Journal of Econometrics 208 (2019) 23–42

Contents lists available at ScienceDirect

Journal of Econometrics

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

Large-dimensional factor modeling based on high-frequency observations[☆]

Markus Pelger [*]

Department of Management Science & Engineering, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Available online 11 October 2018

JEL classification:
C14
C38
C55
C58

Keywords:
Systematic risk
High-dimensional data
High-frequency data
Latent factor model
PCA
Jumps
Semimartingales
Approximate factor model
Number of factors

ABSTRACT

This paper develops a statistical theory to estimate an unknown factor structure based on financial high-frequency data. We derive an estimator for the number of factors and consistent and asymptotically mixed-normal estimators of the loadings and factors under the assumption of a large number of cross-sectional and high-frequency observations. The estimation approach can separate factors for continuous and rare jump risk. The estimators for the loadings and factors are based on the principal component analysis of the quadratic covariation matrix. The estimator for the number of factors uses a perturbed eigenvalue ratio statistic. In an empirical analysis of the S&P 500 firms we estimate four stable continuous systematic factors, which can be approximated very well by a market and industry portfolios. Jump factors are different from the continuous factors.

1. Introduction

Financial economists are now in the fortunate situation of having a huge amount of high-frequency financial data for a large number of assets. Over the past fifteen years the econometric methods to analyze the high-frequency data for a small number of assets have grown exponentially. At the same time the field of large dimensional data analysis has exploded providing us with a variety of tools to analyze a large cross-section of financial assets over a long time horizon. This paper merges these two literatures by developing statistical methods for estimating the systematic pattern in high frequency data for a large cross-section. One of the most popular methods for analyzing large cross-sectional data sets is factor analysis. Some of the most influential economic theories, e.g. the arbitrage pricing theory of Ross (1976) are based on factor models. While there is a well-developed inferential theory for factor models of large dimension with long time horizon and for factor models of small dimension based on high-frequency observations, the inferential theory for large dimensional high-frequency factor models is an area of active research.

[☆] We thank the editor and two anonymous referees for their useful comments. We thank Jason Zhu for excellent research assistance. We are very grateful for comments and helpful discussions with Robert M. Anderson, Martin Lettau, Michael Jansson, Lisa Goldberg, Viktor Todorov, Jianqing Fan, Per Mykland, Noureddine El Karoui, George Tauchen, Torben Andersen, Neil Shephard, Yacine Aït-Sahalia, Dacheng Xiu, Frank Diebold, Mikhail Chernov, Kay Giesecke, Svetlana Bryzgalova, Ken Singleton, Darrell Duffie, John Cochrane, Steve Evans, Olivier Scaillet and the seminar participants at UC Berkeley, Stanford and University of Pennsylvania and conference participants at SoFiE, Informs, FERM, Econometric society and NBER Time-Series meetings. This work was supported by the Center for Risk Management Research at UC Berkeley.

[*] Corresponding author.
E-mail address: mpelger@stanford.edu.

https://doi.org/10.1016/j.jeconom.2018.09.004

This paper develops the statistical inferential theory for approximate factor models of large dimensions based on high-frequency observations. Conventional factor analysis requires a long time horizon, while this methodology also works with short time horizons, e.g. a week. If a large cross-section of firms and sufficiently many high-frequency asset prices are available, we can derive consistent and asymptotically mixed-normal estimators of the latent loadings and factors. These results are obtained for very general stochastic processes, namely Itô semimartingales with jumps, and an approximate factor structure which allows for weak serial and cross-sectional correlation in the idiosyncratic errors. The estimation approach can separate factors for systematic large sudden movements, so-called jumps factors, from continuous factors. Our estimator for the loadings and factors is essentially the well-known principal component based estimator of Bai (2003), where we use properly rescaled increments for the covariance estimation. However, except for very special cases the necessary assumptions and the proofs cannot be mapped into the long-horizon factor model and hence require new derivations.

This paper develops a new diagnostic criterion for the number of factors that requires essentially only the same weak assumptions as the loadings estimator in our model. The basic idea in most estimation approaches is that the systematic eigenvalues of the estimated covariance matrix or quadratic covariation matrix will explode, while the other eigenvalues of the idiosyncratic part will be bounded. Prominent estimators with good performance in simulations[1] impose the additional strong assumptions of random matrix theory that imply that a certain fraction of the small eigenvalues will be bounded from below and above and the largest residual eigenvalues will cluster. We propose the novel idea of perturbing the eigenvalues before analyzing the eigenvalue ratio. As long as the eigenvalue ratio of the perturbed eigenvalues is close to one, the spectrum is due to the residuals. Due to a weaker rate argument and not the strong assumptions of random matrix theory the eigenvalue ratio of perturbed idiosyncratic eigenvalues will cluster. The important contribution of our estimator is that it can estimate the number of continuous, jump and total factors separately and that it can deal with systematic factors that produce only smaller eigenvalues in a finite sample.[2]

We develop an estimator for testing if a set of estimated statistical factors is close to a set of observable economic variables. One drawback of statistical factors is that they are usually not easy to interpret economically. An additional challenge is that factor models are only identified up to invertible transformations. We provide a measure for the distance between two sets of factors based on a total generalized correlation and develop its asymptotic distribution.

In an empirical analysis of the S&P 500 firms with 5 min high-frequency price data we estimate four stable continuous systematic factors, which can be approximated very well by a market and industry portfolios. We can show that the continuous factor structure is very stable in some years, but there is also time variation in the number and structure of factors over longer horizons. For the time period 2007–2012 we estimate four continuous factors which can be approximated very well by a market, oil, finance and electricity factor. From 2003 to 2006 one continuous systematic factor disappears. Systematic jump risk also seems to be different from systematic continuous risk. There seems to be only one stable jump factor, namely a market jump factor.

Our work builds on the fast growing literatures in the two separate fields of large-dimensional factor analysis and high-frequency econometrics. The notion of an "approximate factor model" was introduced by Chamberlain and Rothschild (1983), which allowed for a non-diagonal covariance matrix of the idiosyncratic component. The general case of a static large dimensional factor model is treated in Bai (2003). He develops an inferential theory for factor models for a large cross-section and long time horizons based on a principal component analysis of the sample covariance matrix. As pointed out before for general continuous-time processes we cannot map the high-frequency problem into the long horizon model. Fan et al. (2013) study an approximate factor structure with sparsity. Some of the most relevant estimators for the number of factors in large-dimensional factor models based on long-horizons are the Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013) estimators.[3] The last two estimators perform well in simulations, but their arguments which are based on random matrix theory do not seem to be transferable to our high-frequency problem without imposing unrealistically strong assumptions on the processes.[4] Many of our asymptotic results for the estimation of the quadratic covariation are based on Jacod (2008), where he develops the asymptotic properties of realized power variations and related functionals of semimartingales. Lee and Mykland (2008) and Mancini (2009) introduce a threshold estimator for separating the continuous from the jump variation, which we use in this paper. Bollerslev and Todorov (2010) develop the theoretical framework for high-frequency factor models for a low dimension. Their results are applied empirically in Bollerslev et al. (2016).

So far there are relatively few papers combing high-frequency analysis with high-dimensional regimes, but this is an active and growing literature. Important recent papers include Wang and Zhou (2010) and Tao et al. (2013a,b) who establish results for large sparse matrices estimated with high-frequency observations. Fan et al. (2014) estimate a large-dimensional covariance matrix with high-frequency data for a given factor structure. Aït-Sahalia and Xiu (2017a) develop the inferential theory of principal component analysis applied to a low-dimensional cross-section of high-frequency data. We work in a

[1] E.g. Onatski (2010) and Ahn and Horenstein (2013)

[2] In any finite sample the systematic eigenvalues might be not much larger than the residual noise spectrum as for example illustrated in our empirical analysis. As our diagnostic criterion depends only on the relationship between the eigenvalue of the weakest factor and the largest residual eigenvalues, it can detect systematic factors that are weak in a given finite sample.

