constant. Thus,

$$\begin{aligned}
&\left|f(x_1+y_1,\ldots,x_\ell+y_\ell)\textstyle\prod_{j=1}^{\ell}1_{\{\|x_j+y_j\|\le u_n\}} - f_m(x_1+y_1,\ldots,x_\ell+y_\ell)\right| \\
\le \;& K\textstyle\sum_{s=1}^{\ell}\left(\frac{\|x_s\|^{p+1}}{m}+\|y_s\|^p\wedge u_n^p\right)\prod_{j=1,j\neq s}^{\ell}\left(1+\|x_j\|^p+\|y_j\|^p\wedge u_n^p\right).
\end{aligned} \tag{A.11}$$

Then, using the same decomposition of the process X as in the proof of Lemma A.1, and then setting $x_i = \Delta^n_{i_n+i}X'/\sqrt{\Delta_n}$, $y_i=\Delta^n_{i_n+i}X''/\sqrt{\Delta_n}$ in (A.11) and using the bounds presented in that proof for their conditional expectations, it is not hard to see, applying the law of iterated expectations, that:

$$E|Z_n| \le K(1+K+K\phi_n)(K/m+K\phi_n), \quad \text{for } p\le 2, \quad \text{and}$$

$$E|Z_n| \le K\left(1+K+K\Delta_n^{1-r\varpi+p(\varpi-1/2)}\phi_n\right)(K/m+K\Delta_n^{1-r\varpi+p(\varpi-1/2)}\phi_n) \quad \text{for } p>2,$$

with $\phi_n\to 0$ as $n\to\infty$. Thanks to (A.3), both right-hand-sides tend to 0 by first letting $n\to\infty$ and then $m\to\infty$. This shows that $E|Z_n|$ converges to 0 therefore Z_n converges in probability to 0. This concludes the proof. □

A.3 Proofs

We introduce the following result that characterizes the asymptotic distribution of estimated eigenvalues and normalized eigenvectors. We first introduce some notation. Let B_n be a (q,q)-matrix that consistently estimates a symmetric positive definite matrix Σ, and $\mathbf{\Delta}$ the (q,q)-diagonal matrix with the eigenvalues $\delta_1\ge\ldots\ge\delta_q>0$ of Σ as diagonal elements. Assume that these eigenvalues have the structure in (9). Let Γ be an orthogonal matrix of normalized eigenvectors of Σ, i.e.

$$\Gamma\Gamma' = I_q \quad \text{and} \quad \Gamma'\Sigma\Gamma = \mathbf{\Delta}.$$

Let $A_n = \Gamma' B_n \Gamma$. Note that A_n and B_n have the same eigenvalues $d_1 \geq \cdots \geq d_q$, and let D_n be the diagonal matrix containing those eigenvalues and $\widehat{E}$ an orthogonal matrix of normalized eigenvectors of A_n with nonnegative diagonal elements.

Let $U_n = r_n(A_n - \mathbf{\Delta})$ and $\widehat{H} = r_n(D_n - \mathbf{\Delta})$, with $r_n \to \infty$ as $n \to \infty$. Let $\widehat{E}_{kl}$ and $U_{n,kl}$ denote the (q_k, q_l)-submatrix of $\widehat{E}$ and U_n, respectively, with elements at the intersection of rows and columns with index in $\mathscr{L}_k$ and $\mathscr{L}_l$, respectively; for $k, l = 1, \ldots, r$. Let $\widehat{H}_k$ be defined as $\widehat{E}_{kk}$, but from $\widehat{H}$ and $\widehat{F}_{kl} = r_n \widehat{E}_{kl}$, for $k \neq l$. We have the following result which is a simple adaptation of the results of Anderson (1963) (see also Davis, 1977).

Proposition A.1. *If the eigenvalues of Σ have the structure in (9), almost surely and $U_n \xrightarrow{\mathcal{L}-s} U$, and the functionally unrelated elements of U have a joint distribution that is continuous with respect to the Lebesgue measure in $\mathbb{R}^{q(q+1)/2}$, then, for $k, l = 1, \ldots, r$,*

$$\begin{aligned} \widehat{E}_{kk}\widehat{E}'_{kk} &= I_{q_k} + O_P(r_n^{-2}) \\ U_{n,kk} &= \widehat{E}_{kk}\widehat{H}_k\widehat{E}'_{kk} + O_P(r_n^{-1}) \\ U_{n,kl} &= \lambda_k \widehat{E}_{kk}\widehat{F}'_{lk} + \lambda_l \widehat{F}_{kl}\widehat{E}'_{ll} + O_P(r_n^{-1/2}), \quad k \neq l \\ 0 &= \widehat{E}_{kk}\widehat{F}'_{lk} + \widehat{F}_{kl}\widehat{E}'_{ll} + O_P(r_n^{-1/2}), \quad k \neq l, \end{aligned} \tag{A.12}$$

and $\widehat{E}_{kk}$, $\widehat{H}_k$ and $\widehat{F}_{kl}$ converge stably in law to limiting distributions E_{kk}, H_k and F_{kl}, respectively, uniquely defined in terms of U by the equations:

$$\begin{aligned} E_{kk}E'_{kk} &= I_{q_k} \\ U_{kk} &= E_{kk}H_kE'_{kk} \\ U_{kl} &= \lambda_k E_{kk}F'_{lk} + \lambda_l F_{kl}E'_{ll}, \quad k \neq l \\ 0 &= E_{kk}F'_{lk} + F_{kl}E'_{ll}, \quad k \neq l, \end{aligned} \tag{A.13}$$

where H_k is diagonal and E_{kk} is restricted to have nonnegative diagonal elements; and U_{kl} is defined similarly to $U_{n,kl}$ but from U.

