Journal of Econometrics 211 (2019) 439–460

Contents lists available at ScienceDirect

Journal of Econometrics

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

A rank test for the number of factors with high-frequency data

Xin-Bing Kong[1], Zhi Liu [*,2], Wang Zhou[3]

Nanjing Audit University, China
University of Macau, China
National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 29 May 2016
Received in revised form 6 March 2019
Accepted 10 March 2019
Available online 2 April 2019

JEL classification:
C01
C12
C22

Keywords:
Continuous-time factor model
High-dimensional Itô process
Idiosyncratic process

ABSTRACT

In the literature, consistency of the estimates of the number of factors for large-dimensional factor models had been extensively studied recently. But the second-order property of the estimator has long been unsolved due to lack of limiting distribution of the estimators. In this paper, we propose a rank test of the number of factors using large panel high-frequency data contaminated with microstructure noise. The rank test is realized by forming a fixed number of portfolios which reduce the dimension to a finite number. In the process of constructing portfolios, the number of factors is equal to the rank of the volatility matrix of the diversified portfolios asymptotically. Via estimating the volatility rank of a low-dimensional price dynamics of the portfolios, we establish a central limit theorem of the estimated factor number. We then apply the asymptotic normality to testing on the number of factors. Numerical experiments including the Monte-Carlo simulations and real data analysis justify our theory.

1. Introduction

With the rapid development of information technology, we are now facing huge amount of real time data stream. In finance, tick-by-tick transaction prices of a large number of assets appear as a common occurrence in modern data base. Therefore, recent years have seen increasing interest in statistical inference on the latent structure of large panel high-frequency data, cf., Wang and Zou (2010), Tao et al. (2013), Kim et al. (2018), Liu and Tang (2014), Fan et al. (2016), Aït-Sahalia and Xiu (2017, 2018), Kong (2017) and Kong (2018) and references therein.

A widely accepted model underlying large panel high-frequency data is the following high-dimensional Itô process defined on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$:

$$dX_{it} = \boldsymbol{\beta}_i' d\boldsymbol{Z}_t + dZ_{it}^*, \quad 1 \le i \le p, \tag{1.1}$$

where X_{it} stands for a continuous-time log price trajectory of asset i, $\boldsymbol{\beta}_i$ is a $r \times 1$ vector of factor loadings associated with asset i, $\boldsymbol{Z}_t$ is a $r \times 1$ vector of continuous-time factors, and Z_{it}^* is a specific factor process. Model (1.1) was first introduced and studied in Aït-Sahalia and Xiu (2017) and Pelger (2018). Other variations of model (1.1) can be found in Kong (2017, 2018) allowing for time-varying factor loadings, and Fan et al. (2016) for observable factors. In the present paper, we

* Corresponding author at: University of Macau, China.
E-mail addresses: xinbingkong@126.com (X.-B. Kong), liuzhi@umac.mo (Z. Liu), stazw@nus.edu.sg (W. Zhou).

1 Kong's work is supported by NSF China 11831008, 11571250 and PAPD of Jiangsu Higher Education Institutions.
2 Liu's research is partially supported by FDCT of Macau FDCT127/2016/A3 and FDCT202/2017/A3.
3 Zhou's research was partially supported by the MOE Tier 2 grant MOE2015-T2-2-039 (R-155-000-171-112) in Singapore.

https://doi.org/10.1016/j.jeconom.2019.03.004
0304-4076/

assume that the factor loadings are constant and the continuous-time factors are unobservable. In matrix form, (1.1) can be rewritten as

$$(d\boldsymbol{X}_t)_{p\times 1} = \boldsymbol{\beta}_{p\times r}(d\boldsymbol{Z}_t)_{r\times 1} + (d\boldsymbol{Z}_t^*)_{p\times 1}, \tag{1.2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)'$, $\boldsymbol{X} = (X_1, \ldots, X_p)'$, $\boldsymbol{Z} = (Z_1, \ldots, Z_r)'$, $\boldsymbol{Z}^* = (Z_1^*, \ldots, Z_p^*)'$ and $\boldsymbol{\beta}$ is a matrix of rank r. The discrete-time large-dimensional approximate factor model similar to (1.2) was introduced in Chamberlain and Rothschild (1983), and extensively studied by Bai and Ng (2002) and Fan et al. (2013).

Statistical inference on model (1.1) is a hot topic recently and the availability of high-frequency data facilitates this since it typically does not require stationarity of the log-price process and volatility process. Assuming observable factors and factor number, Fan et al. (2016) proposed a factor-based estimator of the large covariance matrix. Aït-Sahalia and Xiu (2018) investigated extensively into the principal component analysis (PCA) of high-frequency data of fixed dimension. In Aït-Sahalia and Xiu (2017), the authors adopted a global PCA technique and gave PCA-based estimators of the number of common factors, and large covariation matrix as well as its inverse. Extensions can be found in Kong (2017, 2018).

In statistical inference on model (1.1), estimating the number of factors is a key ingredient. However, only consistency of the estimators was established and the second-order asymptotics were still unclear. The factor number provides interpretable economic constructs in practice. For example, in finance and macroeconomics, it stands for the count of non-diversifiable risk sources. In consumer demand theory, it provides crucial information revealing preference. As pointed out in Stock and Watson (2005) and Onatski (2009), there is an ongoing debate in macroeconomics whether the factor number is only two or as large as seven. In finance, the famous Fama–French factor model consists of three or sometimes four factors. In classic econometric model checking, for example testing for CAPM, a first step is to identify whether $r = 1$ in model (1.1). Theoretically, Stock and Watson (2002) showed that given the factor space (including the factor number) is appropriately estimated, the forecast is first-order efficient. Practically, the same paper showed that underestimating or overestimating the factor number deteriorates the performance of forecasting in terms of mean square error. All these raise the question of how to test the hypotheses:

$$H_0 : r = r_0, \quad vs \quad H_1 : r \neq r_0. \tag{1.3}$$

With discrete-time series data, Onatski (2009) considered testing for a one-sided alternative $H_1 : r > r_0$ based on the random matrix theory. An earlier test was proposed in Connor and Korajczyk (1993), in which the factors were assumed to be known. The present paper differs from these two reference papers in several aspects. First, the data structure we used is large-panel high-frequency data and our model is a continuous-time factor model. Second, our test is based on a direct point estimator of the factor number together with an associated central limit theorem of the point estimator, and hence our test is well-suited for two-sided alternative hypotheses. The two tests aforementioned in the literature are from indirect features related to the factor number and therefore only applicable for the one-sided alternative $H_1 : r > r_0$ with discrete time panel data.

In this paper, we introduce a novel estimate of the factor number first and then establish a central limit theorem for the newly proposed estimator. Another feature of the high-dimensional data is the presence of microstructure noise, cf., Zhang et al. (2005). For ease of presentation, we let

$$(\boldsymbol{Y}_t)_{p\times 1} = (\boldsymbol{X}_t)_{p\times 1} + (\boldsymbol{\epsilon}_t)_{p\times 1}, \tag{1.4}$$

where $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt})'$ is an observed vector at time t and $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{pt})'$ is a noise vector. Dai et al. (2018) investigated into the estimation of factor-based large covariation matrices (and their inverses) using noise-contaminated large panel high-frequency data.

Our procedure goes as follows. We first smooth the high-frequency data cross-sectionally by a carefully tuned orthogonal matrix, resulting in a finite number of portfolios. For each portfolio, the major target of this cross-sectional manipulation is to reduce the idiosyncratic errors while maintaining the information strength of r contained in common factor terms. The intuition comes from the financial observation that the idiosyncratic risks are diversifiable in forming a large portfolio. As a byproduct, this cross-sectional transform also helps in eliminating the microstructure noise in spacial direction as the dimension $p \to \infty$. Estimating the number of factors of a large panel transfers to estimation of the rank of a finite number of portfolios. Then, we implement the perturbation approach on the price dynamics of those portfolios to estimate the rank. For more technical details on the perturbation approach, we refer to Kritchman and Nadler (2008), Jacod and Podolskij (2013) and many references therein. The former adopted the matrix perturbation method to determine the number of components in a linear mixture model with high-dimensional noisy data. The latter proposed an estimator of the maximal rank of the matrix-valued volatility process for finite-dimensional Itô semimartingale.

The contribution of this paper is in the following aspects. We provided a clever way of constructing diversified portfolios that sufficiently inherit the information of the factor space of a large panel, and have little to do with the idiosyncratic risk factor and microstructure noise asymptotically. This achieves dimension reduction in the context of testing on the factor number. We provided a rank-based estimator of the factor number that serves as an alternative to existing ones. Though the resulting non-integer valued point estimate is meaningless in the context of determining the number of factors, its distribution provides a measure of significance for testing against a hypothesis, like $r = 1$ in CAPM. We established a central limit theorem of the newly estimated factor number while existing estimators only have consistency. The existing estimators were proposed via minimizing some criterion functions over the set of positive

integers. So the resulting estimators have to be integer-valued which makes the asymptotic normality infeasible. The estimator in the present paper is based on the determinant of the low-dimensional portfolio returns and hence takes value in the real line.

The rest of the present paper is arranged as follows. In Section 2, we introduce our portfolio construction approach. The rank estimator and main asymptotic results are given in Section 3. Section 4 is devoted to Monte-Carlo simulation studies and real data analysis. Section 5 concludes. All proofs are relegated to the appendix.

Throughout the paper, we use $\|a\|$ and $\|A\|$ to denote the Euclidean norm of a vector a and the spectral norm of a matrix A. $\|A\|_F$ stands for the Frobenius norm of a matrix A.

2. Assumptions and methodology

2.1. Assumptions

Before introducing the portfolio weights, we make some technical assumptions that are needed in theoretical analysis.

Assumption 1. The common factor $\boldsymbol{Z}$ and idiosyncratic component $\boldsymbol{Z}^*$ are continuous Itô semimartingales satisfying

$$d\boldsymbol{Z}_t = \boldsymbol{b}_t dt + \boldsymbol{\sigma}_t d\boldsymbol{W}_t, \quad d\boldsymbol{Z}_t^* = \boldsymbol{\sigma}_t^* d\boldsymbol{W}_t^*,$$
$$d\boldsymbol{\sigma}_t = \breve{\boldsymbol{b}}_t dt + \breve{\boldsymbol{\sigma}}_t d\boldsymbol{W}_t + \breve{\boldsymbol{\sigma}}_t^{\perp} d\boldsymbol{W}_t^*,$$

where $\boldsymbol{b}$ is a $r\times 1$ vector of locally bounded adapted processes, $\boldsymbol{\sigma} = (\sigma_{ij})_{r\times r}$ is a $r\times r$ matrix valued adapted locally bounded processes, $\boldsymbol{\sigma}_t^* = diag\{\sigma_{1t}^*, \ldots, \sigma_{pt}^*\}$ is a $p\times p$ diagonal matrix of adapted and uniformly locally bounded processes, $\boldsymbol{W}$ and $\boldsymbol{W}^*$ are respectively r and p variate standard Brownian motions, and $\breve{\boldsymbol{b}}_{r\times r}$, $\breve{\boldsymbol{\sigma}}_{r\times r\times r}$, and $\breve{\boldsymbol{\sigma}}^{\perp}_{r\times r\times p}$ are arrays of adapted and locally bounded processes. Furthermore, for some constant C,

$$E_{\mathcal{F}_t}\left(\|\boldsymbol{b}_{t+s} - \boldsymbol{b}_t\|^{2d} + \|\breve{\boldsymbol{\sigma}}_{t+s} - \breve{\boldsymbol{\sigma}}_t\|^{2d} + \|\breve{\boldsymbol{\sigma}}^{\perp}_{t+s}(\breve{\boldsymbol{\sigma}}^{\perp}_{t+s})' - \breve{\boldsymbol{\sigma}}^{\perp}_{t}(\breve{\boldsymbol{\sigma}}^{\perp}_{t})'\|^{2d}\right) \le Cs^d.$$

Assumption 1 imposes regularity conditions on the drift and volatility coefficients of the price dynamics. This assumption implies that the components of $\boldsymbol{\sigma}$ have $\frac{1}{2}$-Hölder continuous paths, and allows almost arbitrary forms of heteroscedasticity in $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$. The local boundedness condition for $\boldsymbol{\sigma}^*$ is different from the conventional one given in Aït-Sahalia and Jacod (2012) because we let $p\to\infty$ in the present paper. We require that the local bound for σ_i^*'s be independent of i. We exclude jumps from the price dynamics and volatility processes in order not to complicate the model and technicality. We believe that applying the truncation technique in Jing et al. (2014) and Pelger (2018) on our pre-averaged returns before constructing portfolios works. In our empirical analysis, we removed extremely large returns as well as overnight jumps in data preprocessing. But the jump factor is of great importance theoretically and empirically in econometrics, cf., Li et al. (2017). We leave a thorough treatment of jumps to future work.

Next, we give an assumption on the spectrum of the factor loading matrix $\boldsymbol{\beta}$ and factor volatility matrix $\boldsymbol{\sigma}$, which also serves as a partial identifiability condition for factor loadings and factor volatilities.

Assumption 2. $\boldsymbol{\beta}'\boldsymbol{\beta}/p = I_r$, where I_r is a r-dimensional identity matrix and $\int_0^{T/2}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t' dt$ is a $r\times r$ positive definite matrix almost surely with eigenvalues $c > \lambda_1 > \lambda_2 > \cdots > \lambda_r > c^{-1}$ for some constant $c > 0$, and $E(\min_{1\le l\le r}|\lambda_l - \lambda_{l+1}|)^{4d} > c^{-1}$. $\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t'$ for all $T/2 \le t \le T$ are positive definite. $corr(\boldsymbol{W}) = I_r$.

The first condition is the pervasiveness condition that was widely used in the literature. In model (1.2), the factor loading matrix $\boldsymbol{\beta}$ and the spot volatility matrix $\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t'$ are not identified. For any $r\times r$ invertible constant matrix $\boldsymbol{C}$, $\boldsymbol{\beta}\boldsymbol{C}$ and $\boldsymbol{C}^{-1}\boldsymbol{\sigma}\boldsymbol{\sigma}'\boldsymbol{C}^{-1\prime}$ could serve as new factor loading matrix and factor volatility matrix, respectively. In Assumption 2, we normalize the columns of $\boldsymbol{\beta}$ to be orthonormal and leave $\boldsymbol{\sigma}d\boldsymbol{W}_t$ undetermined in scale and rotation. In this way, it is convenient to study the asymptotics of the portfolio weights. We also assume that $\int_0^T\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t' dt$ and $\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t'$ for all $T/2 \le t \le T$ are positive definite which implies that the rank of the spot volatilities of $\boldsymbol{Z}$ at all time instances in $(T/2, T)$ are identical to r. The last condition on the lags of eigenvalues of $\int_0^{T/2}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t' dt$ is an analogue to that in Fan et al. (2013) except that ours is on a random covariation matrix instead of the deterministic covariance matrix. So ours is in the form of expectation, which basically says that the eigenvalues are distinct in moment.

Assumption 3. $corr(\boldsymbol{W}, \boldsymbol{W}^*) = 0$. $|\sum_{i=1}^p\sum_{j=1}^p x_i x_j \rho_{ij}^*| \le C$ for some constant C and any unit vector $x = (x_1, \ldots, x_p)'$, where $\boldsymbol{\rho}^* = (\rho_{ij}^*)_{p\times p} \equiv corr(\boldsymbol{W}^*, \boldsymbol{W}^*)$.

In Assumption 3, we assume for simplicity that $\boldsymbol{W}$ and $\boldsymbol{W}^*$ are independent although it can be relaxed to weak dependence. The second condition on the correlation matrix of $\boldsymbol{W}^*$ is equivalent to $E(x'\boldsymbol{W}^*)^2 \le C$, which indicates that the idiosyncratic errors are cross-sectionally weakly dependent. This includes the sparsity condition for $\boldsymbol{\rho}^*$ where most entries of $\boldsymbol{\rho}^*$ are zeros as a special case, which restricts that W_i^* can only be correlated with a finite number of W_j^*'s that may not be closely located around asset i. We also need technical conditions for $\boldsymbol{\epsilon}$, but we leave it to Section 2.2 for notational convenience.

2.2. Portfolio weights

The target of this subsection is to construct a finite number of portfolios that can sufficiently span the factor space of a large panel. Mathematically, it is equivalent to introducing an (orthogonal) linear transform from R^p to R^d ($r_0 < d \ll p$) so that the high-dimensional noise-contaminated process $\boldsymbol{Y}$ is reduced to a low-dimensional process satisfying **(1)** the irrelevant idiosyncratic error is filtered out to a great extent; and **(2)** the rank of the volatility matrix of the linearly transformed processes (price dynamics of the portfolios) is r asymptotically. Here d is a fixed theoretical upper bound on r, which is typically assumed in the existing PCA-based approach.