[3] There are many alternative methods, e.g. Hallin and Liska (2007), Amengual and Watson (2007) or Kapetanios (2010), but in simulations they do not seem to outperform the above methods.

[4] The Bai and Ng (2002) paper uses an information criterion, while Onatski applies an eigenvalue difference estimator and Ahn and Horenstein an eigenvalue ratio approach. If the first systematic factors are stronger than other weak systematic factors the Ahn and Horenstein method can fail in simulations with realistic values, while the Onatski method can perform better as it focuses only on the residual eigenvalues.

large-dimensional setup which requires the additional structure of a factor model and derive the inferential theory for both the continuous and jump structures. Independently, Aït-Sahalia and Xiu (2017b) study a large-dimensional high-frequency factor model and derive consistent estimators for the factors based on continuous processes. Their paper concentrates on the matrix-wise asymptotic consistency properties of the covariance matrix and its inverse, while our paper focuses on the distribution theory and properties of the factors and loadings. Their main identification is based on a sparsity assumption on the continuous idiosyncratic covariance matrix, while our main identification condition is a bounded eigenvalue condition on the idiosyncratic covariance matrix allowing us to also consider jumps. We also provide an alternative diagnostic criterion for the number of factors and a measure to interpret factors economically.

The rest of the paper is organized as follows. Section 2 introduces the factor model and assumptions. In Section 3 we explain our estimators and present the theoretical results. We show the point-wise consistency of the factors and the loadings and separation into continuous and jump factors. The section also includes the asymptotic mixed-normal distribution of the loadings and a consistent estimator for the covariance matrix in the limiting distribution. In Section 4 we provide a new diagnostic criterion for the number of factors and a measure to compare statistical factors with economic candidate factors. Section 5 provides Monte-Carlo simulation evidence. Section 6 is an empirical application. Concluding remarks are provided in Section 7. All the proofs and additional results are deferred to the Supplementary Appendix.

2. Model setup

Assume the N-dimensional stochastic process $X(t)$ can be explained by a factor model

$$X_i(t) = \Lambda_i^\top F(t) + e_i(t) \qquad i = 1, \dots, N \text{ and } t \in [0, T],$$

where Λ_i is a $K \times 1$ dimensional vector and $F(t)$ is a K-dimensional stochastic process in continuous time. The loadings Λ_i describe the exposure to the systematic factors F, while the residuals e_i are stochastic processes that describe the idiosyncratic component. $X(t)$ will typically be the log-price process. However, we only observe the stochastic process $X(t)$ at discrete time observations $t_0 = 0, t_1 = \Delta_M, t_2 = 2\Delta_M, \dots, t_M = M\Delta_M$, in the interval $[0, T]$, where the time increment is defined as $\Delta_M = t_{j+1} - t_j = \frac{T}{M}$:

$$X_i(t_j) = \Lambda_i^\top F(t_j) + e_i(t_j) \qquad i = 1, \dots, N \text{ and } j = 0, \dots, M,$$

or in vector notation $X(t_j) = \Lambda F(t_j) + e(t_j)$ for $j = 0, \dots, M$ with $\Lambda = (\Lambda_1, \dots, \Lambda_N)^\top$. In our setup the number of cross-sectional observations N and the number of high-frequency observations M are large, while the time horizon T and the number of systematic factors K are fixed. The loadings Λ, factors F, residuals e and number of factors K are unknown and have to be estimated.

All the stochastic processes considered in this paper are locally bounded special Itô semimartingales as specified in Definition 1 in the Appendix:

$$Y(t) = Y(0) + \int_0^t b_s^Y ds + \int_0^t \sigma_s^Y dW_s^Y + \sum_{s \le t} \Delta Y(s),$$

where $\Delta Y(t) = Y(t) - Y(t-)$ denotes the jumps of the process Y. The process consists of a predictable drift term, a continuous martingale with N-dimensional Brownian motion W_t^Y and volatility process σ and a jump part. These particular semimartingales are standard in high-frequency econometrics, see e.g. Aït-Sahalia and Jacod (2014). The dynamics are very general and completely non-parametric. They allow for correlation between the volatility and asset price processes. We only impose some week regularity conditions in Definition 1.[5]

We observe M increments of the N-dimensional stochastic process $X(t)$ in the time interval $[0, T]$. For the time increments $\Delta_M = \frac{T}{M} = t_{j+1} - t_j$ we denote the increments of the stochastic processes by

$$X_{j,i} = X_i(t_{j+1}) - X_i(t_j) \qquad F_j = F(t_{j+1}) - F(t_j) \qquad e_{j,i} = e_i(t_{j+1}) - e_i(t_j).$$

In matrix notation we have

$$\underset{(M \times N)}{X} = \underset{(M \times K)(K \times N)}{F \quad \Lambda^\top} + \underset{(M \times N)}{e}.$$

The sum of squared increments converges to the quadratic covariation $\sum_{j=1}^M X_{j,i} X_{j,k} \xrightarrow{p} [X_i(t), X_k(t)]_T$ for $M \to \infty$ and for all $i, k = 1, \dots, N$. The predictable quadratic covariation $\langle X_i(t), X_k(t) \rangle_T$ is the predictable conditional expectation of $[X_i(t), X_k(t)]_T$, i.e. it is the so-called compensator process. It is the same as the realized quadratic covariation $[X_i(t), X_k(t)]$ for a continuous process, but differs if the processes have jumps. The realized quadratic covariation $[X_i(t), X_k(t)]_T$ and the conditional quadratic covariation $\langle X_i(t), X_k(t) \rangle_T$ are themselves stochastic processes. In order to simplify notation we leave out the time variable t and the terminal time index T for the quadratic covariation if there is no ambiguity.

[5] The model includes many well-known continuous-time models as special cases: for example stochastic volatility models like the CIR or Heston model, the affine class of models in Duffie et al. (2000) and Barndorff-Nielsen and Shephard's (2002) Ornstein–Uhlenbeck stochastic volatility model with jumps or Andersen et al.'s (2002) stochastic volatility model with log-normal jumps generated by a non-homogeneous Poisson process.

Our estimation theory is derived under the assumption of synchronous data with negligible microstructure noise.[6] Using for example 5-minute sampling frequency as commonly advocated in the literature on realized volatility estimation, e.g. Andersen et al. (2001) and the survey by Hansen and Lunde (2006), seems to justify this assumption and still provides enough high-frequency observations to apply our estimator to a monthly horizon.[7]

The key assumption for obtaining a consistent estimator for the loadings and factors is an approximate factor structure. It requires that the factors are systematic in the sense that they cannot be diversified away, while the idiosyncratic residuals are nonsystematic and can be diversified away. The approximate factor structure assumption uses the idea of appropriately bounded eigenvalues of the residual quadratic covariation matrix, which is analogous to Chamberlain and Rothschild (1983) and Chamberlain (1988). Let $\|A\| = (tr(A^\top A))^{1/2}$ denote the norm of a matrix A and $\lambda_i(A)$ the i's largest singular value of the matrix A, i.e. the square-root of the i's largest eigenvalue of $A^\top A$. If A is a symmetric matrix then λ_i is simply the i's largest eigenvalue of A.