The proof of this proposition follows readily from Anderson (1963), with $\sqrt{n}$ replaced by r_n. The stable convergence in law deduced for $\widehat{E}_{kk}$, $\widehat{H}_k$ and $\widehat{F}_{kl}$ follows from the stable convergence in law of U_n. The last two equalities in (A.12) imply that $\widehat{F}_{kl} = O_P(1)$, therefore, $\widehat{E}_{kl} = o_P(1)$ for $k \neq l$. □

Proof of Proposition 3.1: (a) The maximum likelihood estimator $\hat{\lambda}_k$ of λ_k $(k = 1, \ldots, r)$ is obtained by solving the first order condition associated with the log-likelihood function in (14) and it is straightforward to get $\hat{\lambda}_k = \frac{1}{q_k} \sum_{i \in \mathscr{L}_k} \tilde{d}_i$.
(b) Since the log-likelihood in (14) is additively separable in λ_k's, it is maximized under H_0 by

$$C - \frac{n}{2} q_k \log \tilde{\lambda}_k - \frac{n}{2} \sum_{i \in \mathscr{L}_k} \frac{\tilde{d}_i}{\tilde{\lambda}_k}, \quad \text{with} \quad \tilde{\lambda}_k = \frac{1}{q_k} \sum_{i \in \mathscr{L}_k} \tilde{d}_i,$$

where C is the maximum of the part of log-likelihood that depends on λ_s, for $s \neq k$. The unrestricted likelihood is maximized by

$$C - \frac{n}{2} \sum_{i \in \mathscr{L}_k} \log \tilde{\lambda}_i - \frac{n}{2} \sum_{i \in \mathscr{L}_k} \frac{\tilde{d}_i}{\tilde{\lambda}_i}, \quad \text{with} \quad \tilde{\lambda}_i = \tilde{d}_i.$$

The expression of the likelihood ratio criterion in (15) follows by straightforward derivations.

(c) Note that $\tilde{b}$ has finite mean and variance and therefore is $O_P(1)$. As a result,

$$\widetilde{IV}^n - IV_T = \sum_{i=1}^{n} (\Delta_i^n X)(\Delta_i^n X)' - IV_T + O_P(\Delta_n) = \sum_{i=1}^{n} y_i y_i' - IV_T + O_P(\Delta_n),$$

where $y_i = \Delta_i^n X - \Delta_n b \sim N\left(0, \frac{\Delta_n}{T} IV_T\right)$. In this derivation we use the simplification that T/Δ_n is integer equal to n.

Let Γ be the matrix of normalized eigenvectors of IV_T defined such that: $\Gamma'\Gamma = I_q$ and $\Gamma' IV_T \Gamma = \mathbf{\Delta}$, where $\mathbf{\Delta}$ is the diagonal matrix with diagonal vector $\delta = \lambda(IV_T)$. Let $U_n = \frac{1}{\sqrt{\Delta_n}} \left(\Gamma' \widetilde{IV}^n \Gamma - \mathbf{\Delta}\right)$. We have

$$U_n = \frac{1}{\sqrt{\Delta_n}} \sum_{i=1}^{n} \left(z_i - \frac{\mathbf{\Delta}}{n} \right) + o_P(1),$$

with $z_i = (\Gamma' y_i)(\Gamma' y_i)'$. We can easily verify the Lyapunov central limit theorem conditions and deduce:

$$U_n \xrightarrow{d} U, \tag{A.14}$$

where U is normally distributed with mean 0 and covariance: $Cov(u_{ij}, u_{gh}) = \frac{\delta_i \delta_j}{T}(\delta_{ig}\delta_{jh} + \delta_{ih}\delta_{jg})$, with u_{ab} a generic component of U and $\delta_{ab} = 1$ if $a = b$ and 0 otherwise.

We now derive the asymptotic distribution of $\widetilde{LR}_k$. We have

$$\widetilde{LR}_k = -n \left(\sum_{i \in \mathscr{L}_k} \log \tilde{d}_i - q_k \log \left(\frac{1}{q_k} \sum_{i \in \mathscr{L}_k} \tilde{d}_i \right) \right). \tag{A.15}$$

Let $\widehat{H} = \frac{1}{\sqrt{\Delta_n}} \left(\tilde{D} - \mathbf{\Delta}\right)$, where $\tilde{D}$ is the diagonal matrix containing $\tilde{d} = \lambda(\widetilde{IV}^n)$ as diagonal. Let $\widehat{E}$ be the matrix of normalized eigenvectors of $A_n = \Gamma' \widetilde{IV}^n \Gamma$, that is: $\widehat{E}\widehat{E}' = I_q$, $\widehat{E}' A_n \widehat{E} = \tilde{D}$, and $\widehat{E}_{kk}$ $(U_{n,kk}, U_{kk})$ is the submatrix of $\widehat{E}$ (U_n, U) with rows and columns with indexes in $\mathscr{L}_k$.