Before determining the portfolio weights, we first introduce the data structure and some necessary notations. We assume that the data are discretely sampled from $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ where $'$ stands for the transpose, with equal sampling length, $\Delta_n = T/n$, where T is a fixed time horizon and n is the sample size. Mathematically, we will consider the asymptotic regime that $\Delta_n \to 0$. Let $Y_{i,k}$ be the observation of Y_i at time $k\Delta_n$, and $\Delta_j^n Y_i = Y_{i,j} - Y_{i,j-1}$. $\Delta_j^n X_i$'s and $\Delta_j^n \epsilon_i$'s are similarly defined as $\Delta_j^n Y_i$'s with Y_i's replaced by X_i's and ϵ_i's, respectively. Then $\Delta_j^n Y_i = \Delta_j^n X_i + \Delta_j^n \epsilon_i$ for $1 \le i \le p$ and $1 \le j \le n$.

Let $\boldsymbol{Q}$ be a $d \times p$ matrix with orthogonal rows representing portfolio weights. Then the return of the d portfolios are $\boldsymbol{Q}(\boldsymbol{X}_{t+h} - \boldsymbol{X}_t)$. If $\boldsymbol{Q}$ is uncorrelated with the increments of $\boldsymbol{W}$ and $\boldsymbol{W}^*$, $\boldsymbol{Q}(\boldsymbol{Z}^*_{t+h} - \boldsymbol{Z}^*_t)$ is asymptotically negligible compared to $\boldsymbol{Q}\boldsymbol{\beta}(\boldsymbol{Z}_{t+h} - \boldsymbol{Z}_t)$ under Assumptions 2–3. That is the d portfolios are truly diversified. For this reason, we divide the data set $(\Delta_1^n \boldsymbol{Y}, \ldots, \Delta_n^n \boldsymbol{Y})$ into two disjoint subsets of equal size, $\mathcal{Y}_1 = \{\Delta_1^n \boldsymbol{Y}, \ldots, \Delta_{n/2}^n \boldsymbol{Y}\}$ and $\mathcal{Y}_2 = \{\Delta_{n/2+1}^n \boldsymbol{Y}, \ldots, \Delta_n^n \boldsymbol{Y}\}$. We use the first set to determine the portfolio weights $\boldsymbol{Q}$ and the second to calculate portfolio returns using the weights $\boldsymbol{Q}$. We split the sample to ensure that $\boldsymbol{Q}$ is uncorrelated with $\boldsymbol{Z}^*_{t+h} - \boldsymbol{Z}^*_t$ and $\boldsymbol{\epsilon}_t$ in the second subsample, so that $\boldsymbol{Q}$ diversifies $\boldsymbol{Z}^*_t$ and $\boldsymbol{\epsilon}_t$ cross-sectionally under Assumptions 3 and 4. Then technically (A.41) and (A.42) pass through. To determine $\boldsymbol{Q}$, we split the data set $\mathcal{Y}_1$ into $[n/(2k_n')]$ non-overlapping blocks with each block containing k_n' increments of $\boldsymbol{Y}$. Here $[x]$ stands for the largest integer that is smaller than or equal to x. Now for $k = 1, \ldots, [n/(2k_n')]$, we define $\overline{\boldsymbol{Y}}_k = \sum_{j=1}^{k_n'} g_j \Delta_{k,j}^n \boldsymbol{Y}$ where $g_j = g(j/k_n)$ for some smooth function $g(\cdot)$ defined on $[0, 1]$ satisfying $g(0) = g(1) = 0$, and $\Delta_{k,j}^n \boldsymbol{Y} = \boldsymbol{Y}_{t_{k,j}} - \boldsymbol{Y}_{t_{k,j-1}}$ where $t_{k,j} = (k-1)k_n'\Delta_n + j\Delta_n$. Then a bias-uncorrected pre-averaging estimator of the integrated volatility matrix is $IV_n = \overline{\boldsymbol{Y}}\,\overline{\boldsymbol{Y}}'$ where $\overline{\boldsymbol{Y}} = (\overline{\boldsymbol{Y}}_1, \ldots, \overline{\boldsymbol{Y}}_{[n/(2k_n')]})$. A simple algebraic manipulation yields

$$\begin{aligned} IV_n = {} & \boldsymbol{\beta} IV \boldsymbol{\beta}' + \sum_{k=1}^{[n/(2k_n')]} \{\boldsymbol{\beta}(\overline{\boldsymbol{Z}}_k\overline{\boldsymbol{Z}}_k' - \int_0^1 g^2(s)ds \int_{(k-1)k_n'\Delta_n}^{kk_n'\Delta_n} \boldsymbol{\sigma}_t\boldsymbol{\sigma}_t' dt)\boldsymbol{\beta}' + \boldsymbol{\beta}\overline{\boldsymbol{Z}}_k\overline{\boldsymbol{Z}}_k^{*\prime} \\ & + \boldsymbol{\beta}\overline{\boldsymbol{Z}}_k\overline{\boldsymbol{\epsilon}}_k' + \overline{\boldsymbol{Z}}_k^*\overline{\boldsymbol{Z}}_k'\boldsymbol{\beta}' + \overline{\boldsymbol{Z}}_k^*\overline{\boldsymbol{Z}}_k^{*\prime} + \overline{\boldsymbol{Z}}_k^*\overline{\boldsymbol{\epsilon}}_k' + \overline{\boldsymbol{\epsilon}}_k\overline{\boldsymbol{Z}}_k'\boldsymbol{\beta}' + \overline{\boldsymbol{\epsilon}}_k\overline{\boldsymbol{Z}}_k^{*\prime} + \overline{\boldsymbol{\epsilon}}_k\overline{\boldsymbol{\epsilon}}_k'\} \end{aligned} \tag{2.5}$$

where $IV = \int_0^1 g^2(s)ds \int_0^{T/2} \boldsymbol{\sigma}_t\boldsymbol{\sigma}_t' dt$, $\overline{\boldsymbol{Z}}_k$, $\overline{\boldsymbol{Z}}_k^*$ and $\overline{\boldsymbol{\epsilon}}_k$ are similarly defined as $\overline{\boldsymbol{Y}}_k$ except that we replace $\boldsymbol{Y}$ by $\boldsymbol{Z}$, $\boldsymbol{Z}^*$, and $\boldsymbol{\epsilon}$, respectively. The bias-corrected pre-averaging estimator includes an additional term after IV_n correcting $\sum_{k=1}^{[n/(2k_n')]} \overline{\boldsymbol{\epsilon}}_k\overline{\boldsymbol{\epsilon}}_k'$. However, in finite sample, the bias-corrected estimator may not be non-negative definite causing trouble in eigen-decomposition. Therefore in the present paper, we choose to do eigen-decomposition on IV_n. (A.32) in the Appendix shows that the bias due to $\sum_{k=1}^{[n/(2k_n')]} \overline{\boldsymbol{\epsilon}}_k\overline{\boldsymbol{\epsilon}}_k'$ does not affect the spectral property in Theorem 1 and hence all the subsequent technical proofs.

Assumption 4. $\{\epsilon_{jt}\}_{t=1}^n$ is a stationary sequence of centered random variables independent of $\boldsymbol{X}$ and satisfies (1)

$$E\{\frac{1}{\sqrt{n}}\sum_{t=1}^n \epsilon_{jt}\}^{2d} \le C(\sigma_j^\epsilon)^{2d} := C[Var(\epsilon_{j1})]^d < \infty,$$

(2)

$$\| \left(\sum_{i=1}^{k_n'} (g_i - g_{i-1}) \sum_{l=1}^{k_n'} (g_l - g_{l-1}) \frac{1}{p} \sum_{j=1}^p E\{\epsilon_{jt_{k,i}}\epsilon_{jt_{m,l}}\} \right)_{[n/(2k_n')]\times[n/(2k_n')]} \| \le C(k_n')^{-1}$$

and (3)

$$E\{\frac{1}{k_n'}\sum_{i=1}^{k_n'}\sum_{l=1}^{k_n'}\frac{1}{\sqrt{p}}\sum_{j=1}^p [\epsilon_{jt_{k,i}}\epsilon_{jt_{m,l}} - E\epsilon_{jt_{k,i}}\epsilon_{jt_{m,l}}]\}^{2d} \le C.$$

Assumption 4 assumes that $\boldsymbol{\epsilon}_t$'s are cross-sectionally and temporally weakly dependent. Equation (1) of Assumption 4 is a moment (of order $2d$) convergence condition that is satisfied when ϵ_{jt}'s are m-dependent in time and $E(\frac{1}{\sqrt{n}}\sum_{t=1}^n \epsilon_{jt})^{2d} \le C$. Equation (2) of Assumption 4 is also satisfied when ϵ_{jt}'s are m-dependent in time and the second moment of ϵ_{j1} exists. Equation (3) of Assumption 4 is a moment convergence condition in panel (both temporally and cross-sectionally), which is satisfied when all entries of $\boldsymbol{\epsilon}$ are i.i.d. and $E\epsilon_{j1}^{4d} < \infty$. Similar weak dependence conditions for discrete panel of random variables can be found in Bai and Ng (2002) and Fan et al. (2013).

Assumptions 1–4 imply that the first term of (2.5) dominates other terms, and Lemma 1 proves that $\|IV_n/p - \boldsymbol{\beta} IV \boldsymbol{\beta}'/p\| = O_p\{p^{-1/2} + n^{-1/4}\}$. Motivated by this approximation, we could choose $\mathbf{Q}'$ as the $p \times d$ matrix of eigenvectors (listed in columns) of IV_n/p corresponding to the d largest eigenvalues sorted in descending order. Indeed we have the following result.

Theorem 1. *Let $\mathbf{Q}'$ be the $p \times d$ matrix of eigenvectors (listed in columns) of IV_n/p corresponding to the d largest eigenvalues sorted in descending order. Let $\mathbf{Q}_1'$ be the $p \times r$ submatrix of $\mathbf{Q}'$ consisting of the first r columns of $\mathbf{Q}'$ and $\mathbf{Q}_2'$ be the submatrix consisting of the last $d - r$ columns of $\mathbf{Q}'$, and $k_n' \to \infty$. Under Assumptions 1–4,*

$$E\|\mathbf{Q}_1' - \boldsymbol{\beta}\mathbf{H}/\sqrt{p}\|^{2d} I\{\|\boldsymbol{\Lambda}_n^{-1}\| \le C\} \le C(p^{-1}(1 + n^2/k_n'^4) + k_n'\Delta_n + 1/k_n')^d, \tag{2.6}$$

for some $r \times r$ matrix $\mathbf{H}$ satisfying

$$E\|\mathbf{H}'\mathbf{H} - I_r\|^{2d} I\{\|\boldsymbol{\Lambda}_n^{-1}\| \le C\} \le C(p^{-1}(1 + n^2/k_n'^4) + k_n'\Delta_n + 1/k_n')^d, \tag{2.7}$$

and

$$E\|\mathbf{Q}_2\boldsymbol{\beta}/\sqrt{p}\|^{2d} I\{\|\boldsymbol{\Lambda}_n^{-1}\| \le C\} \le C(p^{-1}(1 + n^2/k_n'^4) + k_n'\Delta_n + 1/k_n')^d, \tag{2.8}$$

where $\boldsymbol{\Lambda}_n$ is a $r \times r$ diagonal matrix of the r largest eigenvalues of IV_n/p sorted in descending order.

$$P\left(\|\boldsymbol{\Lambda}_n^{-1}\| \le C\right) \to 1. \tag{2.9}$$

Theorem 1 demonstrates that the carefully tuned asset allocation vectors (rows of $\mathbf{Q}$) have the property that (1) the first r vectors are asymptotically equivalent to $\boldsymbol{\beta}$ up to some orthonormal rotation, and (2) the last $d - r$ vectors are perpendicular to columns of $\boldsymbol{\beta}$ asymptotically. And hence the resulting diversified portfolios could cover the factor space and in particular the first r portfolios span the factor space, asymptotically. The term $O_p(k_n'\Delta_n)$ is due to the discretization error while $O_p(1/k_n')$ is due to the microstructure noise. To balance these two, we can set $k_n' = \theta[\Delta_n^{-1/2}]$ for some $\theta > 0$ and then the optimal rate becomes $p^{-d} + \Delta_n^{d/2}$.

2.3. Portfolio returns and pre-averaging

In last subsection, we have determined the portfolio weights. If we use the portfolio weights in a second stage, the portfolio returns are $\Delta_j^n\tilde{\boldsymbol{Y}}$'s where $\Delta_j^n\tilde{\boldsymbol{Y}} = \mathbf{Q}\Delta_{n/2+j}^n\boldsymbol{Y}$ for $j = 1, \ldots, n/2$. Proposition 1 with $k_n = 1$ demonstrates that although cross-sectional filtering dampens the microstructure noise a lot, it is not enough to achieve the optimal rate in Theorem 2 later, and hence we need to further reduce the microstructure noise. To this end, we use the pre-averaging technique on $\Delta_j^n\tilde{\boldsymbol{Y}}$.

We split the $\Delta_j^n\tilde{\boldsymbol{Y}}$'s into $n' =: [n/(2k_n)]$ non-overlapping blocks with each block containing k_n one step increments of the portfolios. Let $\tilde{\boldsymbol{X}}_t = \mathbf{Q}\boldsymbol{\beta}\boldsymbol{Z}_t$ and $\tilde{\boldsymbol{X}}_t^* = \mathbf{Q}\boldsymbol{Z}_t^*$. The smoothed data in the kth block is obtained via the following weighted average.

$$\begin{aligned}\overline{\tilde{\boldsymbol{Y}}}_k =: \sum_{j=1}^{k_n} g_j \Delta_{k,j}^n\tilde{\boldsymbol{Y}} &= \sum_{j=1}^{k_n} g_j \left(\Delta_{k,j}^n\tilde{\boldsymbol{X}} + \Delta_{k,j}^n\tilde{\boldsymbol{X}}^* + \Delta_{k,j}^n\tilde{\boldsymbol{\epsilon}}\right)\\ &= \overline{\tilde{\boldsymbol{X}}}_k + \overline{\tilde{\boldsymbol{X}}^*}_k + \overline{\tilde{\boldsymbol{\epsilon}}}_k,\end{aligned}$$

for $k = 1, \ldots, [n/(2k_n)]$, where $\Delta_{k,j}^n\boldsymbol{\Theta} = \boldsymbol{\Theta}_{t_{k,j}} - \boldsymbol{\Theta}_{t_{k,j-1}}$ for d-dimensional vectors $\boldsymbol{\Theta} = \tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{\epsilon}}, \tilde{\boldsymbol{X}}^*$.

Proposition 1. *Under the conditions in Theorem 1,*

$$E\|\overline{\tilde{\boldsymbol{X}}}_k/\sqrt{p}\|^{2d} I\{\|\Lambda_n^{-1}\| \le C\} \le C(\frac{k_n}{n})^d,\ E\|\overline{\tilde{\boldsymbol{X}}^*}_k/\sqrt{p}\|^{2d} I\{\|\Lambda_n^{-1}\| \le C\} \le C(\frac{k_n}{pn})^d,$$
$$E\|\overline{\tilde{\boldsymbol{\epsilon}}}_k/\sqrt{p}\|^{2d} I\{\|\Lambda_n^{-1}\| \le C\} \le C(k_n p)^{-d}.$$

For simplicity of presentation and to reduce the number of tuning parameters, we let $k_n = k_n'$. To balance the biases due to the idiosyncratic component and microstructure noise, we choose $k_n = [\theta n^{1/2}]$ for some $\theta > 0$. Then, by Proposition 1, we have

$$\frac{\overline{\tilde{\boldsymbol{Y}}}_k}{\sqrt{pk_n\Delta_n}} = \frac{\overline{\tilde{\boldsymbol{X}}}_k}{\sqrt{pk_n\Delta_n}} + O_p(p^{-1/2}). \tag{2.10}$$

Proposition 1 implies that the constructed portfolios are truly diversified meaning that the idiosyncratic risk term $\tilde{\boldsymbol{X}}^*$ is asymptotically negligible compared with the systematic risk term $\tilde{\boldsymbol{X}}$ which has spot volatility matrix $\mathbf{Q}\boldsymbol{\beta}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t'\boldsymbol{\beta}'\mathbf{Q}'$. Theorem 1 and Assumption 2 imply that $\mathbf{Q}\boldsymbol{\beta}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t'\boldsymbol{\beta}'\mathbf{Q}'$ has rank equal to r asymptotically. When p is large and Δ_n small, (2.10) demonstrates that the number of factors (dimension of factor space) is asymptotically equal to the rank of the volatility matrix of $\tilde{\boldsymbol{Y}}$ (price dynamics of the portfolios). This motivates us to estimate r based on $\{\frac{\overline{\tilde{\boldsymbol{Y}}}_k}{\sqrt{k_n\Delta_n}}\}_{k=1}^{[n/(2k_n)]}$.