Assumption 1 (*Factor Structure Assumptions*)**.**

1. **Locally bounded special Itô semimartingales**
   The K-dimensional common factor F and the N-dimensional residual process e are uniformly locally bounded special Itô semimartingales specified in Definition 1 in the Appendix. In addition each e_i is a square integrable martingale.
2. **Factors and factor loadings**
   The quadratic covariation matrix of the factors Σ_F is positive definite a.s.
$$\sum_{j=1}^{M} F_j F_j^\top \xrightarrow{p} [F, F]_T =: \Sigma_F \qquad \text{and} \qquad \left\| \frac{\Lambda^\top \Lambda}{N} - \Sigma_\Lambda \right\| \to 0,$$
   where Σ_Λ is also positive definite. The loadings are bounded: $\|\Lambda_i\| < \infty$ for all $i = 1, \dots, N$.
3. **Independence of F and e**
   The factor process F and the residual processes e are independent.
4. **Approximate factor structure**
   The largest eigenvalue of the residual quadratic covariation matrix is bounded in probability, i.e. $\lambda_1([e, e]_T) = O_p(1)$. As the predictable quadratic covariation is absolutely continuous, we can define the instantaneous predictable quadratic covariation as $\frac{d\langle e_i, e_k\rangle_t}{dt} =: G_{i,k}(t)$. We assume that the largest eigenvalue of the matrix $G(t)$ is almost surely bounded for all t: $\lambda_1(G(t)) < C$ a.s. for all t for some constant C.
5. **Identification condition** All eigenvalues of $\Sigma_\Lambda \Sigma_F$ are distinct a.s.

The most important part of Assumption 1 is the approximate factor structure in point 4. It implies that the residual risk can be diversified away. Point 1 states that we can use the very general class of uniformly locally bounded special semimartingales. The existence of uniform bounds on all processes is necessary for obtaining the asymptotic results in a large dimensional setup. This is a standard assumption also used in Aït-Sahalia and Xiu (2017a,b) and Fan et al. (2014). The assumption that the residuals are martingales and hence do not have a drift term is only necessary for the asymptotic distribution results. The consistency results do not require this assumption. Point 2 implies that the factors affect an infinite number of assets and hence cannot be diversified away. Point 3 can be relaxed to allow for a weak correlation between the factors and residuals. This assumption is only used to derive the asymptotic distribution of the estimators. The approximate factor structure assumption in point 4 puts a restriction on the correlation of the residual terms. It allows for cross-sectional (and also serial) correlation in the residual terms as long as it is not too strong.[8]

Note that point 4 puts restrictions on both the realized and the conditional quadratic covariation matrix. In the case of continuous residual processes, the conditions on the conditional quadratic covariation matrix are obviously sufficient. However, in our more general setup it is not sufficient to restrict only the conditional quadratic covariation matrix.

The estimation of the factors requires a stronger assumption on the cross-sectional dependence in the residuals:

[6] In Appendix I we extend the model to include microstructure noise and show how the noise affects the largest eigenvalue of the residual matrix. This result can be used to verify if the estimated number of factors changes in the presence of microstructure noise.

[7] Inference on the volatility of a continuous semimartingale under noise contamination can be pursued using smoothing techniques. Several approaches have been developed, prominent ones by Aït-Sahalia et al. (2005b), Barndorff-Nielsen et al. (2008), Zhang et al. (2005), Xiu (2010) and Jacod et al. (2009) in the one-dimensional setting and generalizations for a noisy non-synchronous multi-dimensional setting by Aït-Sahalia et al. (2010), Podolskij and Vetter (2009), Barndorff-Nielsen et al. (2011), Zhang (2011) and Bibinger and Winkelmann (2014) among others. However, neither the microstructure robust estimators nor the non-synchronicity robust estimators can be easily extended to our large dimensional problem. It is beyond the scope of this paper to develop the asymptotic theory for these more general estimators in the context of a large dimensional factor model and we leave this to future research.

[8] We can relax the approximate factor structure assumption. Instead of almost sure boundedness of the predictable instantaneous quadratic covariation matrix of the residuals it is sufficient to assume that

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{k \neq i}^{N} \Lambda_i G_{i,k}(t) \Lambda_k^\top < C \qquad \text{a.s. for all } t.$$

Then, all main results except for Theorems 4 and 6 continue to hold. Under this weaker assumption we do not assume that the diagonal elements of G are almost surely bounded. By Definition 1 the diagonal elements of G are already locally bounded which is sufficient for most of our results.

Assumption 2 (*Weak Dependence of Error Terms*). The row sum of the quadratic covariation of the residuals is bounded in probability:

$$\sum_{i=1}^{N} \|[e_k, e_i]_T\| = O_p(1) \qquad \forall k = 1, \dots, N \qquad \text{for } N \to \infty.$$

Assumption 2 is stronger than $\lambda_1([e, e]_T) = O_p(1)$ in Assumption 1. As the largest eigenvector of a matrix can be bounded by the largest absolute row sum, Assumption 2 implies $\lambda_1([e, e]_T) = O_p(1)$. If the residuals are cross-sectionally independent it is trivially satisfied. However it allows for a weak correlation between the residual processes. For example, if the residual part of each asset is only correlated with a finite number of residuals of other assets, it will be satisfied.[9]

We are also interested in estimating the continuous component, jump component and the volatility of the factors. We can separate the factors into continuous factors that have only a continuous martingale and predictable finite variation part and into jump factors consisting of a jump martingale and predicable finite variation term but no continuous martingale. It is important to include the drift terms in this definition as they correspond to the risk-premium of the continuous respectively jump factors if the arbitrage pricing theory holds.[10] Consider for example a market factor. Bollerslev et al. (2016) have shown that stocks have different loadings with respect to the continuous and the jump movements of the market. They infer that the risk premium with respect to continuous market risk is different from jump market risk. This implies that the drift term in the continuous market factor is different from the drift term in the jump market factor. Without loss of generality we can formulate the model as

$$X(t) = {\Lambda^{C}}^{\top} F^C(t) + {\Lambda^{D}}^{\top} F^D(t) + e(t).$$

F^C denotes the continuous martingales with drift and F^D the jump martingales with corresponding drift. This framework also allows for factors with continuous and jump components that have identical continuous and jump loadings. The number of continuous and jump factors is K^C respectively K^D. In the example of a market factor with loadings that are different for the continuous and jump components, we use the convention $K^C = 1$, $K^D = 1$ and $K = 2$ as the total quadratic covariation matrix has two exploding eigenvalues.

3. Main results

3.1. Estimators

For a given number of factors K our goal is to estimate Λ and F. As in any factor model where only the $M \times N$ matrix X is observed, Λ and F are only identified up to K^2 parameters. Hence, we impose the standard identification assumptions that $\frac{\hat{\Lambda}^\top \hat{\Lambda}}{N} = I_K$ and $\hat{F}^\top \hat{F}$ is a diagonal matrix.[11]

Denote the K largest eigenvalues of $\frac{1}{N} X^\top X$ by V_{MN}. The estimator for the loadings $\hat{\Lambda}$ is defined as the K eigenvectors of V_{MN} multiplied by $\sqrt{N}$. The estimator for the factor increments is $\hat{F} = \frac{1}{N} X \hat{\Lambda}$. Note that $\frac{1}{N} X^\top X$ is an estimator for $\frac{1}{N}[X, X]$ for a finite N. The estimator is essentially principal component analysis applied to the estimated quadratic covariation matrix. The systematic component of $X(t)$ is the part that is explained by the factors and defined as $C(t) = \Lambda F(t)$. The increments of the systematic component $C_{j,i} = F_j \Lambda_i^\top$ are estimated by $\hat{C}_{j,i} = \hat{F}_j \hat{\Lambda}_i^\top$.