From Proposition A.1, we have

$$\widehat{H}_k = O_P(1), \quad U_{n,kk} = \widehat{E}_{kk} \widehat{H}_k \widehat{E}_{kk}' + o_P(1) \quad \text{and} \quad \widehat{E}_{kk} \widehat{E}_{kk}' = I_{q_k} + o_P(1). \tag{A.16}$$

Hence, with $h_i \equiv \widehat{H}_{ii}$ for $i \in \mathscr{L}_k$, we have

$$\begin{aligned}
\widetilde{LR}_k &= -n \left(\sum_{i \in \mathscr{L}_k} \log \left(1 + \tfrac{\sqrt{\Delta_n} h_i}{\lambda_k}\right) - q_k \log \left(1 + \sqrt{\Delta_n} \tfrac{\sum_{i \in \mathscr{L}_k} h_i}{q_k \lambda_k}\right) \right) \\
&= -n \left(\sum_{i \in \mathscr{L}_k} \left[\tfrac{\sqrt{\Delta_n}}{\lambda_k} h_i - \tfrac{\Delta_n}{2\lambda_k^2} \sum_{i \in \mathscr{L}_k} h_i^2 \right] - q_k \left[\tfrac{\sqrt{\Delta_n}}{q_k \lambda_k} \sum_{i \in \mathscr{L}_k} h_i - \tfrac{\Delta_n}{2 q_k^2 \lambda_k^2} \left(\sum_{i \in \mathscr{L}_k} h_i \right)^2 \right] + o_P(\Delta_n) \right) \\
&= \tfrac{n\Delta_n}{2\lambda_k^2} \left(tr(H_k^2) - \tfrac{1}{q_k} \left(tr(H_k)\right)^2 \right) = \tfrac{T}{2\lambda_k^2} \left(tr(U_{n,kk}^2) - \tfrac{1}{q_k} \left(tr(U_{n,kk})\right)^2 \right) + o_P(1),
\end{aligned} \tag{A.17}$$

where the second and third equalities follow from a second order Taylor expansion and Equation (A.16), respectively. Thus $\widetilde{LR}_k$ converges in distribution to

$$\frac{T}{2\lambda_k^2} \left(tr(U_{kk}^2) - \frac{1}{q_k} \left(tr(U_{kk})\right)^2 \right) = \frac{T}{2\lambda_k^2} \left(2 \sum_{i<j \ i,j \in \mathscr{L}_k} u_{ij}^2 + \sum_{i \in \mathscr{L}_k} u_{ii}^2 - \frac{1}{q_k} \left(\sum_{i \in \mathscr{L}_k} u_{ii} \right)^2 \right).$$

Thanks to Equation (A.14), $u_{ij} = u_{ji}$ and is independent of all the other entries of U. Moreover, $u_{ii} \sim$

$N(0, 2\lambda_k^2/T)$ and $u_{ij} \sim N(0, \lambda_k^2/T)$ for $i \neq j$. Therefore, it follows that:

$$\frac{T}{2\lambda_k^2}\left(\sum_{i\in\mathscr{L}_k} u_{ii}^2 - \frac{1}{q_k}\left(\sum_{i\in\mathscr{L}_k} u_{ii}\right)^2\right) \sim \chi^2_{q_k-1}$$

and is independent of

$$\frac{T}{\lambda_k^2}\sum_{i<j\ i,j\in\mathscr{L}_k} u_{ij}^2 \sim \chi^2_{\frac{1}{2}q_k(q_k-1)}.$$

As a result, $\widetilde{LR}_k$ is asymptotically distributed as a $\chi^2_{\frac{1}{2}(q_k-1)(q_k+2)}$.

We now show that $\widetilde{LR}_k$ diverges to infinity under the alternative. Let $\lambda_{k,i}$, for $i = 1, \ldots, q_k$, be the eigenvalues of IV_T in the cluster $\mathscr{L}_k$. Under the alternative, at least two of them are distinct. As previously, let $h_i = \frac{1}{\sqrt{\Delta_n}}\left(\tilde{d}_i - \lambda_{k,i}\right), \quad i \in \mathscr{L}_k$. From Equation (A.15), we have