2.4. Perturbation

(2.10) shows that $\{\frac{\overline{\tilde{Y}}_k}{\sqrt{k_n\Delta_n}}\}_{k=1}^{[n/(2k_n)]}$ can be regarded as noise-free without idiosyncratic component asymptotically. This fits naturally into the framework of Jacod and Podolskij (2013) by replacing the high-frequency returns there by $\{\frac{\overline{\tilde{Y}}_k}{\sqrt{k_n\Delta_n}}\}_{k=1}^{[n/(2k_n)]}$, the pre-averaged portfolio returns. The perturbation method given in this subsection is similar to that in Jacod and Podolskij (2013) in principle, but we have to take care of the biases due to cross-sectional smoothing and temporal pre-averaging. To dig out the information on r, it is natural to study the determinant of d-dimensional square matrix $(\overline{\tilde{Y}}_{j+1},\ldots,\overline{\tilde{Y}}_{j+d})$ which has rank r as $\Delta_n\to 0$ under Assumptions 1–4, see remarks below Theorem 1 and Proposition 1. When $r<d$, $det(\overline{\tilde{Y}}_{j+1},\ldots,\overline{\tilde{Y}}_{j+d})\to 0$ as $p,n\to\infty$ for all j due to a positive rank deficiency $d-r$. To tackle with the rank deficiency, a commonly used technique is to add a full rank matrix with small entries. To demonstrate this, let A and B be two deterministic (just for illustration) d-dimensional square matrix with A and B having rank r and d, respectively. Then, for some constant C_r, we have the following asymptotic expansion

$$det(A+hB)=h^{d-r}C_r+o(h^{d-r}),\quad h\to 0. \tag{2.11}$$

Thus, if $C_r\neq 0$, the rank deficiency (or the rank) of A floats out by the following two-perturbation-scale technique,

$$\frac{det^2(A+chB)}{det^2(A+hB)}\to c^{2(d-r)},\ \text{as } h\downarrow 0,\ \textit{for some } c>0. \tag{2.12}$$

Now, in parallel, we introduce the perturbation approach designed for $\overline{\tilde{Y}}_j$'s. We split the smoothed data set $\{\overline{\tilde{Y}}_j\}_{j=1}^{[n'/(2d)]}$ into $[n'/(2d)]$ non-overlapping blocks with each block containing $2d$ data. Let

$$\boldsymbol{R}_{l,i}^{(1)}=\overline{\tilde{\boldsymbol{Y}}}_{2(l-1)d+i}/\sqrt{p}+\nu_0 h(\tilde{\boldsymbol{W}}_{[2(l-1)d+i]k_n\Delta_n}-\tilde{\boldsymbol{W}}_{[2(l-1)d+i-1]k_n\Delta_n}) \tag{2.13}$$

for $l=1,\ldots,[n'/(2d)]$ and $i=1,\ldots,2d$, where $\tilde{\boldsymbol{W}}$ is a d-dimensional standard Brownian motion and ν_0 in (2.13) is some positive constant and $h=:h_n=\sqrt{k_n\Delta_n}$. Let

$$\boldsymbol{R}_{l,i}^{(2)}=(\overline{\tilde{\boldsymbol{Y}}}_{2(l-1)d+2i-1}+\overline{\tilde{\boldsymbol{Y}}}_{2(l-1)d+2i})/\sqrt{p}+\nu_0\sqrt{2}h(\tilde{\boldsymbol{W}}_{[2(l-1)d+2i]k_n\Delta_n}-\tilde{\boldsymbol{W}}_{[2(l-1)d+2(i-1)]k_n\Delta_n})$$

for $l=1,\ldots,[n'/(2d)]$ and $i=1,\ldots,d$.

3. Estimator and main results

3.1. Estimator

Inspired by (2.11)–(2.13), our basic starting statistics are,

$$S_t^{n,\kappa}=2dk_n\Delta_n\sum_{l=1}^{[n'/(2d)]}det^2(\mathcal{M}_l^{(\kappa)}),\ \kappa=1,2,$$

where

$$\mathcal{M}_l^{(1)}=\frac{1}{\sqrt{k_n\Delta_n}}\left(\boldsymbol{R}_{l,1}^{(1)},\boldsymbol{R}_{l,2}^{(1)},\ldots,\boldsymbol{R}_{l,d}^{(1)}\right)_{d\times d},$$
$$\mathcal{M}_l^{(2)}=\frac{1}{\sqrt{2k_n\Delta_n}}\left(\boldsymbol{R}_{l,1}^{(2)},\ldots,\boldsymbol{R}_{l,d}^{(2)}\right)_{d\times d}.$$

Our theory (see Theorem 2) shows that

$$\frac{S_t^{n,1}}{h^{2(d-r)}}\approx\int_{T/2}^t\Gamma_s ds,\ \frac{S_t^{n,2}}{(\sqrt{2}h)^{2(d-r)}}\approx\int_{T/2}^t\Gamma_s ds,\quad \textit{for}\ \ t\geq T/2, \tag{3.14}$$

where Γ_s is some stochastic process. This raises a natural point estimator of r as

$$\hat{r}=d+\log(S_t^{n,1}/S_t^{n,2})/\log 2.$$

To end this section, we define Γ_t and some related quantities precisely. For $i=1,\ldots,d$ and $\kappa=1,2$, we define, for $t\geq T/2$,

$$\alpha_i^{n,\kappa}(t)=\boldsymbol{Q}\boldsymbol{\beta}\frac{1}{\sqrt{p\kappa k_n\Delta_n}}\sum_{j=1}^{k_n}g_j\boldsymbol{\sigma}_t(\boldsymbol{W}_{t+((i-1)k_n+j)\kappa\Delta_n}-\boldsymbol{W}_{t+((i-1)k_n+j-1)\kappa\Delta_n})$$
$$\beta_i^{n,\kappa}(t)=\nu_0(\tilde{\boldsymbol{W}}_{t-T/2+i\kappa k_n\Delta_n}-\tilde{\boldsymbol{W}}_{t-T/2+(i-1)\kappa k_n\Delta_n})/\sqrt{\kappa k_n\Delta_n}.$$

Let

$$A_t^{(\kappa)} = (\alpha_1^{n,\kappa}(t), \ldots, \alpha_d^{n,\kappa}(t)), \text{ and } B_t^{(\kappa)} = (\beta_1^{n,\kappa}(t), \ldots, \beta_d^{n,\kappa}(t)).$$

Let (a, b) be a pair of non-negative integers and $\gamma_{a,b}(A_t^{(\kappa)}, B_t^{(\kappa)})$ be the sum of determinants over the set of matrices formed by choosing a and b columns from $A_t^{(\kappa)}$ and $B_t^{(\kappa)}$, respectively. Then we define

$$\begin{aligned} \Gamma_t &= E_{\mathcal{F}_t}\gamma_{r,d-r}^2(A_t^{(\kappa)}, B_t^{(\kappa)}) \\ \Gamma_t' &= E_{\mathcal{F}_t}\left(\gamma_{r,d-r}^2(A_t^{(\kappa)}, B_t^{(\kappa)}) - E_{\mathcal{F}_t}\gamma_{r,d-r}^2(A_t^{(\kappa)}, B_t^{(\kappa)})\right)^2 \\ \Gamma_t'' &= E_{\mathcal{F}_t}\left(\gamma_{r,d-r}^2(A_t^{(1)}, B_t^{(1)}) - E_{\mathcal{F}_t}\gamma_{r,d-r}^2(A_t^{(1)}, B_t^{(1)})\right) \\ &\quad \times \left(\gamma_{r,d-r}^2(A_t^{(2)}, B_t^{(2)}) - E_{\mathcal{F}_t}\gamma_{r,d-r}^2(A_t^{(2)}, B_t^{(2)})\right). \end{aligned} \tag{3.15}$$

Γ_t and Γ_t' are actually independent of κ, since $(\alpha_i^{n,1}(t), \beta_i^{n,1}(t))$ and $(\alpha_i^{n,2}(t), \beta_i^{n,2}(t))$ are identical in distribution. From Theorem 1 and remarks below Theorem 1 and Proposition 1, $A_t^{(\kappa)}$ is of rank r for all time t under Assumption 2. The term $B_t^{(\kappa)}$ containing the perturbed Brownian motion complements the rank deficiency of $A_t^{(\kappa)}$ in forming a combined matrix with r columns from $A_t^{(\kappa)}$ and the other $d-r$ columns from $B_t^{(\kappa)}$. As anticipated, we have

Proposition 2. *Under Assumptions 1–3,*

$$\int_{T/2}^{T}\Gamma_s ds > 0, \quad \int_{T/2}^{T}\left(\Gamma_s' - \Gamma_s''\right)ds > 0. \tag{3.16}$$

Proposition 2 demonstrates that the limit of $(S_t^{n,1}, S_t^{n,2})$ is positive making the logarithmic function in defining $\hat{r}$ meaningful as $n, p \to \infty$. This is further verified in our simulation studies.

3.2. Asymptotics on r

Our first theoretic result is on the asymptotics of our basic statistics $S_t^{n,1}$ and $S_t^{n,2}$.

Theorem 2. *Under Assumptions 1–3, if $k_n = [\theta\sqrt{n}]$,*

$$\left|\frac{S_t^{n,\kappa}}{(\sqrt{\kappa}h)^{2(d-r)}} - \int_{T/2}^{t}\Gamma_s ds\right| = O_p(p^{-1}\Delta_n^{-1/2} \vee \Delta_n^{1/4}), \tag{3.17}$$

and in particular if $p\Delta_n^{3/4} \to \infty$,

$$\frac{1}{\sqrt{k_n\Delta_n}}\left(\frac{S_t^{n,1}}{h^{2(d-r)}} - \int_{T/2}^{t}\Gamma_s ds, \frac{S_t^{n,2}}{(\sqrt{2}h)^{2(d-r)}} - \int_{T/2}^{t}\Gamma_s ds\right) \to^{\mathcal{L}_s} \mathcal{U}_t, \tag{3.18}$$

where $\mathcal{L}_s$ stands for stable convergence, and $\mathcal{U}_t \equiv (\mathcal{U}_{1t}, \mathcal{U}_{2t})$ is defined on an extension of $(\Omega, \mathcal{F}, (\mathcal{F})_{t\geq 0}, P)$ and is, conditionally on $\mathcal{F}$, a continuous centered Gaussian martingale with conditional (co)variances

$$V_t(\kappa, \kappa') = \begin{cases} 2d\int_{T/2}^{t}\Gamma_s' ds, & \kappa = \kappa' \\ 2d\int_{T/2}^{t}\Gamma_s'' ds, & \kappa \neq \kappa'. \end{cases} \tag{3.19}$$

In Jacod and Rosenbaum (2013), estimation of the volatility functional $\int_0^t g(\sigma_s^2)ds$ was summarized into two categories. As Γ_t defined in (3.15) is an expectation of some function of conditional normal variables, $S_t^{n,\kappa}$ is a statistic of the first category and hence has no additional bias term. From Theorem 2, the convergence rate of $S_t^{n,\kappa}$ depends simultaneously on n and p. The $O_p(p^{-1}\Delta_n^{-1/2})$ term comes from the bias due to cross-sectionally smoothing the idiosyncratic error and microstructure noise, and the $O_p(\Delta_n^{1/4})$ term is due to the asymptotic variance. It is interesting that the estimation error of the factor space (or the departure of $\mathbf{Q}_1'$ from $\boldsymbol{\beta}$) does not affect the convergence rate as long as $p\Delta_n^{3/4} \to \infty$. The condition $p\Delta_n^{3/4} \to \infty$ demonstrates that the dimensionality is a blessing rather than a curse in the context of estimating the factor number. This is intuitively interpretable since as p increases more strength from other assets are borrowed and more information on the factors are added into the estimation. This blessing of dimension was also found in Kong (2017) and Kong et al. (2015). As an estimator of the volatility functional $\int_{T/2}^t \Gamma_s ds$ of the price trajectories of the constructed finite portfolios, $\frac{S_t^{n,\kappa}}{h^{2(d-r)}}$ achieves the optimal rate $n^{-1/4}$, a standard result in estimating volatility functional with noisy high-frequency data, cf., Jacod et al. (2009) and Christensen et al. (2010), and Xiu (2010). In Assumption 1, we assume that the factor process $\boldsymbol{Z}_t$ contains a finite number of drift components, whose contribution is Δ_n in $\Delta_{kj}^n\tilde{\boldsymbol{X}}$ and $k_n\Delta_n$ in

$\overline{\tilde{\boldsymbol{X}}}_k$. This order is much smaller than $n^{-1/4}$ in Theorem 1 ($d = 1/2$ there) and $\sqrt{k_n\Delta_n}$ in Theorem 2. So the effect of the drift term is negligible here and hereafter in all main theorems on $\hat{r}$.

Next, we present the central limit theorem for the estimated number of common factors. Even in the noise-free setting, there is no central limit theorem for existing estimators of the common factor number in the literature.

Theorem 3. *Under the conditions in Theorem 2 and* $p\Delta_n^{3/4} \to \infty$,

$$\frac{1}{\sqrt{k_n\Delta_n}}(\hat{r} - r) \to^{\mathcal{L}_s} v_t^* \mathcal{N}(0, 1), \tag{3.20}$$

where $\mathcal{N}(0, 1)$ *is independent of* $\mathcal{F}$, *and* $v_t^{*2} = \frac{4d\left(\int_{T/2}^t \Gamma_s' ds - \int_{T/2}^t \Gamma_s'' ds\right)}{(\log 2 \int_{T/2}^t \Gamma_s ds)^2}$.

Theorem 3 not only demonstrates that $\hat{r}$ is consistent with rate $\sqrt{k_n\Delta_n}$, but also provides the asymptotic variance which serves as a measure of accuracy for the estimates. In establishing the central limit theorem, we need the condition $p\Delta_n^{3/4} \to \infty$, but for only the consistency, (3.17) shows that $p\Delta_n^{1/2} \to \infty$ suffices.

Theorem 3 is not applicable in statistical inference since it contains unknown quantities in the limiting conditional variance. Now we construct a consistent estimate of v_t^{*2}. Let

$$V_t^n(\kappa, \kappa') = 4d^2 k_n \Delta_n \sum_{l=1}^{[n'/2d]} det^2(\mathcal{M}_l^{(\kappa)}) det^2(\mathcal{M}_l^{(\kappa')}).$$

Then we have the following theorem.

Theorem 4. *Under the conditions in Theorem 2 and* $p\Delta_n^{1/2} \to \infty$,

$$\frac{V_t^n(\kappa, \kappa')}{(\sqrt{\kappa\kappa'} k_n \Delta_n)^{2(d-r)}} \to^P \begin{cases} 2d\left(\int_{T/2}^t \Gamma_s' ds + \int_{T/2}^t \Gamma_s^2 ds\right) & \kappa = \kappa', \\ 2d\left(\int_{T/2}^t \Gamma_s'' ds + \int_{T/2}^t \Gamma_s^2 ds\right) & \kappa \neq \kappa'. \end{cases} \tag{3.21}$$

This motivates us to estimate the limiting conditional variance by

$$\hat{v}_t^* \equiv \frac{1}{\log 2} \frac{\left(V_t^n(1, 1) + 2^{-2(d-\hat{r})} V_t^n(2, 2) - 2\sqrt{2}^{-2(d-\hat{r})} V_t^n(1, 2)\right)^{1/2}}{S_t^{n,1}}. \tag{3.22}$$

Then, by Theorem 4, $\hat{v}_t^*$ is a consistent estimate of v_t^*, and a straight consequence of Theorems 3 and 4 is the following corollary.

Corollary 1. *Under the conditions in Theorem 3,*

$$\frac{1}{\sqrt{k_n\Delta_n}\hat{v}_t^*}\left(\hat{r} - r\right) \to^{\mathcal{L}_s} \mathcal{N}(0, 1), \tag{3.23}$$

where $\mathcal{N}(0, 1)$ *is independent of* $\mathcal{F}$.

3.3. Testing on r

Now we propose a statistical test for

$$H_0 : r = r_0 \ v.s. \ H_1 : r \neq r_0. \tag{3.24}$$

In particular $r_0 = 1$ is related to testing for CAPM and $r_0 = 3$ to Fama–French 3-factor model. Based on the theory in Corollary 1, a natural testing rule is that we reject the null if

$$\hat{r} > r_0 + z_{1-\alpha/2}\hat{v}_t^*\sqrt{k_n\Delta_n}, \ \text{or} \ \hat{r} < r_0 - z_{1-\alpha/2}\hat{v}_t^*\sqrt{k_n\Delta_n},$$

where $z_{1-\alpha/2}$ is the upper $1 - \alpha/2$ quantile of the standard normal distribution. By Corollary 1, we have the following theorem which guarantees the size and power performance of the test.

Theorem 5. *Under the conditions in Corollary 1, as* $n, p \to \infty$,

$$P\left(\frac{|\hat{r} - r_0|}{\hat{v}_t^*\sqrt{k_n\Delta_n}} > z_{1-\alpha/2} | H_0\right) \to \alpha, \ \text{and} \ P\left(\frac{|\hat{r} - r_0|}{\hat{v}_t^*\sqrt{k_n\Delta_n}} > z_{1-\alpha/2} | H_1\right) \to 1. \tag{3.25}$$

In the context of hypothesis testing, one can replace $\hat{v}_t^*$ by $\tilde{v}_t^*$ which is similarly defined as $\hat{v}_t^*$ except for replacing $\hat{r}$ by r_0. Since $\tilde{v}_t^*$ converges under both hypotheses, though not to the true conditional variance under H_1, Theorem 5 still holds if one replaces $\hat{v}_t^*$ by $\tilde{v}_t^*$.

Table 1
Averaged estimates ($\hat{r}$), true standard errors (s.e.) and estimated standard errors (in the parentheses) of $\hat{r}$; Empirical sizes with nominal levels $\alpha = 5\%$ and 10%.