3.2. Consistency

As pointed out before, the factors F and loadings Λ are not separately identifiable. However, we can estimate them up to an invertible $K \times K$ matrix H. Hence, our estimator $\hat{\Lambda}$ will estimate ΛH and $\hat{F}$ will estimate $F {H^{\top}}^{-1}$. Note that the common component is well-identified and $F \Lambda^\top = F {H^{\top}}^{-1} H^\top \Lambda^\top$.[12]

In our general approximate factor models we require N and M to go to infinity. The rates of convergence will usually depend on the smaller of these two values denoted by $\delta = \min(N, M)$. As noted before we consider a simultaneous limit for N and M and not a path-wise or sequential limit. Without further assumptions the asymptotic results do not hold for a fixed N or M. In this sense the large dimension of our problem, which makes the analysis more complicated, also helps us to obtain more general results and turns the "curse of dimensionality" into a "blessing".

Note that F_j is the increment $F(t_{j+1}) - F(t_j)$ and goes to zero for $M \to \infty$ for almost all increments. It can be shown that in a specific sense we can also consistently estimate the factor increments, but the asymptotic statements will be formulated in terms of the stochastic process F evaluated at a discrete time point t_j. For example $F(T) = \sum_{j=1}^{M} F_j$ denotes the factor

[9] Assumption 2 is similar to the sparsity assumption imposed in Aït-Sahalia and Xiu (2017b). They allow the row sum to grow at a slow rate.

[10] Note that the predictable finite variation part measuring the risk-premium is well-defined if we assume an asset pricing model (for example Chamberlain (1988) and Back (1991).

[11] Λ and F are only identified up to K^2 parameters as $F \Lambda^\top = F A A^{-1} \Lambda^\top$ for any arbitrary invertible $K \times K$ matrix A. Hence, for our estimator we impose the K^2 standard restrictions that $\frac{\hat{\Lambda}^\top \hat{\Lambda}}{N} = I_K$ which gives us $\frac{K(K+1)}{2}$ restrictions and that $\hat{F}^\top \hat{F}$ is a diagonal matrix, which yields another $\frac{K(K-1)}{2}$ restrictions.

[12] For a more detailed discussion see Bai (2003).

process evaluated at time T. Similarly we can evaluate the process at any other discrete time point $T_m = m \cdot \Delta_M$ as long as $m \cdot \Delta_M$ does not go to zero. Essentially m has to be proportional to M. For example, we could chose T_m equal to $\frac{1}{2}T$ or $\frac{1}{4}T$. The terminal time T can always be replaced by the time T_m in all the theorems. The same holds for the common component.[13]

Theorem 1 (*Consistency of Estimators*)**.** *Define the rate $\delta = \min(N, M)$ and the invertible matrix $H = \frac{1}{N}\left(F^\top F\right)\left(\Lambda^\top \hat{\Lambda}\right) V_{MN}^{-1}$. Then the following consistency results hold:*

1. *Consistency of loadings estimator: Under Assumption 1 it follows that*
$$\hat{\Lambda}_i - H^\top \Lambda_i = O_p\left(\frac{1}{\sqrt{\delta}}\right).$$
2. *Consistency of factor estimator and common component: Under Assumptions 1 and 2 it follows that*
$$\hat{F}(T) - H^{-1}F(T) = O_p\left(\frac{1}{\sqrt{\delta}}\right), \qquad \hat{C}_i(T) - C_i(T) = O_p\left(\frac{1}{\sqrt{\delta}}\right).$$
3. *Consistency of quadratic variation: Under Assumptions 1 and 2 and for any stochastic process $Y(t)$ satisfying Definition 1 we have for $\frac{\sqrt{M}}{N} \to 0$ and $\delta \to \infty$:*
$$\sum_{j=1}^{M} \hat{F}_j \hat{F}_j^\top = H^{-1}[F, F]_T H^{-1^\top} + o_p(1), \qquad \sum_{j=1}^{M} \hat{F}_j Y_j = H^{-1}[F, Y]_T + o_p(1)$$
$$\sum_{j=1}^{M} \hat{e}_{j,i}\hat{e}_{j,k} = [e_i, e_k]_T + o_p(1), \qquad \sum_{j=1}^{M} \hat{e}_{j,i} Y_j = [e_i, Y]_T + o_p(1)$$
$$\sum_{j=1}^{M} \hat{C}_{j,i}\hat{C}_{j,k} = [C_i, C_k]_T + o_p(1), \qquad \sum_{j=1}^{M} \hat{C}_{j,i} Y_j = [C_i, Y]_T + o_p(1),$$
for $i, k = 1, \ldots, N$.

3.3. Separating continuous and jump factors

Using a thresholding approach we can separate the continuous and jump movements in the observable process X and estimate the systematic continuous and jump factors. The idea is that with sufficiently many high-frequency observations, we can identify the jumps in $X(t)$ as the movements that are above a certain threshold. This allows us to separate the quadratic covariation matrix of X into its continuous and jump components. Then applying principal component analysis to each of these two matrices we obtain our separate factors. A crucial assumption is that the thresholding approach can actually identify the jumps:

Assumption 3 (*Truncation Identification*)**.** F and e_i have only finite activity jumps. The quadratic covariation matrix of the continuous factors $[F^C, F^C]_T$ and of the jump factors $[F^D, F^D]_T$ are each positive definite a.s. and the matrices $\frac{\Lambda^{C^\top}\Lambda^C}{N}$ and $\frac{\Lambda^{D^\top}\Lambda^D}{N}$ each converge in probability to positive definite matrices.

Assumption 3 has two important parts. First, we require the processes to have only finite jump activity. This means that on every finite time interval there are almost surely only finitely many jumps. The second statement requires each systematic jump factor to jump at least once in the data. This is a straightforward and necessary condition to identify any jump factor. Hence, the main restriction in Assumption 3 is the finite jump activity. For example compound Poisson processes with stochastic intensity rate fall into this category.

Theorem 2 (*Separating Continuous and Jump Factors*)**.** *Assume Assumptions 1 and 3 hold. Set the threshold identifier for jumps as $\alpha\Delta_M^{\bar{\omega}}$ for some $\alpha > 0$ and $\bar{\omega} \in \left(0, \frac{1}{2}\right)$ and define $\hat{X}^C_{j,i} = X_{j,i}\mathbb{1}_{\{|X_{j,i}| \le \alpha\Delta_M^{\bar{\omega}}\}}$ and $\hat{X}^D_{j,i} = X_{j,i}\mathbb{1}_{\{|X_{j,i}| > \alpha\Delta_M^{\bar{\omega}}\}}$.[14] The estimators $\hat{\Lambda}^C$, $\hat{\Lambda}^D$, $\hat{F}^C$ and $\hat{F}^D$ are defined analogously to $\hat{\Lambda}$ and $\hat{F}$, but using $\hat{X}^C$ and $\hat{X}^D$ instead of X. Define $H^C = \frac{1}{N}\left(F^{C^\top}F^C\right)\left(\Lambda^{C^\top}\hat{\Lambda}^C\right)V_{MN}^{C^{-1}}$ and $H^D = \frac{1}{N}\left(F^{D^\top}F^D\right)\left(\Lambda^{D^\top}\hat{\Lambda}^D\right)V_{MN}^{D^{-1}}$.*

[13] This statement only provides a pointwise convergence of processes evaluated at specific times. A stronger statement would be to show weak convergence for the stochastic processes. However, weak convergence of stochastic processes requires significantly stronger assumptions and will in general not be satisfied under our assumptions.

[14] The thresholding approach has first been proposed by Mancini (2009) and Lee and Mykland (2008). Choices of α and $\bar{\omega}$ are standard in the literature (see, e.g. Bollerslev and Todorov (2010)) and are discussed below when implemented in simulations.