$$\begin{aligned}
\widetilde{LR}_k &= -n\left(\sum_{i\in\mathscr{L}_k}\log(\lambda_{k,i}+\sqrt{\Delta_n}h_i) - q_k\log\left(\tfrac{1}{q_k}\sum_{i\in\mathscr{L}_k}(\lambda_{k,i}+\sqrt{\Delta_n}h_i)\right)\right)\\
&= -n\left(\sum_{i\in\mathscr{L}_k}\log\lambda_{k,i} + \sum_{i\in\mathscr{L}_k}\left[\sqrt{\Delta_n}\tfrac{h_i}{\lambda_{k,i}} - \tfrac{1}{2}\Delta_n\tfrac{h_i^2}{\lambda_{k,i}^2} + o_P(\Delta_n)\right]\right.\\
&\quad \left. -q_k\log\left(\frac{\sum_{i\in\mathscr{L}_k}\lambda_{k,i}}{q_k} - q_k\left[\sqrt{\Delta_n}\frac{\sum_{i\in\mathscr{L}_k}h_i}{\sum_{i\in\mathscr{L}_k}\lambda_{k,i}} - \tfrac{1}{2}\Delta_n\left(\frac{\sum_{i\in\mathscr{L}_k}h_i}{\sum_{i\in\mathscr{L}_k}\lambda_{k,i}}\right)^2 + o_P(\Delta_n)\right]\right)\right)\\
&= nq_k\left(\log\frac{\sum_{i\in\mathscr{L}_k}\lambda_{k,i}}{q_k} - \tfrac{1}{q_k}\sum_{i\in\mathscr{L}_k}\log\lambda_{k,i}\right) + O_P(n\sqrt{\Delta_n}),
\end{aligned}$$

where the second equality is obtained from a second order Taylor expansion. Since log is strictly concave, it follows that $\log\frac{\sum_{i\in\mathscr{L}_k}\lambda_{k,i}}{q_k} > \frac{1}{q_k}\sum_{i\in\mathscr{L}_k}\log\lambda_{k,i}$ and $\widetilde{LR}_k \to \infty$ as $n \to \infty$, in probability.

(d) Similar to the proof of (b), under $H_0(\lambda)$, the log-likelihood in (14) is maximized by

$$C - \frac{n}{2}q_k\log\lambda - \frac{n}{2}\sum_{i\in\mathscr{L}_k}\frac{\tilde{d}_i}{\lambda}$$

and the expression of the likelihood ratio criterion in (16) follows easily.

The asymptotic distribution of $\widetilde{LR}_{k,\lambda}$ is obtained similarly to that of $\widetilde{LR}_k$ derived above. We have

$$\begin{aligned}
\widetilde{LR}_{k,\lambda} &= -n\sum_{i\in\mathscr{L}_k}\log\tilde{d}_i + nq_k\log\lambda + \tfrac{n}{\lambda}\sum_{i\in\mathscr{L}_k}\tilde{d}_i - nq_k\\
&= -n\sum_{i\in\mathscr{L}_k}\log\left(1+\tfrac{\sqrt{\Delta_n}h_i}{\lambda}\right) + n\tfrac{\sqrt{\Delta_n}}{\lambda}\sum_{i\in\mathscr{L}_k}h_i\\
&= -n\sum_{i\in\mathscr{L}_k}\left(\tfrac{\sqrt{\Delta_n}h_i}{\lambda} - \tfrac{\Delta_n h_i^2}{\lambda^2} + o_P(\Delta_n)\right) + n\tfrac{\sqrt{\Delta_n}}{\lambda}\sum_{i\in\mathscr{L}_k}h_i\\
&= \tfrac{n\Delta_n}{2\lambda^2}\sum_{i\in\mathscr{L}_k}h_i^2 + o_P(1) = \tfrac{T}{2\lambda^2}tr(U_{n,kk}^2) + o_P(1),
\end{aligned} \tag{A.18}$$

which converges in distribution to $\frac{T}{2\lambda^2}\left(2\sum_{i<j\ i,j\in\mathscr{L}_k}u_{ij}^2 + \sum_{i\in\mathscr{L}_k}u_{ii}^2\right)$. Recalling the distribution of u_{ij} as given in the proof of (c) above, we can claim that $\widetilde{LR}_{k,\lambda}$ converges in distribution to $\chi^2_{\frac{1}{2}q_k(q_k+1)}$.

We next show that $\widetilde{LR}_{k,\lambda}$ diverges under the alternative to $H_0(\lambda)$. Similar calculations to those in the proof

of divergence of $\widetilde{LR}_k$ above lead to

$$\widetilde{LR}_{k,\lambda} = n \sum_{i \in \mathscr{L}_k} \left[\left(\frac{\lambda_{k_i}}{\lambda} - 1 \right) - \log \frac{\lambda_{k_i}}{\lambda} \right] + O_P(n\sqrt{\Delta_n}).$$

Note that $x \mapsto x - 1 - \log x$ is nonnegative on $(0, +\infty)$ and takes value 0 only at $x = 1$. Since under the alternative $\lambda_{k_i}/\lambda \neq 1$ for at least one $i \in \mathscr{L}_k$, we can conclude that $\widetilde{LR}_{k,\lambda} \to \infty$, in probability. $\square$

Proof of Equation (18)**:** Recalling that $\Delta_n^i X \sim N(\Delta_n b, \Delta_n c)$, we have

$$\widetilde{IV}^n = \sum_{i=1}^n \left(\Delta_n^i X - \Delta_n \tilde{b} \right) \left(\Delta_n^i X - \Delta_n \tilde{b} \right)' = \sum_{i=1}^n (\Delta_n^i X)(\Delta_n^i X)' - \Delta_n T \tilde{b}\tilde{b}' = \overline{IV}^n + O_P(\Delta_n).$$