				Empirical sizes	
T	ν_0	$\hat{r}$	s.e.	$\alpha = 5\%$	$\alpha = 10\%$
44	0.30	3.00	0.55 (0.52)	0.06	0.11
	0.35	3.06	0.53 (0.51)	0.08	0.13
	0.40	2.96	0.54 (0.53)	0.05	0.09
	0.45	2.90	0.55 (0.55)	0.05	0.08
60	0.30	3.05	0.47 (0.44)	0.07	0.13
	0.35	2.98	0.45 (0.44)	0.06	0.09
	0.40	2.96	0.47 (0.45)	0.05	0.09
	0.45	2.90	0.46 (0.46)	0.04	0.07

4. Numerical experiments

4.1. Simulation studies

In this section, we conduct simulation studies to check the performance of the theory. The data set is generated from a 500-dimensional stochastic volatility model. The instantaneous volatility processes, σ_{it}^l's are generated independently from the following square-root processes,

$$(\sigma_{it}^l)^2 = 0.03(1 - (\sigma_{it}^l)^2)dt + 0.1|\sigma_{it}^l|dW_{it}^\sigma, \ \ l = 1, \ldots, r. \tag{4.26}$$

The idiosyncratic volatility process also follows the square-root stochastic differential equation.

$$(\sigma_{it}^*)^2 = \frac{0.08}{\sqrt{2}}(0.25\mu_0 - (\sigma_{it}^*)^2)dt + 0.2|\sigma_{it}^*|dW_{it}^{\sigma *}, \tag{4.27}$$

where μ_0 is a parameter controls the relative strength of the noise to signal. We start with $\mu_0 = 1$. We set $\boldsymbol{W}_t^*$ to be cross-sectionally 5-dependent with equal correlation $1/\sqrt{5}$. We set the first r elements of the first column of $\boldsymbol{\beta}$ be $\theta p^{1/2}$ and other elements be zero. For other columns, we set $\beta_{il} = p^{1/2}$ when $(l-1)p/r \le i \le lp/r$ and $2 \le l \le r$. Here θ controls the dominance of the largest eigenvalue of the covariation matrix.

First, we consider the no dominance case where $\theta = 1$. To check the size performance of our test on (3.24), we sample 44 or 66 days (two or three months) every-one-minute data from the continuous-time factor model (1.1), (4.26) and (4.27). We generate the microstructure noises from centered normal distribution with standard error equal to 0.0005, cf., Zhang et al. (2005). We use one month data to determine $\boldsymbol{Q}$ and the remaining data to calculate the portfolio returns. In pre-averaging, we first set $k_n = 5$ since for data up to minute scale the microstructure noise does not affect too much, hence we choose a relatively small k_n as a start. We set $g(x) = x \wedge (1-x)$. The simulations are repeated for 2000 replications. We define a measure of a relative strength of the perturbing Brownian motion to the idiosyncratic component and microstructure noise by the realized variance of the first component of $\nu_0 h\tilde{\boldsymbol{W}}_t$ divided by the $(r+1)$-st diagonal entry of the realized variance matrix of the d vector series $\bar{\tilde{\boldsymbol{Y}}}_k/\sqrt{p}$ (recall the definition of $\boldsymbol{R}_{l,i}^{(1)}$ in (2.13)). We call this relative strength as perturbing ratio.

We first set $r = r_0 = 3$, $d = 5$, and choose $\nu_0 = 0.3, 0.35, 0.4$ and 0.45, and correspondingly the averaged perturbing ratio (averaged over 2000 replications) ranges roughly from 10 to 25. Table 1 reports the estimated factor numbers, true and estimated standard errors and type I error probabilities committed by our test with nominal levels $\alpha = 5\%$ and 10%. We see that the estimated factor numbers are close to the true integer-valued number, that the estimated standard errors are close to the true standard errors, and that the type I errors are well controlled asymptotically. When rounded, all estimates are equal to the true factor number. Table 1 also demonstrates that the results are not much sensitive to the choice of ν_0 when it is in a reasonable range in terms of the perturbing ratio. See also Fig. 1 for illustration where $T = 44$, $\nu_0 \in [0.25, 0.5]$, and other parameters are kept the same as in Table 1. Choosing the optimal ν_0 theoretically is not easy and we leave this problem to our future research work.

Next, we fix $\nu_0 = 0.4$, $T = 66$, and consider $r_0 = 2, 4$ and 5. When $r_0 = 5$ we consider two cases: $d = 5$ and $d = 7$, and when $r_0 = 2, 4$ we fix $d = 5$. Table 2 displays parallel results to Table 1. Once again, the estimated numbers and standard errors are close to the true ones, and the type I errors are under control asymptotically. Again, when rounded, all our estimates are equal to the true factor numbers. As a comparison, in Table 2, we also include the AH estimates (Ahn and Horenstein (2013), a method working for large panel time series data. We see that the AH estimates work very well and equal to the true number exactly in all replications. However, it is not applicable in testing due to lack of second-order asymptotics, yet its consistency in the context of high-frequency data is not yet established.

Next, we check the sensitivity to k_n and d. Again, we fix $T = 66$ and $\nu_0 = 0.4$. First, we fix $k_n = 5$ and let d vary. The results are listed in the upper panel of Table 3. Then, we fix $d = 5$ and let $k_n = 10, 15$, and the results are given in the lower panel of Table 3. Table 3 demonstrates that the estimates are quite robust to d. But with fixed ν_0 they are a bit

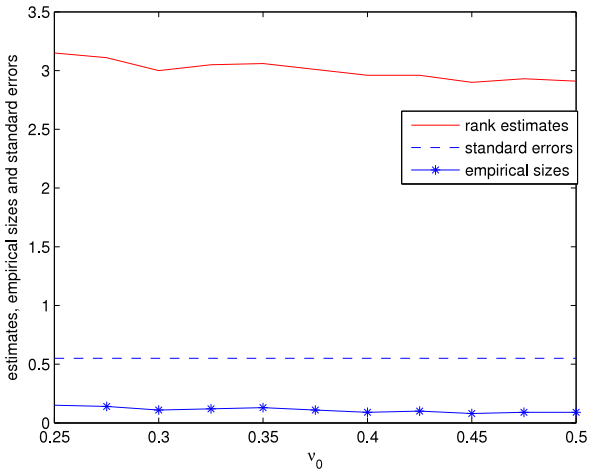

Fig. 1. The rank estimates, standard errors of the rank estimates, and empirical sizes of the test against ν_0.

Table 2
Averaged estimates ($\hat{r}$), true standard errors (s.e.) and estimated standard errors (in the parentheses) of $\hat{r}$; Empirical sizes with nominal levels $\alpha = 5\%$ and 10%. $\nu_0 = 0.4$ and $T = 66$. KLZ stands for our estimate; AH stands for the AH estimate.

	$\hat{r}$			Empirical sizes	
r_0	KLZ	AH	s.e.	$\alpha = 5\%$	$\alpha = 10\%$
2	2.03	2	0.45 (0.43)	0.06	0.10
4	3.88	4	0.46 (0.47)	0.03	0.07
5(d=5)	4.81	5	0.49 (0.49)	0.03	0.07
5(d=7)	4.86	5	0.67 (0.68)	0.05	0.08

Table 3
Averaged estimates ($\hat{r}$), true standard errors (s.e.) and estimated standard errors (in the parentheses) of $\hat{r}$; Empirical sizes with nominal levels $\alpha = 5\%$ and 10%. Upper panel: $k_n = 5$ while $d = 4, 6, 7$; Lower panel: $d = 5$ while $k_n = 10, 15$. $\nu_0 = 0.4$ and $T = 66$.

			Empirical sizes	
d ($k_n = 5$)	$\hat{r}$	s.e.	$\alpha = 5\%$	$\alpha = 10\%$
4	2.91	0.35 (0.35)	0.03	0.08
6	3.01	0.56 (0.53)	0.07	0.11
7	3.03	0.66 (0.62)	0.08	0.13
k_n ($d = 5$)				
10	2.83	0.62 (0.64)	0.04	0.08
15	2.75	0.72 (0.70)	0.03	0.05

Table 4
Averaged estimates ($\hat{r}$) and empirical sizes of our test with nominal levels $\alpha = 5\%$ and 10%; $k_n = 5$, $\nu_0 = 0.4$, $T = 66$, $d = 5$, $r_0 = 3$.

		Rejection rates	
r	$\hat{r}$	$\alpha = 5\%$	$\alpha = 10\%$
1	1.11	0.40	0.71
2	2.03	0.25	0.40
4	3.89	0.71	0.78
5	4.80	0.98	0.99

sensitive to k_n making $\hat{r}$ underestimate the true number and the test a bit conservative. However, all rounded estimates are equal to 3, which demonstrates the robustness of the rounded estimates to both k_n and d, as well as ν_0 shown in Tables 1–3.

Now, we check the power of our test via simulation. We generate data from model (1.1) (4.26) (4.27) with $r = 1, 2, 4, 5$. We set $T = 66$, $k_n = 5$, $\nu_0 = 0.4$, $d = 5$ and $r_0 = 3$. The estimates of r and rejection rates are presented in Table 4, from which we see clear departure from the null $r_0 = 3$.

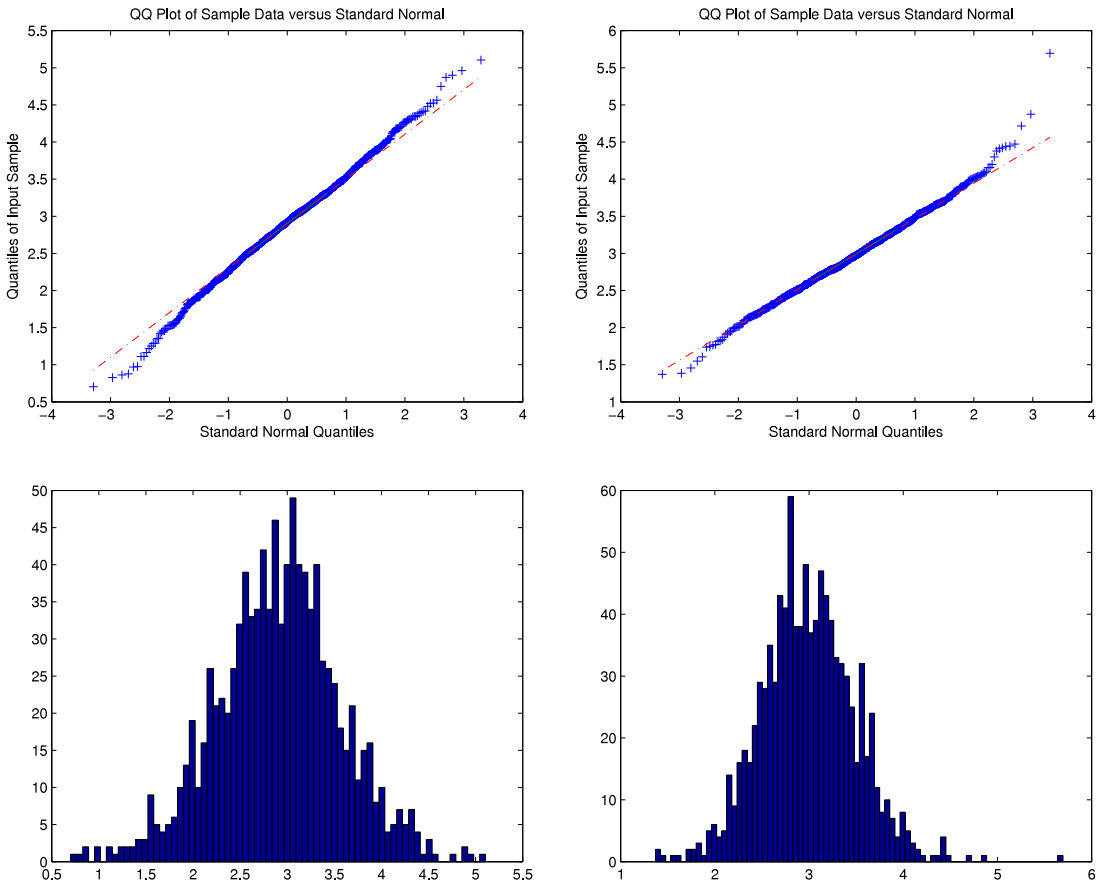

Fig. 2. Q–Q plots and histograms of the rank estimates; Left panels: 2 months; Right panels: 3 months; $r = r_0 = 3$ and $d = 5$.

Next, we check the asymptotic normality of the rank estimates. Fig. 2 draws the Q–Q plots and histograms of the estimates. This shows the normal approximation is quite good which is consistent with Theorem 3.

Finally, we check the robustness of the rank estimate to the signal-noise ratio and dominance of the first few largest eigenvalues. We fix $T = 44$, $\nu_0 = 0.3$, $d = 5$, $r = 3$, but let μ_0 increase from 1 to 5 which implies that the signal-noise ratio is decreasing. In Fig. 3, we draw the curves of the rank estimates and empirical sizes of the test with nominal level 10%, against μ_0. We see that the rank estimates are close to the true number and the rounded estimates equal to the true factor number exactly. The empirical sizes distort from the nominal level as μ_0 increases, but they are well-controlled under 10%. Next, we fix $\mu_0 = 1$ and other parameters as above, but let θ increase from 1 to 3.5 which demonstrates that the largest eigenvalue of the covariation matrix becomes more and more dominant to the weak factors, making the eigenvalue-ratio test tends to estimate r as 1 even if the true number is larger than 1. Fig. 4 plots the rank estimates, the standard errors of the rank estimates, the empirical sizes of the test at nominal level 10%, and the AH estimates $\hat{r}_{AH}$ as well as the standard errors. The rank estimates, the standard errors and empirical sizes are quite robust to the dominance to weak factors, while the AH estimate tends to underestimate r when some factors are weak. The failure of the AH estimator to weak factors was also found in Pelger (2018) and tackled by introducing a perturbed-eigenvalue-ratio estimator.

4.2. Real data analysis

In this section, we implement our test on a real data set consisting of three months (1/4 years) of one-minute data of the S&P 500 constituent stocks. The data set starts from January 6th 2016. Those with missing values are swept away and finally 484 stocks are left in the data set. In the analysis, we also remove extremely large returns and overnight returns from the data set. We consider $k_n = 5, 10, 15$ and set $d = 5$. We use one month data to find the portfolio weights and the remaining two months data to calculate the portfolio returns.

ν_0 serves as a tuning parameter. When ν_0 is small, the factor return dominates the perturbation term. Then $S_t^{n,1} \approx S_t^{n,2}$, and hence our method tends to overestimate the number of factors and the estimates are around d. When ν_0 increases, the rank estimate is expected to trend down. This is well observed in Fig. 5 (left panels). We choose ν_0 from the set of grid points in $[1, 7]$ with grid length 0.1. In order not to veil the signal of the factor component and to be stronger than the

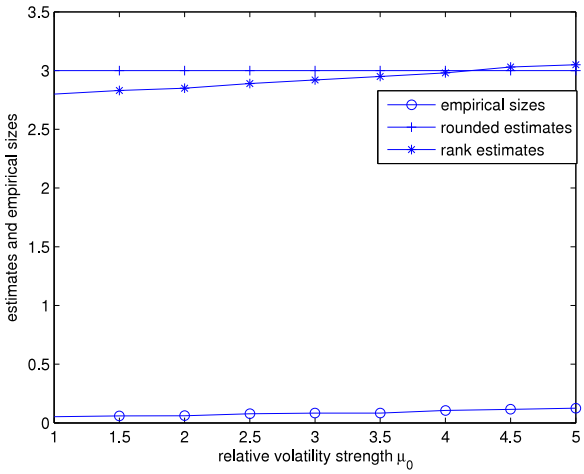

Fig. 3. The rank estimates (*), the rounded rank estimates (+) and empirical sizes (o) against μ_0.

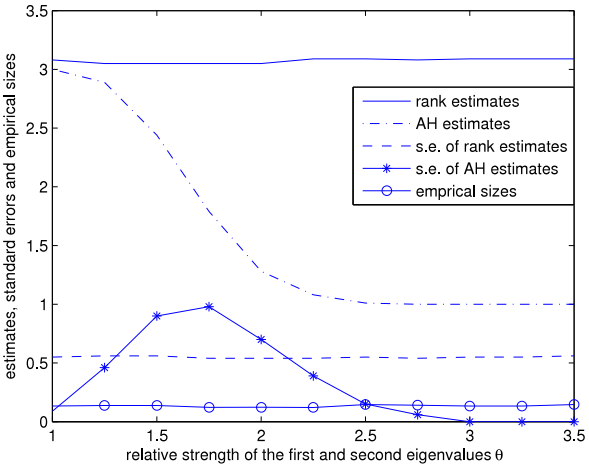

Fig. 4. The rank estimates (-), AH estimates (.-), the standard errors of the rank estimates (–) and AH estimates (*-), and empirical sizes (o-) against μ_0.