1. *The continuous and jump loadings can be estimated consistently:*

$$\hat{\Lambda}_i^C = H^{C^\top}\Lambda_i^C + o_p(1), \qquad \hat{\Lambda}_i^D = H^{D^\top}\Lambda_i^D + o_p(1).$$

2. *Assume that additionally Assumption 2 holds. The continuous and jump factors can only be estimated up to a finite variation bias term*

$$\hat{F}(T)^C = H^{C^{-1}}F(T)^C + o_p(1) + \textit{finite variation term},$$
$$\hat{F}(T)^D = H^{D^{-1}}F(T)^D + o_p(1) + \textit{finite variation term}.$$

3. *Under the additional Assumption 2 we can estimate consistently the covariation of the continuous and jump factors with other processes. Let $Y(t)$ be an Itô-semimartingale satisfying Definition 1. Then we have for $\frac{\sqrt{M}}{N} \to 0$ and $\delta \to \infty$:*

$$\sum_{j=1}^{M} \hat{F}_j^C Y_j = H^{C^{-1}}[F^C, Y]_T + o_p(1), \qquad \sum_{j=1}^{M} \hat{F}_j^D Y_j = H^{D^{-1}}[F^D, Y]_T + o_p(1).$$

The theorem states that we can estimate the factors only up to a finite variation term, i.e. we can only estimate the martingale part of the process correctly. The intuition behind this problem is simple. The truncation estimator can correctly separate the jumps from the continuous martingale part. However, all the drift terms will be assigned to the continuous component. If a jump factor also has a drift term, this will now appear in the continuous part and as this drift term affects infinitely many cross-sectional $X_i(t)$, it cannot be diversified away. This result is important as it shows that using a model with a purely continuous price process based on the argument that jumps have been removed by a thresholding approach can lead to wrong inferential asymptotics.

3.4. Asymptotic distribution

The estimator for the loadings converges stably in law to a mixed Gaussian limit.[15]

Theorem 3 (*Asymptotic Distribution of Loadings*). *Assume Assumptions 1 and 2 hold and define $\delta = \min(N, M)$. Then*

$$\sqrt{M}\left(\hat{\Lambda}_i - H^\top \Lambda_i\right) = V_{MN}^{-1}\left(\frac{\hat{\Lambda}^\top \Lambda}{N}\right)\sqrt{M}F^\top e_i + O_p\left(\frac{\sqrt{M}}{\delta}\right).$$

If $\frac{\sqrt{M}}{N} \to 0$, then

$$\sqrt{M}(\hat{\Lambda}_i - H^\top \Lambda_i) \xrightarrow{L-s} N\left(0, V^{-1}Q\Gamma_i Q^\top V^{-1}\right),$$

where V is the diagonal matrix of eigenvalues of $\Sigma_\Lambda^{\frac{1}{2}}\Sigma_F\Sigma_\Lambda^{\frac{1}{2}}$ and $\operatorname*{plim}_{N,M\to\infty}\frac{\hat{\Lambda}^\top\Lambda}{N} = Q = V^{\frac{1}{2}}\Upsilon^\top\Sigma_F^{\frac{1}{2}}$ with Υ being the eigenvectors of V. The entry $\{l, g\}$ of the $K \times K$ matrix Γ_i is given by

$$\Gamma_{i,l,g} = \int_0^T \sigma_{F^l,F^g}\sigma_{e_i}^2 ds + \sum_{s\le T}\Delta F^l(s)\Delta F^g(s)\sigma_{e_i}^2(s) + \sum_{s'\le T}\Delta e_i^2(s')\sigma_{F^g,F^l}(s').$$

F^l denotes the lth component of the K dimensional process F and σ_{F^l,F^g} are the entries of its $K \times K$ dimensional volatility matrix.

The asymptotic expansion is very similar to the conventional factor analysis in Bai (2003), but the limiting distributions of the loadings are obviously different. The mode of convergence is stable convergence in law, which is stronger than simple convergence in distribution.[16] The asymptotic variance is random combining the volatility and jump processes of the factors and idiosyncratic component.[17] Here we can see very clearly how the results from high-frequency econometrics impact the estimators in our factor model.

The asymptotic covariance matrix for the estimator of the loadings can be estimated consistently under the same weak assumptions and is obviously very different from the long-horizon framework:

[15] In order to obtain a mixed Gaussian limit distribution for the loadings we need to assume that there are no common jumps in σ_F and e_i and in σ_{e_i} and F. Without this assumption the estimator for the loadings still converges at the same rate, but it is not mixed-normally distributed any more. Note that Assumption 1 requires the independence of F and e, which implies the no common jump assumption.

[16] For more details see Aït-Sahalia and Jacod (2014).

[17] The long-horizon model of Bai (2003) has an asymptotic variance of the form $\Gamma_i = plim_{T\to\infty}\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}E\left[F_tF_se_{s,i}e_{t,i}\right]$ for stationary processes F and e. The Online Appendix H has an extensive discussion about the differences between these two frameworks.

Theorem 4 (*Feasible Estimator of Covariance Matrix of Loadings*). *Assume Assumptions 1 and 2 hold and $\frac{\sqrt{M}}{N} \to 0$. Define the asymptotic covariance matrix of the loadings as $\Theta_{\Lambda,i} = V^{-1}Q\Gamma_i Q^\top V^{-1}$. Take any sequence of integers $k \to \infty$, $\frac{k}{M} \to 0$. Denote by $I(j)$ a local window of length $\frac{2k}{M}$ around j. Define the $K \times K$ matrix $\hat{\Gamma}_i$ by*

$$\begin{aligned}\hat{\Gamma}_i =& M\sum_{j=1}^{M}\left(\frac{\hat{X}_j^C\hat{\Lambda}}{N}\right)\left(\frac{\hat{X}_j^C\hat{\Lambda}}{N}\right)^\top\left(\hat{X}_{j,i}^C - \frac{\hat{X}_j^C\hat{\Lambda}}{N}\hat{\Lambda}_i\right)^2 \\ &+ \frac{M}{2k}\sum_{j=k+1}^{M-k}\left(\frac{\hat{X}_j^D\hat{\Lambda}}{N}\right)\left(\frac{\hat{X}_j^D\hat{\Lambda}}{N}\right)^\top\left(\sum_{h\in I(j)}\left(\hat{X}_{h,i}^C - \frac{\hat{X}_h^C\hat{\Lambda}}{N}\hat{\Lambda}_i\right)^2\right) \\ &+ \frac{M}{2k}\sum_{j=k+1}^{M-k}\left(\hat{X}_{j,i}^D - \frac{\hat{X}_j^D\hat{\Lambda}}{N}\hat{\Lambda}_i\right)^2\left(\sum_{h\in I(j)}\left(\frac{\hat{X}_h^C\hat{\Lambda}}{N}\right)\left(\frac{\hat{X}_h^C\hat{\Lambda}}{N}\right)^\top\right).\end{aligned}$$

Then a feasible estimator for $\Theta_{\Lambda,i}$ is $\hat{\Theta}_{\Lambda,i} = V_{MN}^{-1}\hat{\Gamma}_i V_{MN}^{-1} \xrightarrow{p} \Theta_{\Lambda,i}$ and

$$\sqrt{M}\hat{\Theta}_{\Lambda,i}^{-1/2}(\hat{\Lambda}_i - H^\top\Lambda_i) \xrightarrow{D} N(0, I_K).$$

Under the same assumptions we can derive an asymptotic expansion for the estimator of the factors. The asymptotic mixed-normality of the factors need the substantially stronger assumptions which are collected in Appendix E.