To prove the second equality, observe that

$$\widetilde{IV}^n - \overline{IV}^n = \sum_{i=1}^n (y_i - \Delta_n \tilde{b})(y_i - \Delta_n \tilde{b})' 1_{\{\|y_i\| > \alpha \Delta_n^{\varpi}\}} + O_P(\Delta_n),$$

where $y_i \equiv \Delta_n^i X$. Hence,

$$\frac{1}{\sqrt{\Delta_n}} (\widetilde{IV}^n - \widehat{IV}^n) = \sqrt{\Delta_n} c^{1/2} \left(\sum_{i=1}^n \Delta_n^{-1} c^{-1/2} (y_i - \Delta_n \tilde{b})(y_i - \Delta_n \tilde{b})' c^{-1/2} 1_{\{\|y_i\| > \alpha \Delta_n^{\varpi}\}} \right) c^{1/2} + O_P(\sqrt{\Delta_n}).$$

The first term in the right hand side is then of the same order of magnitude as

$$\sqrt{\Delta_n} \sum_{i=1}^n z_i z_i' 1_{\{\|y_i\| > \alpha \Delta_n^{\varpi}\}} + O_P(\sqrt{\Delta_n}),$$

where z_i's are $N(0, I_q)$. By The triangle and the Cauchy-Schwarz inequalities, we have

$$\left\| \sqrt{\Delta_n} \sum_{i=1}^n z_i z_i' 1_{\{\|y_i\| > \alpha \Delta_n^{\varpi}\}} \right\| \leq \sqrt{\Delta_n} \sum_{i=1}^n \|z_i\|^2 1_{\{\|y_i\| > \alpha \Delta_n^{\varpi}\}}.$$

To conclude the claimed order of magnitude, it suffices to show that the right hand side of the above inequality converges in absolute mean to 0. By the Cauchy-Schwarz and Markov inequalities, there exists a (generic) constant $C > 0$ such that, for any $\ell > 0$,

$$E \left(\sqrt{\Delta_n} \sum_{i=1}^n \|z_i\|^2 1_{\{\|y_i\| > \alpha \Delta_n^{\varpi}\}} \right) \leq C \sqrt{\Delta_n} \sum_{i=1}^n P \left(\|y_i\| > \alpha \Delta_n^{\varpi} \right)^{1/2} \leq C \Delta_n^{\frac{1}{2} - \frac{\ell \varpi}{2}} \sum_{i=1}^n E \left(\|y_i\|^{\frac{\ell}{2}} \right).$$

Note that

$$E \left(\|y_i\|^{\frac{\ell}{2}} \right) = E \| \sqrt{\Delta_n} c^{1/2} z_i + \Delta_n \tilde{b} \|^{\frac{\ell}{2}} \leq C \left(\Delta_n^{\frac{\ell}{4}} E \|z_i\|^{\frac{\ell}{2}} + \Delta_n^{\frac{\ell}{2}} E \|\tilde{b}\|^{\frac{\ell}{2}} \right),$$

where we have used the Cauchy-Schwarz and the C_r inequalities.
Hence, the leading term of $\Delta_n^{\frac{1}{2} - \frac{\ell \varpi}{2}} \sum_{i=1}^n E \left(\|y_i\|^{\frac{\ell}{2}} \right)$ is at most of order $n \Delta_n^{\frac{1}{2} - \frac{\ell \varpi}{2} + \frac{\ell}{4}} = T \Delta_n^{-\frac{1}{2} - \frac{\ell \varpi}{2} + \frac{\ell}{4}}$ and the expected convergence to 0 is warranted since we can find ℓ such that $\varpi < \frac{1}{2} - \frac{1}{\ell}$. $\square$

Proof of Theorem 3.1: Since $\mathcal{U}_n \equiv \frac{1}{\sqrt{\Delta_n}} \left(\Gamma' \widehat{IV}^n \Gamma - \mathbf{\Delta} \right) = \Gamma' \left(\frac{1}{\sqrt{\Delta_n}} \left(\widehat{IV}^n - IV_T \right) \right) \Gamma$ converges stably in law to $\mathcal{U}_T$, which has a continuous distribution, Proposition A.1 can be applied as in the proof of Proposition 3.1 and we can claim that the expansion of $\widetilde{LR}_k$ in Equation (A.17) also holds for LR_k and that of $\widetilde{LR}_{k,\lambda}$ in Equation (A.18) holds for $LR_{k,\lambda}$. We can therefore write

$LR_k = \frac{T}{2\lambda_k^2}\left(tr\left(\mathcal{U}_{n,kk}^2\right) - \frac{1}{q_k}\left(tr\left[\mathcal{U}_{n,kk}\right]\right)^2\right) + o_P(1)$, and $LR_{k,\lambda} = \frac{T}{2\lambda^2} tr\left(\mathcal{U}_{n,kk}^2\right) + o_P(1)$, where $\mathcal{U}_{n,kk}$ is the submatrix of $\mathcal{U}_n$ at the intersection of rows and columns with indexes in $\mathscr{L}_k$. The claimed result then holds by the continuous mapping theorem.