Table 5
Estimated factor numbers and test statistics testing for $H_0 : r_0 = 1$, $H_0 : r_0 = 2$ and $H_0 : r_0 = 3$ with empirically optimized ν_0. $\hat{r}_p = \hat{r}_{AH} = 1$.

k_n	$\hat{r}$	STA $r_0 = 1$	STA $r_0 = 2$	STA $r_0 = 3$
5	0.91	0.80	−3.00	−4.25
10	1.11	0.16	1.56	−3.41
15	0.97	−0.54	−2.68	−4.40

idiosyncratic component, we tune ν_0 such that the perturbing ratio (replace r by Pelger (2018)'s estimate) ranges from 1 to 4.

With the aid of Pelger (2018)'s estimator, we define a mean-square-error like measure as $G(\nu_0) = (\hat{r} - \hat{r}_p)^2 + k_n \Delta_n \hat{v}_T^{*2}$, where $\hat{r}_p$ stands for Pelger (2018)'s estimate. The first square term in $G(\nu_0)$ controls the bias while the second term refers to the asymptotic variance. Now, we empirically determine ν_0 as the minimizer of $G(\nu_0)$ for ν_0 lying in the grid so that the perturbing ratio is in [1, 4]. Table 5 reports the estimated factor numbers and test statistics for $H_0 : r = 1$, $H_0 : r = 2$ and $H_0 : r = 3$ with empirically optimized ν_0. It turns out $\hat{r}_p = \hat{r}_{AH} = 1$ for all k_n's. A comparison of the statistics with different k_n, we see that the results are affected by the microstructure noise. But with a slight smoothing, all results seem reasonable. Our empirical lesson from Table 5 is that the data evidence supports the CAPM model while significantly reject the Fama–French three factor model.

Fig. 5 plots the estimated number, and the mean squared error against ν_0 for perturbing ratio in [1, 4] in the left panels. The right panels draw test statistics against ν_0 for $H_0 : r = 1$, $H_0 : r = 3$ and $H_0 : r = 4$. We see that in the neighborhood

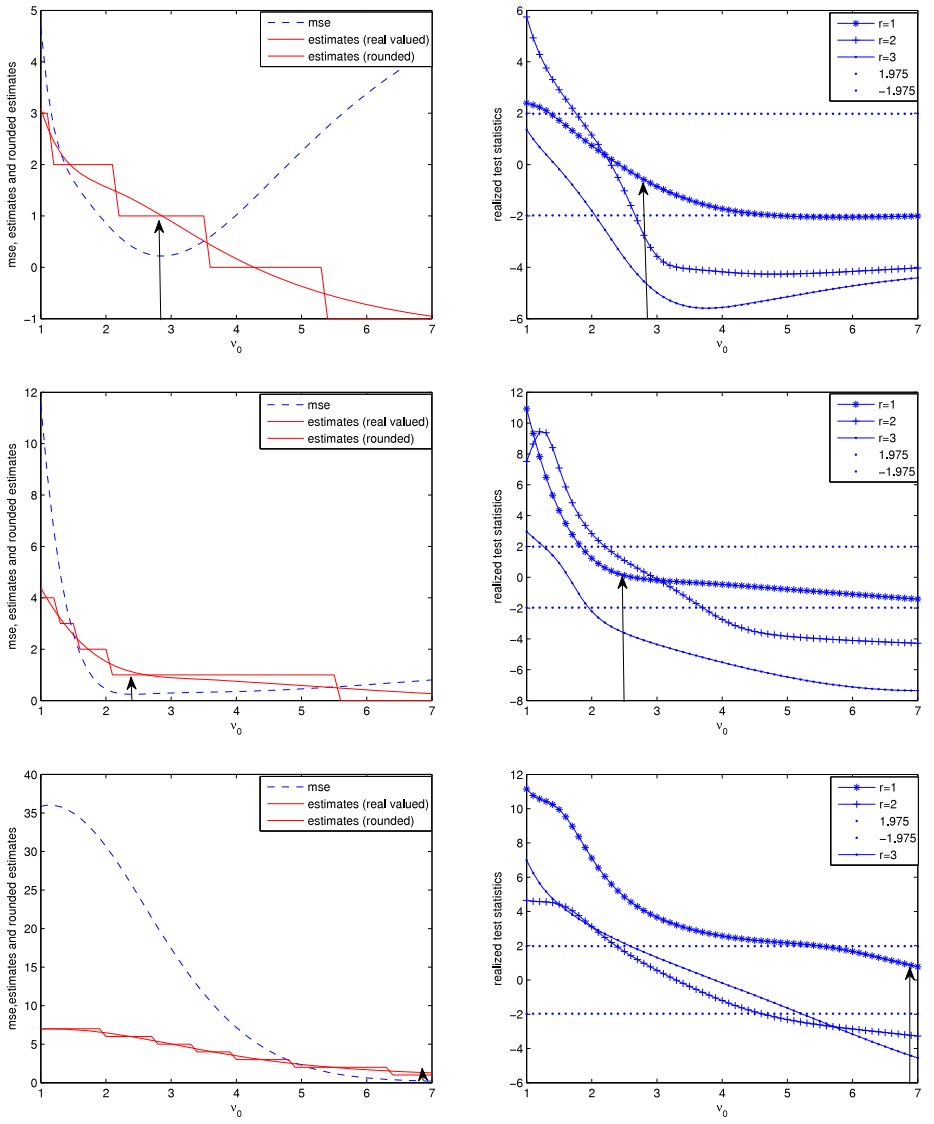

Fig. 5. In the left panels display the (rounded) rank estimates (-), mean squared error (- -) as functions of ν_0, where the top, middle and bottom panels correspond to $k_n = 15$, 10 and 5, respectively; In the right panels display the realized test statistics against ν_0 for $H_0 : r_0 = 1$ (*), $H_0 : r_0 = 2$ (+) and $H_0 : r_0 = 3$ (.), where the top, middle and bottom panels correspond to $k_n = 15$, 10 and 5, respectively. The arrows point to the optimal ν_0 minimizing the mse.

of the minima of the mean squared error curve the rounded estimates are 1 which is consistent to the Pelger (2018) and AH estimates. The test statistic curves show clear evidence to reject the two or three factor models and favor the CAPM model. The tests can be regarded as a second-order verification of the Pelger (2018) and AH estimates since observed from the test statistics the data supports $\hat{r}_p = \hat{r}_{AH} = 1$ for all k_n.

5. Conclusions

In this paper, we proposed a portfolio construction approach to reduce a large panel to a finite number of portfolios. We showed that these portfolios are asymptotically diversified. Based on a hybrid use of this factor space invariant dimension reduction technique and the perturbation technique, we proposed a rank test of the number of factors. There are some directions to extend this work. First, since the constructed principal portfolios could span the factor space of the large panel high-frequency return data, they could be useful to predict future indices, asset prices, or even volatilities. As in Stock and Watson (2002), one can simply regress the return features of the indices or assets of interest on the lagged returns (or absolute returns) of the principal portfolios. Second, the present paper did not consider jump factors which are empirically found and theoretically important in econometrics, cf., Jing et al. (2012b,a), Kong et al. (2015), Li et al. (2017)

and Pelger (2018). One could consider testing for continuous or jump factors (or both) when price jumps are present. Third, we did not consider observable factors if any other than the latent factors. In this case, determining diversifiable portfolio weights would be interesting and more challenging. Fourth, choosing ν_0 in a theoretically sound data-driven way is not easy to reach. Finally, the empirical results in Aït-Sahalia and Xiu (2017) show that the number of factors can be time-varying. By separating the whole sample into pieces and estimating the factor numbers piece by piece, it is possible to propose a test for the stability of the factor numbers. We leave all these problems to future works.

Appendix

In the sequel, C stands for a generic constant that may take different value at different appearance. By the standard localization method, we assume throughout the proof that

$$\max_{1\le i,j\le r}|\sigma_{ij}|+\max_{1\le i\le p}|\sigma^*_{ii}|\le C. \tag{A.28}$$

Let $\boldsymbol{\Gamma}=\int_0^{T/2}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t'dt$ and $\boldsymbol{H}=\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}'\boldsymbol{\beta}'\mathbf{Q}_1'\boldsymbol{\Lambda}_n^{-1}/\sqrt{p}$, where

$$\overline{\boldsymbol{Z}}=(\overline{\boldsymbol{Z}}(1),\ldots,\overline{\boldsymbol{Z}}([n/(2k_n')]))=(\overline{\boldsymbol{Z}}_1,\ldots,\overline{\boldsymbol{Z}}_r)'.$$

Let $\overline{\boldsymbol{Z}}^*=(\overline{\boldsymbol{Z}}^*(1),\ldots,\overline{\boldsymbol{Z}}^*([n/(2k_n')]))=(\overline{\boldsymbol{Z}}^*_1,\ldots,\overline{\boldsymbol{Z}}^*_p)'$ and

$$\overline{\boldsymbol{\epsilon}}=(\overline{\boldsymbol{\epsilon}}(1),\ldots,\overline{\boldsymbol{\epsilon}}([n/(2k_n')]))=(\overline{\boldsymbol{\epsilon}}_1,\ldots,\overline{\boldsymbol{\epsilon}}_p)'.$$

Let $\boldsymbol{\Lambda}$ be a $r\times r$ diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Gamma}$ sorted in descending order. We start with a lemma of its own interest.

Lemma 1. *Under Assumptions 1–4,*

$$E\parallel\{\boldsymbol{\beta}\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{*\prime}+\boldsymbol{\beta}\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{\epsilon}}'+\overline{\boldsymbol{Z}}^*\,\overline{\boldsymbol{Z}}'\boldsymbol{\beta}'+\overline{\boldsymbol{Z}}^*\,\overline{\boldsymbol{Z}}^{*\prime}+\overline{\boldsymbol{Z}}^*\,\overline{\boldsymbol{\epsilon}}'+\overline{\boldsymbol{\epsilon}}\,\overline{\boldsymbol{Z}}'\boldsymbol{\beta}'+\overline{\boldsymbol{\epsilon}}\,\overline{\boldsymbol{Z}}^{*\prime}+\overline{\boldsymbol{\epsilon}}\,\overline{\boldsymbol{\epsilon}}'\}$$
$$\mathbf{Q}_1'\boldsymbol{\Lambda}_n^{-1}/p\parallel^{2d}I(\|\boldsymbol{\Lambda}_n^{-1}\|\le C)\le C\{p^{-d}(1+(\frac{n}{k_n'^2})^{2d})+(k_n'\Delta_n)^d+(k_n')^{-d}\}.$$

Proof. We prove the lemma term by term. For the first term, by Assumption 2 on $\boldsymbol{\beta}$,

$$E\|\frac{\boldsymbol{\beta}\overline{\boldsymbol{Z}}\overline{\boldsymbol{Z}}^{*\prime}}{p}\|^{2d}\le CE\|\frac{\overline{\boldsymbol{Z}}\overline{\boldsymbol{Z}}^{*\prime}}{\sqrt{p}}\|^{2d}\le C\sum_{l=1}^{r}E\|\frac{1}{p}\sum_{j=1}^{p}(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k)\overline{\boldsymbol{Z}}_j^*(k))^2\|^d$$
$$\le C\sum_{l=1}^{r}\frac{1}{p^d}\sum_{j_1=1}^{p}\cdots\sum_{j_d=1}^{p}E(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k)\overline{\boldsymbol{Z}}_{j_1}^*(k))^2\cdots(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k)\overline{\boldsymbol{Z}}_{j_d}^*(k))^2$$
$$\le C\max_{1\le m,l\le d}E(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k)\overline{\boldsymbol{Z}}_m^*(k))^{2d}\le C(k_n'\Delta_n)^d, \tag{A.29}$$

where the last inequality is due to the Burkhölder–Davis–Gundy inequality. For the second term, by Assumption 2 on $\boldsymbol{\beta}$ again,

$$E\|\frac{\boldsymbol{\beta}\overline{\boldsymbol{Z}}\overline{\boldsymbol{\epsilon}}'}{p}\|^{2d}\le E\|\frac{\overline{\boldsymbol{Z}}\overline{\boldsymbol{\epsilon}}'}{\sqrt{p}}\|^{2d}\le C\sum_{l=1}^{r}E\|\frac{1}{p}\sum_{j=1}^{p}(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k)\overline{\boldsymbol{\epsilon}}_j(k))^2\|^d$$
$$\le C\max_{1\le l,m\le d}E(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k)\overline{\boldsymbol{\epsilon}}_m(k))^{2d}.$$

Let $\overline{\boldsymbol{Z}}_l(k,1)=\sum_{i=1}^{k_n'}g_i\int_{t_{k,i-1}}^{t_{k,i}}b_{lt}dt$ and $\overline{\boldsymbol{Z}}_l(k,2)=\sum_{i=1}^{k_n'}g_i\int_{t_{k,i-1}}^{t_{k,i}}\boldsymbol{\sigma}_{lt}'d\boldsymbol{W}_t$ where b_{lt} and $\boldsymbol{\sigma}_{lt}'$ are the lth row of $\boldsymbol{b}_t$ and $\boldsymbol{\sigma}_t$, respectively. By independence between $\boldsymbol{Z}$ and $\boldsymbol{\epsilon}$ and (1) of Assumption 4,

$$E(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k,1)\overline{\boldsymbol{\epsilon}}_m(k))^{2d}\le C\Delta_n^d,\ E(\sum_{k=1}^{[n/(2k_n')]}\overline{\boldsymbol{Z}}_l(k,2)\overline{\boldsymbol{\epsilon}}_m(k))^{2d}\le C(k_n')^{-d}.$$

The above two equations show that

$$E\|\frac{\boldsymbol{\beta}\overline{\boldsymbol{Z}}\overline{\boldsymbol{\epsilon}}'}{p}\|^{2d}\le C(k_n')^{-d}. \tag{A.30}$$

For the third term, its upper bound is the same as that of the first one because it is only the transpose of the first term. For the fourth one,

$$\begin{aligned}
&\|\frac{\overline{\boldsymbol{Z}}^*\overline{\boldsymbol{Z}}^{*\prime}}{p}\|^{2d} = \{\|\frac{\overline{\boldsymbol{Z}}^*\overline{\boldsymbol{Z}}^{*\prime}}{p}\|^2\}^d = \{\frac{\sup_{\|x\|=1} x'\overline{\boldsymbol{Z}}^*\overline{\boldsymbol{Z}}^{*\prime}x}{p}\}^d \\
&= \{\sup_{\|x\|=1}\sum_{k=1}^{[n/(2k_n')]}(\frac{\sum_{j=1}^p x_j\overline{\boldsymbol{Z}}_j^*(k)}{\sqrt{p}})^2\}^d \le C\{\sup_{\|x\|=1}\sum_{k=1}^{[n/(2k_n')]}E_{\mathcal{F}_{t_{k,0}}}(\frac{\sum_{j=1}^p x_j\overline{\boldsymbol{Z}}_j^*(k)}{\sqrt{p}})^2\}^d \\
&\quad + C\{\sup_{\|x\|=1}\sum_{k=1}^{[n/(2k_n')]}[(\frac{\sum_{j=1}^p x_j\overline{\boldsymbol{Z}}_j^*(k)}{\sqrt{p}})^2 - E_{\mathcal{F}_{t_{k,0}}}(\frac{\sum_{j=1}^p x_j\overline{\boldsymbol{Z}}_j^*(k)}{\sqrt{p}})^2]\}^d \\
&=: C\{\sup_{\|x\|=1} M_{1,x}\}^d + C\{\sup_{\|x\|=1}\sum_{k=1}^{[n/(2k_n')]} M_{2,x}(k)\}^d,
\end{aligned}$$

where $x = (x_1, \ldots, x_p)'$ is a p-vector. By (A.28), Assumption 3 and Itô's isometry,

$$\{\sup_{\|x\|=1} M_{1,x}\}^d \le C\sup_{\|x\|=1}|\frac{1}{p}\sum_{j_1=1}^p\sum_{j_2=1}^p x_{j_1}x_{j_2}\rho_{j_1j_2}^*|^d \le Cp^{-d}.$$

Notice that $\sum_{k=1}^{[n/(2k_n')]} M_{x,2}(k)$ is the end point of a continuous-time martingale and that

$$\begin{aligned}
&\max_{\|x\|=1}\sum_{k=1}^{[n/(2k_n')]} E_{\mathcal{F}_{t_{k,0}}}M_{x,2}^2(k) \\
&\le \sum_{k=1}^{[n/(2k_n')]}\{E_{\mathcal{F}_{t_{k,0}}}(\frac{1}{\sqrt{p}}\sum_{j=1}^p x_j\overline{\boldsymbol{Z}}_j^*(k))^4 + (E_{\mathcal{F}_{t_{k,0}}}(\frac{1}{\sqrt{p}}\sum_{j=1}^p x_j\overline{\boldsymbol{Z}}_j^*(k))^2)^2\} \le C\frac{k_n'\Delta_n}{p^2},
\end{aligned}$$

where the last inequality is due to Assumption 3. Then by change of time and the Burkhölder–Davis–Gundy inequality for continuous-time martingales,

$$E\{\max_{\|x\|=1}\sum_{k=1}^{[n/(2k_n')]} M_{x,2}(k)\}^d \le C(\frac{k_n'\Delta_n}{p^2})^{d/2}.$$