Theorem 5 (*Asymptotic Distribution of the Factors*). *Assume Assumptions 1 and 2 hold. Then*

$$\sqrt{N}\left(\hat{F}(T) - H^{-1}F(T)\right) = \frac{1}{\sqrt{N}}e_T\Lambda H + O_P\left(\frac{\sqrt{N}}{\sqrt{M}}\right) + O_p\left(\frac{\sqrt{N}}{\delta}\right).$$

If Assumptions 4 and 5 hold and $\frac{\sqrt{N}}{M} \to 0$ or only Assumption 4 holds and $\frac{N}{M} \to 0$:

$$\sqrt{N}\left(\hat{F}(T) - H^{-1}F(T)\right) \xrightarrow{L-s} N\left(0, Q^{-1^\top}\Phi_T Q^{-1}\right)$$

with $\Phi_T = \underset{N\to\infty}{plim}\frac{\Lambda^\top[e,e]_T\Lambda}{N}$.

The asymptotic distribution is driven by a cross-sectional average of the martingale processes $e(T)$. It should not come as a surprise that the central limit theorems impose restrictions on the tail behavior of the idiosyncratic processes. Note that we study the asymptotic distribution of the factor processes evaluated at some terminal time, i.e. the cumulative sum of increments, which is different from the conventional long-horizon models and another reason why stronger assumptions are necessary.

The central limit theorem for the common components in Appendix E combines the asymptotic distribution of the loading and factor estimates and hence requires similar assumptions as for the factors. Depending on the asymptotic relationship between N and M either the loading or factor distribution dominates.

4. Further discussions

4.1. A diagnostic criterion for the number of factors

We propose a consistent estimator for the number of total, continuous and jump factors. Intuitively the large eigenvalues are associated with the systematic factors and hence the problem of estimating the number of factors is roughly equivalent to deciding which eigenvalues are considered to be large with respect to the rest of the spectrum. Our arguments are based on the result that the first K "systematic" eigenvalues of $X^\top X$ are $O_p(N)$, while the nonsystematic eigenvalues are $O_p(1)$. A straightforward estimator for the number of factors considers the eigenvalue ratio of two successive eigenvalues and associates the number of factors with a large eigenvalue ratio. However, without very strong assumptions we cannot bound the small eigenvalues from below, which could lead to exploding eigenvalue ratios in the nonsystematic spectrum. We propose a perturbation method to avoid this problem.[18] As long as the eigenvalue ratios of the perturbed eigenvalues cluster, we are in the nonsystematic spectrum. We are in the systematic spectrum when the clustering stops and the perturbed eigenvalue ratio is large.

[18] A different approach of using perturbed matrices for rank testing has been proposed in Jacod and Podolskij (2013).

Theorem 6 (*Estimator for Number of Factors*). *Assume Assumptions 1 and 3 hold,* $\frac{\log(N)}{M} \rightarrow 0$ *and idiosyncratic jumps are independent of the continuous part in the idiosyncratic process* $e(t)$. *Denote the ordered eigenvalues of* $X^\top X$ *by* $\lambda_1 \geq \cdots \geq \lambda_N$. *Choose a slowly increasing sequence* $g(N, M)$ *such that* $\frac{g(N,M)}{N} \rightarrow 0$ *and* $g(N, M) \rightarrow \infty$. *Define perturbed eigenvalues*

$$\hat{\lambda}_k = \lambda_k + g(N, M)$$

and the perturbed eigenvalue ratio statistics:

$$ER_k = \frac{\hat{\lambda}_k}{\hat{\lambda}_{k+1}} \qquad \text{for } k = 1, \ldots, N-1.$$

Define $\hat{K}(\gamma) = \max\{k \leq N-1 : ER_k > 1+\gamma\}$ *for* $\gamma > 0$. *If* $ER_k < 1+\gamma$ *for all k, then set* $\hat{K}(\gamma) = 0$. *Then we have for any* $\gamma > 0$ *that* $\hat{K}(\gamma) \xrightarrow{p} K$. *Denote the ordered eigenvalues of* $\hat{X}^{C\top}\hat{X}^C$ *by* $\lambda_1^C \geq \cdots \geq \lambda_N^C$ *and analogously for* $\hat{X}^{D\top}\hat{X}^D$ *by* $\lambda_1^D \geq \cdots \lambda_N^D$. *Define* $\hat{K}^C(\gamma)$ *and* $\hat{K}^D(\gamma)$ *as above but using* λ_i^C *respectively* λ_i^D. *Then we have for any* $\gamma > 0$ *that* $\hat{K}^C(\gamma) \xrightarrow{p} K^C$ *and* $\hat{K}^D(\gamma) \xrightarrow{p} K^D$, *where* K^C *is the number of continuous factors and* K^D *is the number of jump factors.*

Some of the most relevant estimators for the number of factors in large-dimensional factor models based on long-horizons are the Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013) estimators. The Bai and Ng (2002) paper uses an information criterion, while Onatski applies an eigenvalue difference estimator and Ahn and Horenstein an eigenvalue ratio approach. In simulations the last two estimators seem to perform well.[19] Our estimator combines elements of the Ahn and Horenstein estimator as we analyze eigenvalue ratios and elements of the Onatski estimator as we use a clustering argument. In contrast to these two approaches our results are not based on random matrix theory. Under the strong assumptions of random matrix theory a certain fraction of the small eigenvalues will be bounded from below and above[20] and the largest residual eigenvalues will cluster. Onatksi analyses the difference in eigenvalues. As long as the eigenvalue difference is small, it is likely to be part of the residual spectrum because of the clustering effect. The first time the eigenvalue difference is above a threshold, it indicates the beginning of the systematic spectrum. The Ahn and Horenstein method looks for the maximum in the eigenvalue ratios. As the smallest systematic eigenvalue is unbounded, while up to a certain index the nonsystematic eigenvalues are bounded from above and below, consistency follows. However, if the first systematic factor is more dominant than the other systematic factors the Ahn and Horenstein method can fail to detect the less dominant factors in a finite sample. In this sense the clustering argument of Onatksi is more appealing as it focusses on the residual spectrum and tries to identify when the spectrum is unlikely to be due to residual terms.[21] For the same reason our perturbed eigenvalue ratio estimator performs well in simulations with dominant and weaker factors.

The need for developing our estimator was motivated by the empirical analysis of the 5 min returns of the 500 companies in the S&P 500 from 2003–2012 in Section 6. The Onatski approach predicts 3 to 4 factors for the different time periods. These first four factors are stable over time and have an economically meaningful interpretation. Unfortunately, the Onatski estimator applied to high-frequency data requires very strong assumption which is not satisfied by the data and could not separate the continuous and jump factors.[22] Our estimator provides economically meaningful results similar to the Onatski approach, but under much weaker and realistic assumptions. The Ahn and Horenstein method would consistently predict only a single dominant factor, while the Bai and Ng (2002) methodology provides very unstable results that erratically fluctuate between 1 and 16 factors.

Our estimator depends on two choice variables: the perturbation g and the cutoff γ. In contrast to Bai and Ng, Onatski or Ahn and Horenstein we do not need to choose some upper bound on the number of factors. Although consistency follows for any g or γ satisfying the necessary conditions, the finite sample properties will obviously depend on them. As a first step for understanding the factor structure we recommend plotting the perturbed eigenvalue ratio statistic. In all our simulations the transition from the idiosyncratic spectrum to the systematic spectrum is very apparent. Based on simulations a good choice for the perturbation is $g = \sqrt{N} \cdot median(\{\lambda_1, \ldots, \lambda_N\})$. In the simulations we also test different specifications for g, e.g. $\log(N) \cdot median(\{\lambda_1, \ldots, \lambda_N\})$. Our estimator is very robust to the choice of the perturbation value. A more delicate issue is the cutoff γ. Simulations suggest that γ between 0.05 and 0.2 performs very well. If we apply our estimator without a perturbation, it will still be consistent, but requires stronger assumptions as Proposition 4 in Appendix F shows.