The proof of divergence of these test statistics follows the exact same lines as the proof of their respective counterparts of Proposition 3.1. □

Proof of Theorem 3.2: Under the conditions on the theorem, $\mathcal{U}_n = \frac{1}{\sqrt{\Delta_n}}\left(\Gamma'\widehat{IV}^n\Gamma - \mathbf{\Delta}\right)$ converges stably in law to $\mathcal{U}_T = \Gamma'\mathcal{W}_T\Gamma$; see Equation (20). Assume without loss of generality that the eigenvalues of IV_T have the structure in Equation (9) and let $\mathscr{L}_k$ $(k = k_1, \ldots, r)$ be the clusters of the $q - Q$ smallest eigenvalues of IV_T. We have:

$$\sum_{i=Q+1}^{q} d_i = \sum_{k=k_1}^{r}\sum_{i\in\mathscr{L}_k} d_i = \sum_{k=k_1}^{r}\left(\sum_{i\in\mathscr{L}_k}\left[d_i - \lambda_k\right] + q_k\lambda_k\right).$$

Using the notation leading to Proposition A.1, with $S_n = \widehat{IV}^n$, $U_n = \mathcal{U}_n$, etc., this proposition allows us to claim that $\frac{1}{\sqrt{\Delta_n}}\sum_{i\in\mathscr{L}_k}(d_i - \lambda_k) = tr(\widehat{H}_k) = tr(\mathcal{U}_{n,kk}) + o_P(1)$. Thus,

$$\frac{1}{\sqrt{\Delta_n}}\sum_{i=Q+1}^{q} d_i = \sum_{k=k_1}^{r} tr(\mathcal{U}_{n,kk}) + \frac{1}{\sqrt{\Delta_n}}\sum_{k=k_1}^{r} q_k\lambda_k + o_P(1) = tr\left(\mathcal{U}_{n,Q+1:q,Q+1:q}\right) + \frac{1}{\sqrt{\Delta_n}}\sum_{k=k_1}^{r} q_k\lambda_k + o_P(1).$$

Similarly, $\frac{1}{\sqrt{\Delta_n}}\sum_{i=1}^{q} d_i = tr\left(\mathcal{U}_n\right) + \frac{1}{\sqrt{\Delta_n}}\sum_{k=1}^{r} q_k\lambda_k + o_P(1)$. As a result,

$$T_n = tr\left(\mathcal{U}_{n,Q+1:q,Q+1:q}\right) - \pi\cdot tr\left(\mathcal{U}_n\right) + \frac{1}{\sqrt{\Delta_n}}\left(\sum_{k=k_1}^{r} q_k\lambda_k - \pi\cdot\sum_{k=1}^{r} q_k\lambda_k\right) + o_P(1).$$

(a) If the null hypothesis holds with equality, the claimed asymptotic distribution is obtained thanks to the continuous mapping theorem. (b) If the null holds with strict inequality, T_n diverges to $-\infty$. (c) If the null does not hold, T_n diverges to $+\infty$, thus (c). □

Proof of Theorem 3.3: The proof of this theorem follows the same lines as that of Theorem 3.1 and uses the continuity of the asymptotic distribution $\mathcal{U}_T^\rho$ in (26). □

Proofs of Proposition 4.1 and Corollary 4.1: We rely on Hounyo (2017, Ths. 3.1 and 3.2) to establish these results. For this, it suffices to check the Condition A in Hounyo (2017). That is:
(i) For $k, l, k', l' = 1, \ldots, q$,

$$\frac{n}{2}\sum_{i=1}^{n-1}\left(y_{i,k}y_{i,l} - y_{i+1,k}y_{i+1,l}\right)\left(y_{i,k'}y_{i,l'} - y_{i+1,k'}y_{i+1,l'}\right)$$

converges in probability to $\int_0^T \left[c_s^{kk'}c_s^{ll'} + c_s^{kl'}c_s^{lk'}\right] ds$,
(ii) $n^{1+\epsilon}\sum_{i=1}^{n}|y_{i,k}y_{i,l}|^{2+\epsilon} = O_P(1)$, for $k, l = 1, \ldots, q$ and some $\epsilon > 2$ and
(iii) $\frac{\sqrt{n}^{\frac{2+\epsilon}{1+\epsilon}}}{n} = o\left(1\right)$. Both (i) and (ii) follow from Jacod and Protter (2012, Th. 9.4.1) while (iii) is obvious. □

Proof of Theorem 4.1: U_{kk}^* and $U_{Q+1:q,Q+1:q}^*$ can both be written as $\widehat{\Gamma}_0' S_n^* \widehat{\Gamma}_0$ with $\widehat{\Gamma}_0 = \Gamma\widehat{E}_0$, where $\widehat{E}_0$ is a matrix equal to the collection of columns of $\widehat{E}$ indexed by $\mathscr{L}_k$ or $\bigcup_{k=k_1}^{r}\mathscr{L}_k = \{Q+1, \ldots, q\}$. Thanks to the arguments leading to Equation (28), we can claim that LR_k^*, $LR_{k,\lambda}^*$ and Z_n^* converge in distribution to

the same limit distributions as those of LR_k, $LR_{k,\lambda}$ and Z_n as given in Theorem 3.1(a) and (b) and Theorem 3.3(a), respectively. The claimed uniform consistencies in part (a), (b) and (c) of Theorem 4.1 follow from the continuity of these asymptotic distributions. □