Combining this with the upper bound for $M_{x,1}$,

$$E\|\frac{\overline{\boldsymbol{Z}}^*\overline{\boldsymbol{Z}}^{*\prime}}{p}\|^{2d} \le Cp^{-d}. \tag{A.31}$$

The upper bound for the sixth term is the same as that of the second. We proceed to prove the last inequality first and come back to the proof for the fifth and seventh terms later. For the last term,

$$E\|\frac{\overline{\boldsymbol{\epsilon}}\overline{\boldsymbol{\epsilon}}'}{p}\|^{2d} \le E\|\frac{\overline{\boldsymbol{\epsilon}}'\overline{\boldsymbol{\epsilon}}}{p} - E\frac{\overline{\boldsymbol{\epsilon}}'\overline{\boldsymbol{\epsilon}}}{p}\|_F^{2d} + \|E\frac{\overline{\boldsymbol{\epsilon}}'\overline{\boldsymbol{\epsilon}}}{p}\|^{2d}.$$

By (2) of Assumption 4 and the condition on the function g,

$$\|E\frac{\overline{\boldsymbol{\epsilon}}'\overline{\boldsymbol{\epsilon}}}{p}\|^{2d} = \|E\left(\frac{1}{p}\sum_{j=1}^p\sum_{i=1}^{k_n'}(g_i - g_{i-1})\epsilon_j(t_{k,i})\sum_{l=1}^{k_n'}(g_l - g_{l-1})\epsilon_j(t_{m,l})\right)_{k,m=1}^{[n/(2k_n')]}\|^{2d} \le C(k_n')^{-2d}.$$

By (3) of Assumption 4

$$\begin{aligned}
&E\|\frac{\overline{\boldsymbol{\epsilon}}'\overline{\boldsymbol{\epsilon}}}{p} - E\frac{\overline{\boldsymbol{\epsilon}}'\overline{\boldsymbol{\epsilon}}}{p}\|_F^{2d} = E\{\sum_{k=1}^{[n/(2k_n')]}\sum_{j=1}^{[n/(2k_n')]}(\frac{\overline{\boldsymbol{\epsilon}}'(k)\overline{\boldsymbol{\epsilon}}(j) - E\overline{\boldsymbol{\epsilon}}'(k)\overline{\boldsymbol{\epsilon}}(j)}{p})^2\}^d \\
&\le C[\frac{n}{2k_n'}]^{2d}\max_{1\le m\le d}E(\frac{\overline{\boldsymbol{\epsilon}}'(k_m)\overline{\boldsymbol{\epsilon}}(j_m) - E\overline{\boldsymbol{\epsilon}}'(k_m)\overline{\boldsymbol{\epsilon}}(j_m)}{p})^{2d} \\
&\le C[\frac{n}{2k_n'}]^{2d}\max_{1\le m\le d}p^{-d}E(\frac{1}{\sqrt{p}}\sum_{j=1}^p\{\overline{\boldsymbol{\epsilon}}_j(k_m)\overline{\boldsymbol{\epsilon}}_j(j_m) - E\overline{\boldsymbol{\epsilon}}_j(k_m)\overline{\boldsymbol{\epsilon}}_j(j_m)\})^{2d} \le C(\frac{n}{k_n'^2})^{2d}p^{-d}.
\end{aligned}$$

The above two equations prove that

$$E\|\frac{\overline{\boldsymbol{\epsilon}}\overline{\boldsymbol{\epsilon}}'}{p}\|^{2d} \le C\{(\frac{n}{k_n'^2})^{2d}p^{-d} + (k_n')^{-2d}\}. \tag{A.32}$$

(A.32) and (A.31) together and the Hölder inequality imply that

$$E\{\|\frac{\overline{\boldsymbol{Z}}^*\overline{\boldsymbol{\epsilon}}'}{p}\|^{2d} + \|\frac{\overline{\boldsymbol{\epsilon}}\overline{\boldsymbol{Z}}^{*\prime}}{p}\|^{2d}\} \le C\{(\frac{n}{k_n'^2})^{d}p^{-d} + (k_n')^{-d}p^{-d/2}\}. \tag{A.33}$$

Combining the results for all terms proves Lemma 1.

Proof of Theorem 1. By the definition of $\mathbf{Q}_1$, we have $IV_n\mathbf{Q}_1'/p = \mathbf{Q}_1'\boldsymbol{\Lambda}_n$ which yields $\mathbf{Q}_1' = IV_n\mathbf{Q}_1'\boldsymbol{\Lambda}_n^{-1}/p$. By (2.5),

$$\begin{aligned}\mathbf{Q}_1' = &\{\boldsymbol{\beta}IV\boldsymbol{\beta}' + \boldsymbol{\beta}(\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}' - IV)\boldsymbol{\beta}' + \boldsymbol{\beta}\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{Z}}^{*\prime} + \boldsymbol{\beta}\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{\epsilon}}' + \overline{\boldsymbol{Z}}^*\overline{\boldsymbol{Z}}'\boldsymbol{\beta}' \\ &+ \overline{\boldsymbol{Z}}^*\,\overline{\boldsymbol{Z}}^{*\prime} + \overline{\boldsymbol{Z}}^*\,\overline{\boldsymbol{\epsilon}}' + \overline{\boldsymbol{\epsilon}}\,\overline{\boldsymbol{Z}}'\boldsymbol{\beta}' + \overline{\boldsymbol{\epsilon}}\,\overline{\boldsymbol{Z}}^{*\prime} + \overline{\boldsymbol{\epsilon}}\,\overline{\boldsymbol{\epsilon}}'\}\mathbf{Q}_1'\boldsymbol{\Lambda}_n^{-1}/p,\end{aligned} \tag{A.34}$$

where $IV = \int_0^1 g^2(s)ds\int_0^{T/2}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t'dt$. Let $\boldsymbol{H} = IV\boldsymbol{\beta}'\mathbf{Q}_1'\boldsymbol{\Lambda}_n^{-1}/\sqrt{p}$, then the first equation in Theorem 1 is proved by Lemma 1 and the standard econometric theory on realized variance matrix showing that

$$E\|\overline{\boldsymbol{Z}}\overline{\boldsymbol{Z}}' - IV\|^{2d} \le C(k_n\Delta_n)^d, \tag{A.35}$$

cf., Christensen et al. (2010).

Next, we proceed to prove the second equation. Let $\tilde{\mathbf{Q}}_1$ and $\boldsymbol{\Lambda}$ be respectively the $p \times r$ eigenvector matrix and eigenvalue matrix of $\boldsymbol{\beta}IV\boldsymbol{\beta}'/p$. By the triangular inequality,

$$\begin{aligned}&\|\boldsymbol{H}'\boldsymbol{H} - I_r\|^{2d} = \|(\boldsymbol{\Lambda}_n^{-1}\mathbf{Q}_1\boldsymbol{\beta}IV\boldsymbol{\beta}')(\boldsymbol{\beta}IV\boldsymbol{\beta}'\mathbf{Q}_1'\boldsymbol{\Lambda}_n^{-1})/p^2 - I_r\|^{2d} \\ \le\ &\|(\boldsymbol{\Lambda}_n^{-1}\mathbf{Q}_1\boldsymbol{\beta}IV\boldsymbol{\beta}')(\boldsymbol{\beta}IV\boldsymbol{\beta}'\mathbf{Q}_1'\boldsymbol{\Lambda}_n^{-1}) - (\boldsymbol{\Lambda}_n^{-1}\tilde{\mathbf{Q}}_1\boldsymbol{\beta}IV\boldsymbol{\beta}')(\boldsymbol{\beta}IV\boldsymbol{\beta}'\tilde{\mathbf{Q}}_1'\boldsymbol{\Lambda}_n^{-1})\|^{2d}/p^{4d} \\ &+ \|(\boldsymbol{\Lambda}_n^{-1}\tilde{\mathbf{Q}}_1\boldsymbol{\beta}IV\boldsymbol{\beta}')(\boldsymbol{\beta}IV\boldsymbol{\beta}'\tilde{\mathbf{Q}}_1'\boldsymbol{\Lambda}_n^{-1}) - (\boldsymbol{\Lambda}^{-1}\tilde{\mathbf{Q}}_1\boldsymbol{\beta}IV\boldsymbol{\beta}')(\boldsymbol{\beta}IV\boldsymbol{\beta}'\tilde{\mathbf{Q}}_1'\boldsymbol{\Lambda}^{-1})\|^{2d}/p^{4d} \\ =:\ &H_1 + H_2.\end{aligned}$$

By Lemma 1 and (A.35), the $SIN(\theta)$ theorem in Davis and Kahan (1970) and Weyl's eigenvalue theorem,

$$E\|\mathbf{Q}_1 - \tilde{\mathbf{Q}}_1\|^{2d} \le E(\frac{\|IV_n/p - \boldsymbol{\beta}IV\boldsymbol{\beta}'/p\|}{\min_{1\le l\le r}|\lambda_l - \lambda_{l-1}|})^{2d} \le C\{(k_n')^{-d} + (k_n'\Delta_n)^d + p^{-d}(1 + (\frac{n}{k_n'^2})^{2d})\} \tag{A.36}$$

$$E\|\boldsymbol{\Lambda}_n - \boldsymbol{\Lambda}\|^{2d} \le CE\|\frac{IV_n - \boldsymbol{\beta}IV\boldsymbol{\beta}'}{p}\|^{2d} \le C\{(k_n')^{-d} + (k_n'\Delta_n)^d + p^{-d}(1 + (\frac{n}{k_n'^2})^{2d})\}. \tag{A.37}$$

(A.37) and Assumption 2 prove the last equation of the theorem and that

$$E(\|\boldsymbol{\Lambda}_n\|^{2d} + \|\boldsymbol{\Lambda}_n^{-1}\|^{2d} + \|\boldsymbol{\Lambda}\|^{2d} + \|\boldsymbol{\Lambda}^{-1}\|^{2d}) \le C. \tag{A.38}$$

(A.37) and Assumption 2 show that

$$\begin{aligned}&E\|\boldsymbol{\Lambda}_n^{-1} - \boldsymbol{\Lambda}^{-1}\|^{2d}I(\boldsymbol{\Lambda}_n^{-1} \le C) \le E\|\boldsymbol{\Lambda}_n^{-1}\|^{2d}\|\boldsymbol{\Lambda}^{-1}\|^{2d}\|\boldsymbol{\Lambda}_n - \boldsymbol{\Lambda}\|^{2d}I(\boldsymbol{\Lambda}_n^{-1} \le C) \\ \le\ &C\{(k_n')^{-d} + (k_n'\Delta_n)^d + p^{-d}(1 + (\frac{n}{k_n'^2})^{2d})\}.\end{aligned} \tag{A.39}$$

By (A.36), (A.39) and Assumption 2,

$$\begin{aligned}&E(H_1)I(\|\boldsymbol{\Lambda}_n^{-1}\| \le C) \le C\{(k_n')^{-d} + (k_n'\Delta_n)^d + p^{-d}(1 + (\frac{n}{k_n'^2})^{2d})\}, \\ &E(H_2)I(\|\boldsymbol{\Lambda}_n^{-1}\| \le C) \le C\{(k_n')^{-d} + (k_n'\Delta_n)^d + p^{-d}(1 + (\frac{n}{k_n'^2})^{2d})\}.\end{aligned}$$

This proves the second equation of the theorem.

By the first and second equations of Theorem 1, we have on $\{\|\boldsymbol{\Lambda}_n^{-1}\| \le C\}$,

$$\begin{aligned}E\|\mathbf{Q}_2\boldsymbol{\beta}/\sqrt{p}\|^{2d} &\le E(\|\mathbf{Q}_2\boldsymbol{\beta} - \mathbf{Q}_2\boldsymbol{\beta}\boldsymbol{H}\boldsymbol{H}'\|/\sqrt{p})^{2d} + E\|\mathbf{Q}_2(\boldsymbol{\beta}\boldsymbol{H}/\sqrt{p} - \mathbf{Q}_1')\boldsymbol{H}'\|^{2d} \\ &\le C\{(k_n')^{-d} + (k_n'\Delta_n)^d + p^{-d}(1 + (\frac{n}{k_n'^2})^{2d})\}.\end{aligned}$$

This proves the third equation of the theorem. ■

By Theorem 1, without loss of generality, we assume that $\{\|\boldsymbol{\Lambda}_n\| \le C\}$ throughout the proof.

Proof of Proposition 1. Simple manipulation yields

$$\|\widetilde{\overline{\boldsymbol{X}}}_k/\sqrt{p}\|^{2d} = \|\overline{\boldsymbol{Z}}_k'\boldsymbol{\beta}'(\mathbf{Q}_1'\mathbf{Q}_1 + \mathbf{Q}_2'\mathbf{Q}_2)\boldsymbol{\beta}\overline{\boldsymbol{Z}}_k/p\|^d. \tag{A.40}$$

The Burkhölder–Davis–Gundy inequality and the standard econometric theory on pre-averaging estimator (cf., Christensen et al. (2010) and Jacod et al. (2009) imply that $E|\overline{\boldsymbol{Z}}_k'\boldsymbol{\beta}'\boldsymbol{\beta}\overline{\boldsymbol{Z}}_k/p|^d \le C(k_n\Delta_n)^d$, then

$$E|\overline{\boldsymbol{Z}}_k'\boldsymbol{\beta}'\mathbf{Q}_1'\mathbf{Q}_1\boldsymbol{\beta}\overline{\boldsymbol{Z}}_k/p|^d \le C(k_n\Delta_n)^d.$$

For the second term of (A.40), the third equation of Theorem 1 proves that

$$E|\overline{\boldsymbol{Z}}_k'\boldsymbol{\beta}'\mathbf{Q}_2'\mathbf{Q}_2\boldsymbol{\beta}\overline{\boldsymbol{Z}}_k/p|^d I(\|\boldsymbol{\Lambda}_n^{-1}\| \le C) \le C(k_n\Delta_n)^d\{(k_n')^{-d} + (k_n'\Delta_n)^d + p^{-d}(1 + (\frac{n}{k_n'^2})^{2d})\}.$$

Let $\mathbf{Q}_{l\cdot}' = (Q_l(1), \ldots, Q_l(p))$ be the lth row of $\mathbf{Q}$. Assumption 3 and (A.28) prove that

$$E\|\mathbf{Q}_{l\cdot}'\overline{\boldsymbol{Z}}^*(k)/\sqrt{p}\|^{2d} = E\{E_{\mathcal{F}_{T/2}}\|\mathbf{Q}_{l\cdot}'\overline{\boldsymbol{Z}}^*(k)/\sqrt{p}\|^{2d}\} \le C(k_n\Delta_n/p)^d, \tag{A.41}$$

By Assumption 3 and (A.28) again,

$$E\|\mathbf{Q}_l'\overline{\boldsymbol{\epsilon}}_k/\sqrt{p}\|^{2d} \le C/(pk_n)^d. \tag{A.42}$$

(A.41) and (A.42) finish the proof of the last two equations of the proposition. ■

Proof of Proposition 2. A direct application of Lemma 3.1 of Jacod and Podolskij (2013) with a slight modification in adapting the local smoothing weights, g_j's, results in the proposition.