4.2. Identifying the factors

This section develops an estimator for testing if a set of estimated statistical factors is close to a set of observable economic variables. One drawback of statistical factors is that they are usually not easy to interpret economically. In the case of only

[19] See for example the numerical simulations in Onatski (2010) and Ahn and Horenstein (2013).

[20] See Ahn and Horenstein (2013) Lemma A.9.

[21] Onatksi's estimator requires the distance between the idiosyncratic eigenvalues to converge to zero, while our perturbed eigenvalue ratio would also work if there is a finite gap between the idiosyncratic eigenvalues.

[22] Zheng and Li (2011) prove a Marchenko–Pastur type equation for the estimated integrated covariance matrix of high-dimensional diffusion processes. They require strong structural assumptions on the volatility process and exclude jumps.

one factor, one could measure correlations with other factors or more generally regress this factor on a set of candidate factors and report a R^2 measure. Our estimator generalizes this idea to a multivariate setup.

As we have already noted before, factor models are only identified up to invertible transformations. Two sets of factors represent the same factor model if the factors span the same vector space. When trying to interpret estimated factors by comparing them with economic factors, we need a measure to describe how close two vector spaces are to each other. As proposed by Bai and Ng (2006) the generalized correlation is a natural candidate measure.[23] Let F be our K-dimensional set of factor processes and G be a K_G-dimensional set of economic candidate factor processes. We want to test if a linear combination of the candidate factors G can approximate the true factors F. The first generalized correlation is the highest correlation that can be achieved through a linear combination of the factors F and the candidate factors G. For the second generalized correlation we first project out the subspace that spans the linear combination for the first generalized correlation and then determine the highest possible correlation that can be achieved through linear combinations of the remaining $K-1$ respectively K_G-1 dimensional subspaces. This procedure continues until we have calculated the $min(K, K_G)$ generalized correlation. Mathematically the generalized correlations are the square root of the $min(K, K_G)$ largest eigenvalues of the matrix $[F, G]^{-1}[F, F][G, G]^{-1}[G, F]$. If $K = K_G = 1$ it is simply the correlation as measured by the quadratic covariation. If for example for $K = K_G = 3$ the generalized correlations are $\{1, 0.5, 0.5\}$ it implies that there exists a linear combination of the three factors in G that can replicate one of the three factors in F, while the other two uncorrelated factors in G each have a 50% correlation with the remaining two true factors.[24] We show that under general conditions the estimated factors $\hat{F}$, $\hat{F}^C$ and $\hat{F}^D$ can be used instead of the true unobserved factors.

Unfortunately, in this high-frequency setting there does not seem to exist a theory for confidence intervals for the individual generalized correlations.[25] However, we have developed an asymptotic distribution theory for the sum of squared generalized correlations, which we label as total generalized correlation. With the total generalized correlation we can measure how close a set of economic factors is to a set of statistical factors.

The total generalized correlation denoted by $\bar{\rho}$ is defined as the sum of the squared generalized correlations $\bar{\rho} = \sum_{k=1}^{\min(K_F,K_G)} \rho_k^2$. It is equal to

$$\bar{\rho} = trace\left([F, F]^{-1}[F, G][G, G]^{-1}[G, F]\right).$$

The estimator for the total generalized correlation is defined as

$$\hat{\bar{\rho}} = trace\left((\hat{F}^\top \hat{F})^{-1}(\hat{F}^\top G)(G^\top G)^{-1}(G^\top \hat{F})\right).$$

As the trace operator is a differentiable function and the quadratic covariation estimator is asymptotically mixed-normally distributed we can apply a delta method argument to show that $\sqrt{M}(\hat{\bar{\rho}} - \bar{\rho})$ is asymptotically mixed-normally distributed as well. The higher the generalized correlation, the closer are the candidate factors to the statistical factors.

Theorem 7 (*Asymptotic Distribution for Total Generalized Correlation*)**.** *Assume $F(t)$ is a factor process as in Assumption 1. Denote by $G(t)$ a K_G-dimensional process satisfying Definition 1. The process G is either (i) a well-diversified portfolio of X, i.e. it can be written as $G(t) = \frac{1}{N}\sum_{i=1}^N w_i X_i(t)$ with $\|w_i\|$ bounded for all i or (ii) G is independent of the residuals $e(t)$. Furthermore assume that $\frac{\sqrt{M}}{N} \to 0$ and $\bar{\rho} < \min(K_G, K)$. The $M \times K_G$ matrix of increments is denoted by G. Assume that*[26]

$$\sqrt{M}\left(\begin{pmatrix} F^\top F & F^\top G \\ G^\top F & G^\top G \end{pmatrix} - \begin{pmatrix} [F, F] & [F, G] \\ [G, F] & [G, G] \end{pmatrix}\right) \xrightarrow{L-s} N(0, \Pi).$$

Then

$$\sqrt{M}\left(\hat{\bar{\rho}} - \bar{\rho}\right) \xrightarrow{L-s} N(0, \Xi) \qquad and \qquad \frac{\sqrt{M}}{\sqrt{\Xi}}\left(\hat{\bar{\rho}} - \bar{\rho}\right) \xrightarrow{D} N(0, 1)$$

with $\Xi = \xi^\top \Pi \xi$ and ξ is equal to

$$vec\begin{pmatrix} -\left([F, F]^{-1}[F, G][G, G]^{-1}[G, F][F, F]^{-1}\right)^\top & [F, F]^{-1}[F, G][G, G]^{-1} \\ [G, G]^{-1}[G, F][F, F]^{-1} & -\left([G, G]^{-1}[G, F][F, F]^{-1}[F, G][G, G]^{-1}\right)^\top \end{pmatrix}.$$

[23] The generalized correlation is also known as canonical correlation.

[24] Although labeling the measure as a correlation, we do not demean the data. This is because the drift term essentially describes the mean of a semimartingale and when calculating or estimating the quadratic covariation it is asymptotically negligible. Hence, the generalized correlation measure is based only on inner products and the generalized correlations correspond to the singular values of the matrix $[F, G]$ if F and G are orthonormalized with respect to the inner product $[., .]$.

[25] It is well-known that if F and G are observed and *i.i.d.* normally distributed then $\frac{\sqrt{M}(\hat{\rho}_k^2 - \rho_k^2)}{2\rho_k(1-\rho_k^2)} \xrightarrow{D} N(0, 1)$ for $k = 1, \ldots, \min(K_F, K_G)$ where ρ_k is the kth generalized correlation. The result can also be extended to elliptical distributions. However, the normalized increments of stochastic processes that can realistically model financial time series are neither normally nor elliptically distributed. Hence, we cannot directly make use of these results as for example in Bai and Ng (2006).

[26] The statement should be read as $\sqrt{M}\left(vec\left(\begin{pmatrix} F^\top F & F^\top G \\ G^\top F & G^\top G \end{pmatrix}\right) - vec\left(\begin{pmatrix} [F, F] & [F, G] \\ [G, F] & [G, G] \end{pmatrix}\right)\right) \xrightarrow{L-s} N(0, \Pi)$, where vec is the vectorization operator. Inevitably the matrix Π is singular due to the symmetric nature of the quadratic covariation. A proper formulation avoiding the singularity uses $vech$ operators and elimination matrices (See Magnus (1988)).

In Proposition 5 in the Appendix we present a feasible test statistic for the estimated continuous factors. A feasible test for the jump factors can be derived analogously. The assumption that G has to be a well-diversified portfolio of the underlying asset space is satisfied by essentially all economic factors considered in practice, e.g. the market factor or the value, size and momentum factors. Hence, practically it does not impose a restriction on the testing procedure. This assumption is only needed to obtain the same distribution theory for the quadratic covariation of G with the estimated factors as with the true factors.