Proof of Theorem 5.1: Because $E^*(\eta_m) = 1$ and $Var^*(\eta_m) = 1/2$, it not hard to see that $E^*(S^*_{nt}) = 0$ and, the (q^2, q^2)-matrix $V^*_n = Var^*(vec(S^*_{nt}))$ has its (ab, cd) element, for $a, b, c, d = 1, \ldots, q$, given by:

$$Covar^*(S^*_{nt,ab}, S^*_{nt,cd}) = \frac{1}{2k_n\Delta_n^2}\sum_{m=0}^{k_n-1}(\mathcal{Z}^{ab}_m - \mathcal{Z}^{ab}_{m+1})(\mathcal{Z}^{cd}_m - \mathcal{Z}^{cd}_{m+1})'.$$

By Lemma A.2, this quantity converges in probability to $c_t^{ac}c_t^{bd} + c_t^{ad}c_t^{bc}$ corresponding to the asymptotic variance of S_{nt}. Therefore, it remains to show that S^*_{nt} is asymptotically normally distributed with respect to the bootstrap measure P^*. We will show that $vech\,(S^*_{nt})$ is asymptotically normal. We can see that the probability limit of V^*_n is equal to $Var(y \otimes y)$ with $y \sim N(0, c_t)$ and '$\otimes$' the Kronecker product. Since c_t is almost surely positive definite, we can rely on Magnus and Neudecker (1979, Th. 4.3(v)) to claim that $\tilde{V}^*_n \equiv Var^*\left[vech\,(S^*_{nt})\right]$ is positive definite with probability approaching 1. Thus, it suffices to show that:

$$Z^*_n \equiv \tilde{V}^{*^{-1/2}}_n vech(S^*_{nt}) \xrightarrow{d^*} N\left(0, I_{\frac{q(q+1)}{2}}\right), \quad \text{in probability.}$$

For this, we rely on the modified Cramer-Wold device [see Pauly (2011)] by showing that for any $\lambda \in D$ a countably dense subset of the unit sphere in $\mathbb{R}^{\frac{q(q+1)}{2}}$, $\lambda' Z^*_n \equiv \sum_{m=0}^{k_n-1} z_{n,m} \xrightarrow{d^*} N(0,1)$, in probability; with $z_{n,m} = \lambda' \tilde{V}^{*^{-1/2}}_n vech(\mathcal{Z}^*_m - \mathcal{Z}_m)$. Since $E^*(\lambda' Z^*_n) = 0$ and $Var^*(\lambda' Z^*_n) = 1$, it suffices to verify the Lyapunov's condition: for some $\delta > 0$,

$$\sum_{m=0}^{k_n-1} E^* \left\|z_{n,m}\right\|^{2+\delta} \to 0, \quad \text{in probability as } n \to \infty, \tag{A.19}$$

where the norm sign denotes the Euclidean norm. By the Cauchy-Schwarz inequality, we have:

$$\|z_{n,m}\|^{2+\delta} \le k_n^{-\frac{\delta}{2}}(\lambda'\tilde{V}^{*^{-1}}_n\lambda)^{(2+\delta)/2}\frac{1}{k_n}\|vech(\mathcal{Z}^*_m - \mathcal{Z}_m)\|^{2+\delta} \le k_n^{-\frac{\delta}{2}}(\lambda'\tilde{V}^{*^{-1}}_n\lambda)^{(2+\delta)/2}\frac{1}{k_n}\|\mathcal{Z}^*_m - \mathcal{Z}_m\|^{2+\delta}.$$

Note that $\lambda'\tilde{V}^{*^{-1}}_n\lambda \xrightarrow{P} \lambda'\tilde{V}^*\lambda = O_P(1)$ since $\tilde{V}^* \equiv \text{plim}\tilde{V}^*_n$ is nonsingular almost surely. Also,

$$\|\mathcal{Z}^*_m - \mathcal{Z}_m\|^{2+\delta} = \|\mathcal{Z}_{m+1} - \mathcal{Z}_m\|^{2+\delta}|1-\eta_m|^{2+\delta} \le 2^{1+\delta}[\|\mathcal{Z}_{m+1}\|^{2+\delta} + \|\mathcal{Z}_m\|^{2+\delta}](1 + |\eta_m|^{2+\delta}),$$

where the last inequality follows from the Jensen's inequality. Thus, taking $\delta : 0 < \delta < \epsilon$, we have:

$$E^*\|z_{n,m}\|^{2+\delta} \le O_P(1)k_n^{-\delta/2}\frac{1}{k_n}\left[\|\mathcal{Z}_m\|^{2+\delta} + \|\mathcal{Z}_{m+1}\|^{2+\delta}\right] = O_P(1)k_n^{-\delta/2}\frac{1}{k_n}\left[\|y_m\|^{4+4\delta} + \|y_{m+1}\|^{4+2\delta}\right],$$

where the $O_P(1)$ term keeps its magnitude uniformly over $m = 0, \ldots, k_n - 1$ and
$y_m = \Delta^n_{i+1+m}X1_{\{\|\Delta^n_{i+1+m}X\| \le \alpha\Delta^{\varpi}\}}$. Hence,

$$\sum_{m=0}^{k_n-1} E^*\|z_{n,m}\|^{2+\delta} \le O_P(1)k_n^{-\delta/2}\frac{1}{k_n}\sum_{m=0}^{k_n-1}\left(\|y_m\|^{4+4\delta} + \|y_{m+1}\|^{4+2\delta}\right).$$