Preliminary calculus and decompositions Now, we introduce some more notations and useful decompositions. For short, we write $t_{i,j}^{n,l} = T/2 + (2(l-1)+i)k_n\Delta_n + j\Delta_n$, $\Delta_{i,j}^{n,l}\boldsymbol{\Theta} = \boldsymbol{\Theta}_{t_{i,j}^{n,l}} - \boldsymbol{\Theta}_{t_{i,j-1}^{n,l}}$ for $\boldsymbol{\Theta} = \boldsymbol{W}$, $\boldsymbol{\epsilon}$. Simple algebraic manipulations yield

$$\frac{\boldsymbol{R}_{l,i}^{(\kappa)}}{\sqrt{\kappa}h} = \alpha_{l,i}^{n,\kappa} + \sqrt{\kappa}h\beta_{l,i}^{n,\kappa} + \delta\gamma_{l,i}^{n,\kappa} + (\sqrt{\kappa}h)^2\delta_{l,i}^{n,\kappa}, \tag{A.43}$$

for $i = 1, \ldots, d$, $l = 1, \ldots, [n'/2d]$ and $\kappa = 1, 2$, where $\delta = \sqrt{\frac{1}{p}}$ and

$$\begin{aligned}
\alpha_{l,i}^{n,1} &= \frac{1}{\sqrt{pk_n\Delta_n}}\mathbf{Q}\boldsymbol{\beta}\boldsymbol{\sigma}_{t_{0,0}^{n,l}}\sum_{j=1}^{k_n} g_j\Delta_{i,j}^{n,l}\boldsymbol{W} =: \mathbf{Q}\boldsymbol{\beta}/\sqrt{p}\alpha_{l,i}^{*n,1},\\
\alpha_{l,i}^{n,2} &= \frac{1}{\sqrt{2pk_n\Delta_n}}\mathbf{Q}\boldsymbol{\beta}\boldsymbol{\sigma}_{t_{0,0}^{n,l}}\sum_{j=1}^{k_n} g_j(\Delta_{i,2j-1}^{n,l}\boldsymbol{W} + \Delta_{i,2j}^{n,l}\boldsymbol{W}) =: \mathbf{Q}\boldsymbol{\beta}/\sqrt{p}\alpha_{l,i}^{*n,2},\\
\boldsymbol{\beta}_{l,i}^{n,1} &= \frac{1}{\sqrt{p}k_n\Delta_n}\mathbf{Q}\boldsymbol{\beta}\{\sum_{j=1}^{k_n} g_j[\boldsymbol{b}_{t_{0,0}^{n,l}}\Delta_n + \int_{t_{i,j-1}^{n,l}}^{t_{i,j}^{n,l}}\int_{t_{0,0}^{n,l}}^{t}(\breve{\boldsymbol{\sigma}}_{t_{0,0}^{n,l}}d\boldsymbol{W}_u + \breve{\boldsymbol{\sigma}}_{t_{0,0}^{n,l}}^{\perp}d\boldsymbol{W}_u^*)d\boldsymbol{W}_t]\}\\
&\quad + \nu_0(\tilde{\boldsymbol{W}}_{t_{i,0}^{n,l}} - \tilde{\boldsymbol{W}}_{t_{i-1,0}^{n,l}})/h =: \tilde{\boldsymbol{\beta}}_{l,i}^{n,1} + M_{l,i}^{n,1},\\
\boldsymbol{\beta}_{l,i}^{n,2} &= \frac{1}{2\sqrt{p}k_n\Delta_n}\mathbf{Q}\boldsymbol{\beta}\{\sum_{j=1}^{k_n} g_j[\boldsymbol{b}_{t_{0,0}^{n,l}}2\Delta_n + \int_{t_{i,2(j-1)}^{n,l}}^{t_{i,2j}^{n,l}}\int_{t_{0,0}^{n,l}}^{t}(\breve{\boldsymbol{\sigma}}_{t_{0,0}^{n,l}}d\boldsymbol{W}_u + \breve{\boldsymbol{\sigma}}_{t_{0,0}^{n,l}}^{\perp}d\boldsymbol{W}_u^*)d\boldsymbol{W}_t]\}\\
&\quad + \nu_0(\tilde{\boldsymbol{W}}_{t_{2i,0}^{n,l}} - \tilde{\boldsymbol{W}}_{t_{2(i-1),0}^{n,l}})/(\sqrt{2}h) =: \tilde{\boldsymbol{\beta}}_{l,i}^{n,2} + M_{l,i}^{n,2},\\
\gamma_{l,i}^{n,1} &= \frac{1}{h}\mathbf{Q}\{\sum_{j=1}^{k_n} g_j[\int_{t_{i,j-1}^{n,l}}^{t_{i,j}^{n,l}}\boldsymbol{\sigma}_t^*d\boldsymbol{W}_t^* + \Delta_{i,j}^{n,l}\boldsymbol{\epsilon}]\},\\
\gamma_{l,i}^{n,2} &= \frac{1}{\sqrt{\kappa}h}\mathbf{Q}\{\sum_{j=1}^{k_n} g_j[\int_{t_{i,2(j-1)}^{n,l}}^{t_{i,2j}^{n,l}}\boldsymbol{\sigma}_t^*d\boldsymbol{W}_t^* + \Delta_{i,2j}^{n,l}\boldsymbol{\epsilon} + \Delta_{i,2j-1}^{n,l}\boldsymbol{\epsilon}]\},\\
\delta_{l,i}^{n,1} &= \frac{1}{\sqrt{p}h^3}\mathbf{Q}\boldsymbol{\beta}\{\sum_{j=1}^{k_n} g_j\int_{t_{i,j-1}^{n,l}}^{t_{i,j}^{n,l}}[(\boldsymbol{b}_t - \boldsymbol{b}_{t_{0,0}^{n,l}})dt\\
&\quad + \int_{t_{0,0}^{n,l}}^{t}(\breve{\boldsymbol{b}}_u du + (\breve{\boldsymbol{\sigma}}_u - \breve{\boldsymbol{\sigma}}_{t_{0,0}^{n,l}})d\boldsymbol{W}_u + (\breve{\boldsymbol{\sigma}}_u^{\perp} - \breve{\boldsymbol{\sigma}}_u^{\perp})d\boldsymbol{W}_u^*)]d\boldsymbol{W}_t\},
\end{aligned}$$

$$\delta_{l,i}^{n,2} = \frac{1}{\sqrt{p}(\sqrt{2}h)^3}\boldsymbol{Q}\boldsymbol{\beta}\{\sum_{j=1}^{k_n} g_j \int_{t_{l,2(j-1)}^{n,l}}^{t_{i,2j}^{n,l}} [(\boldsymbol{b}_t - \boldsymbol{b}_{t_{0,0}^{n,l}})dt + \int_{t_{0,0}^{n,l}}^{t} (\breve{\boldsymbol{b}}_u du + (\breve{\boldsymbol{\sigma}}_u - \breve{\boldsymbol{\sigma}}_{t_{0,0}^{n,l}})d\boldsymbol{W}_u + (\breve{\boldsymbol{\sigma}}_u^{\perp} - \breve{\boldsymbol{\sigma}}_u^{\perp})d\boldsymbol{W}_u^*)]d\boldsymbol{W}_t\}.$$

By Theorem 1, and repeated use of the Burkhölder–Davis–Gundy inequality and Hölder's inequality,

$$E_{\mathcal{F}_{l-1}}\left(\|\alpha_{l,i}^{n,\kappa}\|^{2d} + \|\beta_{l,i}^{n,\kappa}\|^{2d} + \|\gamma_{l,i}^{n,\kappa}\|^{2d} + \|\delta_{l,i}^{n,\kappa}\|^{2d}\right) \le C. \tag{A.44}$$

For $m \ge 1$, let $\mathcal{P}_m$ be the set of all multi-integers $\boldsymbol{p} = (p_1, \ldots, p_m)$ with $p_1 + \cdots + p_m = d$, and $\mathcal{I}_{\boldsymbol{p}}$ the set of all partitions $\boldsymbol{I} = (I_1, \ldots, I_m)$ of $\{1, \ldots, d\}$ such that I_j contains exactly p_j points. Let $G_{A_1,\ldots,A_m}^{\boldsymbol{I}}$ be the matrix whose ith column is the ith column of A_j when $i \in I_j$. Define

$$\gamma_{a,b,c,d}(A, B, C, D) = \sum_{\boldsymbol{I}\in\mathcal{I}_{(a,b,c,d)}} det(G_{A,B,C,D}^{\boldsymbol{I}})$$

and for short

$$\gamma_{a,b}(A, B) = \sum_{\boldsymbol{I}\in\mathcal{I}_{(a,b,0,0)}} det(G_{A,B,C,D}^{\boldsymbol{I}}).$$

Now we directly cite Lemma 6.1 of Jacod and Podolskij (2013).

Lemma 2 (*JP (2013)*). *For any $m \ge 1$ and $A_1, \ldots, A_m \in \mathcal{M}$,*

$$det\,(A_1 + \cdots + A_m) = \sum_{\boldsymbol{p}\in\mathcal{P}_m}\sum_{\boldsymbol{I}\in\mathcal{I}_{\boldsymbol{p}}} det(G_{A_1,\ldots,A_m}^{\boldsymbol{I}}), \tag{A.45}$$

where $\mathcal{M}$ is the set of $d \times d$ matrices.

For $\kappa = 1, 2$, let $A_l^{(\kappa)} = (\alpha_{l,1}^{n,\kappa}, \ldots, \alpha_{l,d}^{n,\kappa})$, $B_l^{(\kappa)} = (\beta_{l,1}^{n,\kappa}, \ldots, \beta_{l,d}^{n,\kappa})$, $C_l^{(\kappa)} = (\gamma_{l,1}^{n,\kappa}, \ldots, \gamma_{l,d}^{n,\kappa})$ and $D_l^{(\kappa)} = (\delta_{l,1}^{n,\kappa}, \ldots, \delta_{l,d}^{n,\kappa})$. We further decompose $B_l^{(\kappa)}$ as $\tilde{B}_l^{(\kappa)} + P_l^{(\kappa)}$ where

$$\tilde{B}_l^{(\kappa)} = (\tilde{\boldsymbol{\beta}}_{l,1}^{n,\kappa}, \ldots, \tilde{\boldsymbol{\beta}}_{l,d}^{n,\kappa}) \quad P_l^{(\kappa)} = (M_{l,1}^{n,\kappa}, \ldots, M_{l,d}^{n,\kappa}).$$

so that $P_l^{(\kappa)}$ is the matrix of perturbing Brownian motion terms. Let $P_{l,1}^{(\kappa)}$ and $P_{l,2}^{(\kappa)}$ be the submatrices of $P_l^{(\kappa)}$ consisting of the first r and last $d-r$ rows, respectively. Let $\alpha_{l,i,j}^{n,\kappa}$, $\tilde{\beta}_{l,i,j}^{n,\kappa}$ and $M_{l,i,j}^{n,\kappa}$ be the jth element of the d-vectors $\alpha_{l,i}^{n,\kappa}$, $\tilde{\beta}_{l,i}^{n,\kappa}$ and $M_{l,i}^{n,\kappa}$. By definition and Theorem 1,

$$E\{(\alpha_{l,i,j}^{n,\kappa})^{2d} + (\tilde{\beta}_{l,i,j}^{n,\kappa})^{2d}\} \le C(n^{-d/2} + p^{-d}), \quad \text{for } j = r+1, \ldots, d-r. \tag{A.46}$$

Now we have the following lemma.

Lemma 3. *Under Assumptions 1–4,*

$$\begin{aligned} &2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} E_{\mathcal{F}_{l-1}}\{\frac{\gamma_{r,d-r}^2(A_l^{(\kappa)}, \sqrt{\kappa}hB_l^{(\kappa)})}{(\sqrt{\kappa}h)^{2(d-r)}} - \gamma_{r,d-r}^2(A_l^{(\kappa)}, P_l^{(\kappa)})\} \\ &= O_p(n^{-1/2} + p^{-1/2}) + o_p(n^{-1/4}) \\ &\quad 2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} E_{\mathcal{F}_{l-1}}\gamma_{r,d-r}^2(A_l^{(\kappa)}, P_l^{(\kappa)}) \\ &= \int_{T/2}^{t} \Gamma_s ds + O_p(n^{-1/2} + p^{-1/2}) + o_p(n^{-1/4}). \end{aligned} \tag{A.47}$$

Proof. Let $S = (s_{ij})_{d\times d}$ and $U = (u_{ij})_{d\times d}$ be two arbitrary matrices that are composed of choosing r columns from $A_l^{(\kappa)}$ and $d-r$ columns from $B_l^{(\kappa)}$. Then $\gamma_{r,d-r}^2(A_l^{(\kappa)}, B_l^{(\kappa)})$ is the sum of all products of the form $det(S)det(U)$. By definition of the determinant

$$det(S) = \sum_{(j_1,\ldots,j_d)\in\{1,\ldots,d\}} (-1)^{1+j_1+2+j_2+\cdots+d+j_d} s_{1j_1}\cdots s_{dj_d}.$$

(A.44) shows that $E(s_{1j_1}\cdots s_{rj_r})\le C$. Now we take a closer look at s_{kj_k} for $k=r+1,\ldots,d$, which can take values of $\alpha^{n,\kappa}_{l,i,k}$ or $\tilde\beta^{n,\kappa}_{l,i,k}+M^{n,\kappa}_{l,i,k}$ for some $i=1,\ldots,d$. Simple calculation yields

$$\prod_{k=r+1}^{d} s_{kj_k}=\prod_{k=r+1}^{r+r^*}\alpha^{n,\kappa}_{l,i_k,j_k}\prod_{k=r^*+1}^{d}(\tilde\beta^{n,\kappa}_{l,i_k,j_k}+M^{n,\kappa}_{l,i_k,j_k})=:\prod_{k=r+1}^{r+r^*}\alpha^{n,\kappa}_{l,i_k,j_k}\sum_{r_0^*=1}^{d-r^*}s^*_{r_0^*},\tag{A.48}$$

where $(j_{r+1},\ldots,j_d)\in\{r+1,\ldots,d\}$ and $s^*_{r_0^*}$ is the term that contains r_0^* terms of the form $\tilde\beta^{n,\kappa}_{l,i_k,j_k}$. By (A.46), if $r^*+r_0^*\ge 2$, $E(\prod_{k=r+1}^d s_{kj_k})\le C(n^{-1/2}+p^{-1})$, then the expectation of the absolute value of all summands in $det(S)$ with $r^*+r_0^*\ge 2$ is less than $C(n^{-1/2}+p^{-1})$ and denote the sum of all summands in $det(S)$ with $r^*+r_0^*\ge 2$ by $det(S)_{l1}$. If $r^*+r_0^*=1$, then either $(r^*,r_0^*)=(1,0)$ or $(r^*,r_0^*)=(0,1)$. If $r^*+r_0^*=0$, then $(r^*,r_0^*)=(0,0)$. Let the sum of all summands in $det(S)$ with $(r^*,r_0^*)=(1,0)$, $(r^*,r_0^*)=(0,1)$ and $(r^*,r_0^*)=(0,0)$ by $det(S)_{l2}$, $det(S)_{l3}$ and $det(S)_{l4}$, respectively. Then $det(S)=\sum_{m=1}^4 det(S)_{lm}$ where l is a subscript having the same meaning as the l in the statement of the lemma. Similarly, we have the decomposition $det(U)=\sum_{m=1}^4 det(U)_{lm}$ where $det(U)_{lm}$'s are similarly defined as $det(S)_{lm}$ except that we replace S by U in the argument for S. Now

$$\begin{aligned}
&2dk_n\Delta_n\sum_{l=1}^{[n'/(2d)]}E_{\mathcal{F}_{l-1}}\frac{\gamma^2_{r,d-r}(A_l^{(\kappa)},\sqrt{\kappa}hB_l^{(\kappa)})}{(\sqrt{\kappa}h)^{2(d-r)}}=2dk_n\Delta_n\sum_{l=1}^{[n'/(2d)]}E_{\mathcal{F}_{l-1}}\gamma^2_{r,d-r}(A_l^{(\kappa)},B_l^{(\kappa)})\\
&=2dk_n\Delta_n\sum_{l=1}^{[n'/(2d)]}\sum_{(S,U)}E_{\mathcal{F}_{l-1}}det(S)det(U)\\
&=2dk_n\Delta_n\sum_{l=1}^{[n'/(2d)]}\sum_{(S,U)}E_{\mathcal{F}_{l-1}}\{det(S)_{l1}det(U)\\
&\quad+[det(S)_{l2}+det(S)_{l3}][det(U)_{l1}+det(U)_{l2}+det(U)_{l3}]\\
&\quad+[det(S)_{l2}+det(S)_{l3}]det(U)_{l4}+det(S)_{l4}det(U)_{l1}\\
&\quad+det(S)_{l4}[det(U)_{l2}+det(U)_{l3}]+det(S)_{l4}det(U)_{l4}\}=:\sum_{m=1}^{6}L_m.
\end{aligned}$$

By definition of $det(S)_{lm}$'s and $det(U)_{lm}$'s, (A.46) and (A.48), $E|L_1+L_2+L_4|\le C/\sqrt{n}$. For L_3, by again (A.48) the summand $(det(S)_{l2}+det(S)_{l3})det(U)_{l4}$ is an odd function of $M^{n,\kappa}_{l,i_k,j_k}$ for some $j_k\in\{r+1,\ldots,d\}$ and $i_k\in\{1,\ldots,d\}$, that being said, $E_{\mathcal{F}_{t^{n,\kappa}_{l,0}}}(det(S)_{l2}+det(S)_{l3})det(U)_{l4}=0$ and L_3 is a sum of martingale differences. Then by calculating the conditional variance of L_3, $L_3=O_p(n^{-1/2})$. Similarly, $L_5=O_p(n^{-1/2})$. By definition of Γ_s and standard econometric theory on estimating volatility functionals (cf., Jacod and Rosenbaum (2013)), the discretization error $L_6-\int_{T/2}^T\Gamma_s ds=o_p(n^{-1/4})$. Combining the above results finishes the proof of the lemma. ■

For short, let $\gamma^{(\kappa,l)}_{r-a,d-r+k,a-k-j,j}$ be $\gamma^{(\kappa,l)}_{r-a,d-r+k,a-k-j,j}(A_l^{(\kappa)},\sqrt{\kappa}hB_l^{(\kappa)},\delta C_l^{(\kappa)},(\sqrt{\kappa}h)^2D_l^{(\kappa)})$.

Lemma 4. *Under Assumptions 1–4, if $k_n=[\theta\sqrt{n}]$ and $\delta h^{-1}=o_p(1)$,*

$$2d\sqrt{k_n\Delta_n}\sum_{l=1}^{[n'/(2d)]}\frac{\gamma^{(\kappa,l)}_{r-a,d-r+k,a-k-j,j}\gamma^{(\kappa,l)}_{r-a',d-r+k',a-k'-j',j'}}{(\sqrt{\kappa}h)^{2(d-r)}}=O_p(\delta^2h^{-3})+o_p(1),$$

unless $(a,k,j)=(a',k',j')=0$, where all subindices under $\gamma^{(\kappa,l)}$ are non-negative.

Proof. By (A.44), we have

$$\begin{aligned}
&E_{\mathcal{F}_{l-1}}|\gamma^{(\kappa,l)}_{r-a,d-r+k,a-k-j,j}\gamma^{(\kappa,l)}_{r-a',d-r+k',a'-k'-j',j'}|/h^{2(d-r)}\\
&\le Ch^{k+k'}h^{2(j+j')}\delta^{a+a'-(k+k')-(j+j')}=:O_{\kappa,l}.
\end{aligned}\tag{A.49}$$

We prove this lemma by singling out several cases.

Case 1: $j+j'\ge 1$. In this case, $O_{\kappa,l}=O_p(h^2)$ which proves the lemma, and hence we only need to evaluate

$$\sqrt{k_n\Delta_n}\sum_{l=1}^{[n'/(2d)]}\gamma^{(\kappa,l)}_{r-a,d-r+k,a-k}\gamma^{(\kappa,l)}_{r-a',d-r+k',a'-k'}/(\sqrt{\kappa}h)^{2(d-r)}$$

in the sequel.

Case 2: $j = j' = 0$, $a + a' \geq 1$ and $k + k' \geq 0$. In this case, $O_{\kappa,l} \leq Ch^{a+a'}(\delta h^{-1})^{a+a'-k-k'}$. When $a + a' - k - k' > 0$, the lemma is obvious. Otherwise, $a + a' = k + k'$ and $O_{\kappa,l} = ch^{k+k'}$ and it suffices to consider $0 \leq k + k' \leq 1$. When $k + k' = 0$, $a + a' = 0$, which is a controversy to $a + a' \geq 1$. When $k + k' = 1$, $a + a' = 1$, then $(a, k) = (0, k)$ and $(a', k') = (1, 1 - k)$, or $(a, k) = (1, k)$ and $(a', k') = (0, 1 - k)$. In either case, $a - k < 0$ or $a' - k' < 0$ except that $k$or $k' = 0$. By symmetry, assume that $k = 0$, then $(a, k) = (0, 0)$ and $(a', k') = (1, 1)$, and then we need to evaluate

$$\sqrt{k_n\Delta_n}\sum_{l=1}^{[n'/(2d)]}\gamma_{r,d-r}^{(\kappa,l)}\gamma_{r-1,d-r+1}^{(\kappa,l)}/(\sqrt{\kappa}h)^{2(d-r)},$$

which is $O_p(h)$ by the martingale central limit theorem because $\gamma_{r,d-r}^{(\kappa,l)}\gamma_{r-1,d-r+1}^{(\kappa,l)}$ is either an odd function of $\boldsymbol{W}$ or $\boldsymbol{W}^*$.

Case 3: $j = j' = 0$, $a + a' \geq 1$ and $k + k' < 0$. The lemma is true obviously since $O_{\kappa,l} \leq C(h(\delta h^{-1})^2)$.

Case 4: Combining the results as before in cases 1–3, it suffices to consider $(a, k) = (0, k)$ and $(a', k') = (0, k')$, that is to evaluate

$$\sqrt{k_n\Delta_n}\sum_{l=1}^{[n'/(2d)]}\gamma_{r,d-r+k,-k}^{(\kappa,l)}\gamma_{r,d-r+k',-k'}^{(\kappa,l)}/(\sqrt{\kappa}h)^{2(d-r)},$$

which is $o_p(1)$ if $-(k, k') = (0, 1)$ or $-(k, k') = (1, 0)$ because $\gamma_{r,d-r+k,-k}^{(\kappa,l)}\gamma_{r,d-r+k',-k'}^{(\kappa,l)}$ is an odd function of $\tilde{\boldsymbol{W}}$. When $-k - k' \geq 2$, the left hand side of the lemma is $O_p(\delta^2 h^{-3})$. This proves the lemma. ■

Lemma 4 demonstrates that the principal term is $2dh\sum_{l=1}^{[n'/(2d)]}\frac{(\gamma_{r,d-r}^{(\kappa,l)})^2}{h^{2(d-r)}}$ if $\delta h^{-1} = o_p(1)$. Lemma 5 concludes this.

Lemma 5. *Under Assumptions 1–4, if $k_n = [\theta\sqrt{n}]$ and $\delta h^{-1} = o_p(1)$,*

$$2dh^2\sum_{l=1}^{[n'/(2d)]}\left(\frac{det^2(\mathcal{M}_l^{(\kappa)})}{(\sqrt{\kappa}h)^{2(d-r)}} - \frac{\gamma_{r,d-r}^2(A_l^{(\kappa)}, \sqrt{\kappa}hB_l^{(\kappa)})}{(\sqrt{\kappa}h)^{2(d-r)}}\right) = O_p((\delta h^{-1})^2) + o_p(h). \tag{A.50}$$

Proof. By the decompositions for $\boldsymbol{R}_{l,i}^{(\kappa)}$'s, we have

$$\mathcal{M}_l^{(\kappa)} = A_l^{(\kappa)} + \sqrt{\kappa}hB_l^{(\kappa)} + \delta C_l^{(\kappa)} + (\sqrt{\kappa}h)^2 D_l^{(\kappa)}. \tag{A.51}$$

Then Lemma 5 is a direct consequence of Lemmas 2 and 4. ■

Let $\zeta_{l,t}^{n,\kappa} = \gamma_{r,d-r}^2(A_l^{(\kappa)}, B_l^{(\kappa)})$. Now, we have the following lemma.

Lemma 6. *Under Assumptions 1–3, if $k_n = [\theta\sqrt{n}]$,*

$$2d\sqrt{k_n\Delta_n}\sum_{l=1}^{[n'/(2d)]}\left(\zeta_{l,t}^{n,1} - E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,1}, \zeta_{l,t}^{n,2} - E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,2}\right) \to^{\mathcal{L}_s} \mathcal{U}_t. \tag{A.52}$$

Proof. By simple calculation, we have

$$\begin{aligned}
&\sum_{l=1}^{[n'/(2d)]} 4d^2k_n\Delta_n E_{\mathcal{F}_{l-1}}\left(\zeta_{l,t}^{n,\kappa} - E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa}\right)^2 \\
&= 2d\sum_{l=1}^{[n'/(2d)]} E_{\mathcal{F}_{l-1}}\left(\gamma_{r,d-r}^2(A_l^{(\kappa)}, P_l^{(\kappa)}) - E_{\mathcal{F}_{l-1}}\gamma_{r,d-r}^2(A_l^{(\kappa)}, P_l^{(\kappa)})\right)^2 2dk_n\Delta_n \\
&= 2d\int_{T/2}^t \Gamma_s' ds + o_p(1).
\end{aligned} \tag{A.53}$$

and

$$\begin{aligned}
&\sum_{l=1}^{[n'/(2d)]} 4d^2k_n\Delta_n E_{\mathcal{F}_{l-1}}\left(\zeta_{l,t}^{n,1} - E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,1}\right)\left(\zeta_{l,t}^{n,2} - E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,2}\right) \\
&= 2d\sum_{l=1}^{[n'/(2d)]} 2dk_n\Delta_n E_{\mathcal{F}_{l-1}}\left(\gamma_{r,d-r}^2(A_l^{(1)}, P_l^{(1)}) - E_{\mathcal{F}_{l-1}}\gamma_{r,d-r}^2(A_l^{(1)}, P_l^{(1)})\right)
\end{aligned}$$

$$\times \left(\gamma^2_{r,d-r}(A_l^{(2)}, P_l^{(2)}) - E_{\mathcal{F}_{l-1}}\gamma^2_{r,d-r}(A_l^{(2)}, P_l^{(2)})\right) + o_p(1)$$
$$= 2d\int_{T/2}^{t} \Gamma_s'' ds + o_p(1). \tag{A.54}$$

A.1. Proof of the main results

Proof of Theorem 2. The first equation is a direct consequence of Lemma 3–5. The second equation is a straight result of Lemma 3–6.

Proof of Theorem 3. Simple algebraic manipulation yields

$$\hat{r} - r = \frac{\left(\log \frac{S_t^{n,1}}{h^{2(d-r)}} - \log \int_{T/2}^{t} \Gamma_s ds\right) - \left(\log \frac{S_t^{n,2}}{(\sqrt{2}h)^{2(d-r)}} - \log \int_{T/2}^{t} \Gamma_s ds\right)}{\log 2}. \tag{A.55}$$

Then by the delta method and the first equation of Theorem 2, the limiting distribution of $\frac{1}{\sqrt{k_n\Delta_n}}(\hat{r} - r)$ is equivalent to that of

$$\frac{\left(\frac{S_t^{n,1}}{h^{2(d-r)}} - \int_{T/2}^{t} \Gamma_s ds\right) - \left(\frac{S_t^{n,2}}{(\sqrt{2}h)^{2(d-r)}} - \int_{T/2}^{t} \Gamma_s ds\right)}{\sqrt{k_n\Delta_n}\int_{T/2}^{t} \Gamma_s ds \log 2}, \tag{A.56}$$

which converges stably to $\mathcal{N}(0, \frac{4d(\int_{T/2}^{t} \Gamma_s' ds - \int_{T/2}^{t} \Gamma_s'' ds)}{(\log 2 \int_{T/2}^{t} \Gamma_s ds)^2})$ due to the second equation of Theorem 2. ■

Proof of Theorem 4. By (A.51) and $\delta h^{-1} = o(1)$,

$$2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} \left(\frac{det^2(\mathcal{M}_l^{(\kappa)})det^2(\mathcal{M}_l^{(\kappa')})}{(\sqrt{\kappa\kappa'}h^2)^{2(d-r)}} - \gamma^2_{r,d-r}(A_l^{(\kappa)}, B_l^{(\kappa)})\gamma^2_{r,d-r}(A_l^{(\kappa')}, B_l^{(\kappa')})\right) = o_p(1). \tag{A.57}$$

By the argument and decomposition $det(S) = \sum_{m=1}^{4} det(S)_{lm}$ in Lemma 3,

$$2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} \left(\gamma^2_{r,d-r}(A_l^{(\kappa)}, B_l^{(\kappa)})\gamma^2_{r,d-r}(A_l^{(\kappa')}, B_l^{(\kappa')}) - \zeta_{l,t}^{n,\kappa}\zeta_{l,t}^{n,\kappa'}\right) = o_p(1). \tag{A.58}$$

By the martingale central limit theorem, we have

$$2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} \left(\zeta_{l,t}^{n,\kappa}\zeta_{l,t}^{n,\kappa'} - E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa}\zeta_{l,t}^{n,\kappa'}\right) = O_p(\sqrt{k_n\Delta_n}). \tag{A.59}$$

By (A.57) and (A.59), it suffices to prove that

$$2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa}\zeta_{l,t}^{n,\kappa'} \to^P \begin{cases} \int_{T/2}^{t} \Gamma_s' ds + \int_{T/2}^{t} \Gamma_s^2 ds & \kappa = \kappa' \\ \int_{T/2}^{t} \Gamma_s'' ds + \int_{T/2}^{t} \Gamma_s^2 ds & \kappa \neq \kappa' \end{cases} \tag{A.60}$$

By (3.15),

$$\begin{aligned} & 2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa}\zeta_{l,t}^{n,\kappa'} \\ = \; & 2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} \left(E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa}\zeta_{l,t}^{n,\kappa'} - E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa}E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa'}\right. \\ & \left. + E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa}E_{\mathcal{F}_{l-1}}\zeta_{l,t}^{n,\kappa'}\right) \\ = \; & o_p(1) + \begin{cases} 2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} \left(\Gamma'_{T/2+2(l-1)dk_n\Delta_n} + \Gamma^2_{T/2+2(l-1)dk_n\Delta_n}\right) & \kappa = \kappa' \\ 2dk_n\Delta_n \sum_{l=1}^{[n'/(2d)]} \left(\Gamma''_{T/2+2(l-1)dk_n\Delta_n} + \Gamma^2_{T/2+2(l-1)dk_n\Delta_n}\right) & \kappa \neq \kappa' \end{cases} \\ \to^P \; & \begin{cases} \int_{T/2}^{t} \Gamma_s' ds + \int_{T/2}^{t} \Gamma_s^2 ds & \kappa = \kappa' \\ \int_{T/2}^{t} \Gamma_s'' ds + \int_{T/2}^{t} \Gamma_s^2 ds & \kappa \neq \kappa'. \end{cases} \end{aligned} \tag{A.61}$$

Proof of Theorem 5. Simple manipulation yields

$$\frac{\hat{r}-r_0}{\hat{v}_t^*\sqrt{k_n\Delta_n}} = \mathcal{N}(0,1) + \frac{r-r_0}{\hat{v}_t^*\sqrt{k_n\Delta_n}}. \tag{A.62}$$

Under H_0, $\frac{r-r_0}{\hat{v}_t^*\sqrt{k_n\Delta_n}} = 0$, while under H_1, $|\frac{r-r_0}{\hat{v}_t^*\sqrt{k_n\Delta_n}}| \to \infty$ in probability. This proves Theorem 5.

References

Ahn, S., Horenstein, A., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.
Aït-Sahalia, Y., Jacod, J., 2012. High Frequency Financial Econometrics. Princeton University Press.
Aït-Sahalia, Y., Xiu, D., 2017. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. J. Econometrics 201, 384–399.
Aït-Sahalia, Y., Xiu, D., 2018. Principal component analysis of high frequency data. J. Amer. Statist. Assoc. (in press).
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.
Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure, and mean–variance analysis on large asset markets. Econometrica 51, 1281–1304.
Christensen, K., Kinnebrock, S., Podolskij, M., 2010. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. J. Econometrics 159, 116–133.
Connor, G., Korajczyk, R., 1993. A test for the number of factors in an approximate factor model. J. Finance 48, 1263–1291.
Dai, C., Lu, K., Xiu, D., 2018. Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data. J. Econometrics 208 (1), 43–79.
Davis, Kahan, 1970. The rotation of eigenvectors by a perturbation, *III**. SIAM J. Numer. Anal. 7, 1–46.
Fan, J., Furger, A., Xiu, D., 2016. Incorporating global industrial classification standard into portfolio allocation: a simple factor-based large covariance matrix estimator with high frequency data. J. Bus. Econom. Statist. 34, 489–503.
Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc. Ser. B Stat. Methodol. 75, 603–680.
Jacod, J., Li, Y., Mykland, P.A., Podolskij, M., Vetter, M., 2009. Microstructure noise in the continuous case: The pre-averaging approach. Stochastic Process. Appl. 119, 2249–2276.
Jacod, J., Podolskij, M., 2013. A test for the rank of the volatility process: the random perturbation approach. Ann. Statist. 41, 2391–2427.
Jacod, J., Rosenbaum, M., 2013. Quarticity and other functionals of volatility: Efficient estimation. Ann. Statist. 41, 1462–1484.
Jing, B.Y., Kong, X.B., Liu, Z., 2012a. Modeling high frequency financial data by pure jump models. Ann. Statist. 40, 759–784.
Jing, B.Y., Kong, X.B., Liu, Z., Mykland, P., 2012b. On the jump activity index on semimartingales. J. Econometrics 166, 213–223.
Jing, B.Y., Liu, Z., Kong, X.B., 2014. On the estimation of integrated volatility with jumps and microstructure noise. J. Bus. Econom. Statist. 32, 457–467.
Kim, D., Kong, X.B., Li, C., Wang, Y., 2018. Adaptive thresholding for large volatility matrix estimation based on high-frequency financial data. J. Econometrics 203, 69–79.
Kong, X.B., 2017. On the number of common factors with high frequency data. Biometrika 104, 397–410.
Kong, X.B., 2018. On the integrated systematic and idiosyncratic volatility with large panel high-frequency data. Ann. Statist. 46 (3), 1077–1108.
Kong, X.B., Liu, Z., Jing, B.Y., 2015. Testing for pure-jump processes for high-frequency data. Ann. Statist. 43, 847–877.
Kritchman, S., Nadler, B., 2008. Determining the number of components in a factor model from limited noisy data. Chemometr. Intell. Lab. Syst. 94, 19–32.
Li, J., Todorov, V., Tauchen, G., 2017. Jump regressions. Econometrica 85, 173–195.
Liu, C., Tang, C.Y., 2014. A quasi-maximum likelihood approach to covariance matrix with high frequency data. J. Econometrics 180, 217–232.
Onatski, A., 2009. Testing hypotheses about the number of factors in large factor models. Econometrica 77, 1447–1479.
Pelger, M., 2018. Large-dimensional factor modeling based on high-frequency observations. J. Econometrics 208, 23–42.
Stock, J., Watson, M., 2002. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97, 1167–1179.
Stock, J., Watson, M., 2005. Implications of dynamic factor models for VAR analysis. NBER Working Paper No. 11467.
Tao, M., Wang, Y., Zhou, H.H., 2013. Optimal sparse volatility matrix estimation for high dimensional Itô processes with measurement errors. Ann. Statist. 41, 1816–1864.
Wang, Y., Zou, J., 2010. Vast volatility matrix estimation for high-frequency financial data. Ann. Statist. 38, 943–978.
Xiu, D., 2010. Quasi-maximum likelihood estimation of volatility with high frequency data. J. Econometrics 159, 235–250.
Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. J. Amer. Statist. Assoc. 100, 1394–1411.