Now, choose δ small enough so that $\varpi > \frac{4+2\delta-2}{2(4+2\delta-r)} = \frac{1+\delta}{4-r+2\delta}$.
Lemma A.2 ensures that both $(1/k_n)\sum_{m=0}^{k_n-1}\|y_m\|^{4+4\delta}$ and $(1/k_n)\sum_{m=0}^{k_n-1}\|y_{m+1}\|^{4+4\delta}$ are $O_P(1)$ and this establishes (A.19). The statement of the lemma follows from the fact that $vech(S_{nt})$ and $vech(S^*_{nt})$ have the same limit law that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{q(q+1)/2}$. □

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Aït-Sahalia, Y. and J. Jacod (2014). *High-frequency financial econometrics*. Princeton University Press.

Aït-Sahalia, Y. and D. Xiu (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics 201*(2), 384–399.

Aït-Sahalia, Y. and D. Xiu (2019). Principal component analysis of high-frequency data. *Journal of the American Statistical Association 114*(525), 287–303.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics 34*(1), 122–148.

Ang, A., J. Liu, and K. Schwarz (2020). Using stocks or portfolios in tests of factor models. *Journal of Financial and Quantitative Analysis 55*(3), 709–750.

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

Barndorff-Nielsen, O. E. and N. Shephard (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica 72*(3), 885–925.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–1304.

Chen, D., P. A. Mykland, and L. Zhang (2019). The five trolls under the bridge: Principal component analysis with asynchronous and noisy high frequency data. *Journal of the American Statistical Association 115*(532), 1–18.

Chen, Q. and Z. Fang (2019). Inference on functionals under first order degeneracy. *Journal of Econometrics 210*(2), 459–481.

Chu, K.-W. E. (1990). On multiple eigenvalues of matrices depending on several parameters. *Journal on Numerical Analysis 27*(5), 1368–1385.

Cochrane, J. H. (1996). A cross-sectional test of an investment-based asset pricing model. *Journal of Political Economy 104*(3), 572–621.

Cook, R. D. and C. M. Setodji (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association 98*(462), 340–351.

Cooper, M. J., H. Gulen, and M. J. Schill (2008). Asset growth and the cross-section of stock returns. *The Journal of Finance 63*(4), 1609–1651.

Davis, A. (1977). Asymptotic theory for principal component analysis: Non-normal case. *Australian & New Zealand Journal of Statistics 19*(3), 206–212.

Dovonon, P., S. Gonçalves, and N. Meddahi (2013). Bootstrapping realized multivariate volatility measures. *Journal of Econometrics 172*(1), 49–65.

Eaton, M. L. and D. E. Tyler (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Annals of Statistics 19*(1), 260–271.

Gomes, J., L. Kogan, and L. Zhang (2003). Equilibrium cross section of returns. *Journal of Political Economy 111*(4), 693–732.

Hiriart-Urruty, J.-B. and D. Ye (1995). Sensitivity analysis of all eigenvalues of a symmetric matrix. *Numerische Mathematik 70*(1), 45–72.

Hounyo, U. (2017). Bootstrapping integrated covariance matrix estimators in noisy jump–diffusion models with non-synchronous trading. *Journal of Econometrics 197*(1), 130–152.

Hounyo, U., S. Gonçalves, and N. Meddahi (2017). Bootstrapping pre-averaged realized volatility under market

microstructure noise. *Econometric Theory 33*(4), 791–838.

Jacod, J. and M. Podolskij (2013). A test for the rank of the volatility process: the random perturbation approach. *Annals of Statistics 41*(5), 2391–2427.

Jacod, J. and P. Protter (2012). *Discretization of processes*. Springer-Verlag, Berlin Heidelberg.

Jagannathan, R. and Y. Wang (2007). Lazy investors, discretionary consumption, and the cross-section of stock returns. *The Journal of Finance 62*(4), 1623–1661.

Magnus, J. R. (1985). On differentiating eigenvalues and eigenvectors. *Econometric Theory 1*(2), 179–191.

Magnus, J. R. and H. Neudecker (1979). The commutation matrix: some properties and applications. *The Annals of Statistics 7*(2), 381–394.

Muirhead, R. J. (1982). *Aspects of multivariate statistical analysis*. John Wiley & Sons.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Onatski, A., M. J. Moreira, and M. Hallin (2013). Asymptotic power of sphericity tests for high-dimensional data. *Annals of Statistics 41*(3), 1204–1231.

Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics 5*, 41–52.

Pelger, M. (2019). Large-dimensional factor modeling based on high-frequency observations. *Journal of Econometrics 208*(1), 23–42.

Todorov, V. and T. Bollerslev (2010). Jumps and betas: A new framework for disentangling and estimating systematic risks. *Journal of Econometrics 157*(2), 220–235.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *Annals of Statistics 9*(4), 725–736.

Waternaux, C. M. (1976). Asymptotic distribution of the sample roots for a nonnormal population. *Biometrika 63*(3), 639–645.