We have ruled out the special case of $\rho = min(K, K_G)$, which implies that the candidate factors are simply a rotation of the true factors, i.e. $G = \tilde{H}F$ for a full-rank $K \times K$ matrix $\tilde{H}$. This "corner case" leads to super-consistency in the estimation of the generalized correlation, similar to a unit-root case.

Proposition 1 (*Super-Consistency of Generalized Correlation*). *Assume Assumptions 1 and 2 hold and $G = \tilde{H}F$ for a full-rank $K \times K$ matrix $\tilde{H}$. Then*

$$\hat{\bar{\rho}} = \bar{\rho} + O_p\left(\frac{1}{\delta}\right).$$

If in addition Assumption 3 holds then $\hat{\bar{\rho}}^C = \bar{\rho}^C + O_p\left(\frac{1}{\delta}\right)$ and $\hat{\bar{\rho}}^D = \bar{\rho}^D + O_p\left(\frac{1}{\delta}\right)$.

The inferential theory for this special case is beyond the scope of this paper.[27] In most practical applications the candidate factors are only a noisy approximation of the true factors which puts us into the setup of Theorem 7.

5. Simulations

This section considers the finite sample properties of our estimators through Monte-Carlo simulations. In the first subsection we use Monte-Carlo simulations to analyze the distribution of our estimators for the loadings, factors and common components. In the second subsection we provide a simulation study of the estimator for the number of factors and compare it to the most popular estimators in the literature.

Our benchmark model is a Heston-type stochastic volatility model with jumps. In the general case we assume that the K factors and N residual processes are modeled as

$$\begin{aligned} dF_k(t) &= (\mu - \sigma^2_{F_k}(t))dt + \rho_F \sigma_{F_k}(t) dW_{F_k}(t) + \sqrt{1-\rho_F^2}\sigma_{F_k}(t) d\tilde{W}_{F_k}(t) + J_{F_k} dN_{F_k}(t) \\ d\sigma^2_{F_k}(t) &= \kappa_F\left(\alpha_F - \sigma^2_{F_k}(t)\right)dt + \gamma_F \sigma_{F_k}(t) d\tilde{W}_{F_k}(t) \\ de_i(t) &= \rho_e \sigma_{e_i}(t) dW_{e_i}(t) + \sqrt{1-\rho_e^2}\sigma_{e_i}(t) d\tilde{W}_{e_i}(t) + J_{e_i} dN_{e_i}(t) - \mathbb{E}[J_{e_i}]\nu_e dt \\ d\sigma^2_{e_i}(t) &= \kappa_e\left(\alpha_e - \sigma^2_{e_i}(t)\right)dt + \gamma_e \sigma_{e_i}(t) d\tilde{W}_{e_i}(t). \end{aligned}$$

The Brownian motions $W_F, \tilde{W}_F, W_e, \tilde{W}_e$ are assumed to be independent. We set the parameters to values typically used in the literature: $\kappa_F = \kappa_e = 5$, $\gamma_F = \gamma_e = 0.5$, $\rho_F = -0.8$, $\rho_e = -0.3$, $\mu = 0.05$, $\alpha_F = \alpha_e = 0.1$. The jumps are modeled as a compound Poisson process with intensity $\nu_F = \nu_e = 6$ and normally distributed jumps with $J_{F_k} \sim N(-0.1, 0.5)$ and $J_{e_i} \sim N(0, 0.5)$. The time horizon is normalized to $T = 1$.

In order to separate continuous from discontinuous movements we use the threshold $3\hat{\sigma}_X(j)\Delta_M^{0.48}$.[28] The spot volatility is estimated using Barndorff-Nielsen and Shephard's (2002) bi-power volatility estimator on a window of $\sqrt{M}$ observations. Under certain assumptions the bi-power estimator is robust to jumps estimating the volatility consistently.

In order to capture cross-sectional correlations we formulate the dynamics of X as $X(t) = \Lambda F(t) + Ae(t)$, where the matrix A models the cross-sectional correlation. If A is an identity matrix, then the residuals are cross-sectionally independent. The empirical results suggest that it is very important to distinguish between strong and weak factors. Hence the first factor is multiplied by the scaling parameter $\sigma_{dominant}$. If $\sigma_{dominant} = 1$ then all factors are equally strong. In practice, the first factor has the interpretation of a market factor and has a significantly larger variance than the other weaker factors. Hence, a realistic model with several factors should set $\sigma_{dominant} > 1$.

The loadings Λ are drawn from independent standard normal distributions. All Monte-Carlo simulations have 1000 repetitions. We first simulate a discretized model of the continuous time processes with 2000 time steps representing the true model and then use the data which is observed on a coarser grid with $M = 50$,100, 250 or 500 observations. Our results are robust to changing the number of Monte-Carlo simulations or using a finer time grid for the "true" process.

[27] Pelger and Xiong (2018) develop the inferential theory for the corner case in the long-horizon framework.

[28] Compare e.g. with Aït-Sahalia and Xiu (2017b) or Bollerslev et al. (2013). ω is typically chosen between 0.47 and 0.49 and the results are insensitive to this choice. Intuitively we classify all increments as jumps that are beyond 3 standard deviations of a local estimator of the stochastic volatility.

Table 1
Mean and standard deviations of estimated correlation coefficients between $\hat{F}(T)$ and $F(T)$ and $\hat{\Lambda}_i$ and Λ_i based on 1000 simulations.

	N=200, M=250					N=100, M=100				
	Case 1			Case 2	Case 3	Case 1			Case 2	Case 3
	Total	Cont.	Jump			Total	Cont.	Jump		
Corr. $F(T)$	0.994	0.944	0.972	0.997	0.997	0.986	0.789	0.943	0.994	0.997
SD $F(T)$	0.012	0.065	0.130	0.001	0.000	0.037	0.144	0.165	0.002	0.000
Corr. Λ	0.995	0.994	0.975	0.998	0.998	0.986	0.966	0.949	0.994	0.998
SD Λ	0.010	0.008	0.127	0.001	0.000	0.038	0.028	0.157	0.002	0.000
	N=500, M=50					N=50, M=500				
	Case 1			Case 2	Case 3	Case 1			Case 2	Case 3
	Total	Cont.	Jump			Total	Cont.	Jump		
Corr. $F(T)$	0.997	0.597	0.926	0.999	0.999	0.973	0.961	0.954	0.988	0.990
SD $F(T)$	0.006	0.196	0.151	0.001	0.000	0.067	0.028	0.141	0.005	0.002
Corr. Λ	0.979	0.921	0.906	0.987	0.990	0.991	0.997	0.974	0.999	0.999
SD Λ	0.027	0.051	0.175	0.005	0.002	0.053	0.002	0.128	0.001	0.000

5.1. Asymptotic distribution theory

In this subsection we consider only one factor in order to assess the properties of the limiting distribution, i.e. $K = 1$ and $\sigma_{dominant} = 1$. We consider three different cases:

1. **Case 1: Benchmark model with jumps.** The correlation matrix A is a Toeplitz matrix with parameters $(1, 0.2, 0.1)$, i.e. it is a symmetric matrix with diagonal elements 1 and the first two off-diagonals have elements 0.2 respectively 0.1.
2. **Case 2: Benchmark model without jumps.** This model is identical to case 1 but without the jump component in the factors and residuals.
3. **Case 3: Toy model.** Here all the stochastic processes are standard Brownian motions $X(t) = \Lambda W_F(t) + W_e(t)$. After rescaling case 3 is identical to the simulation study considered in Bai (2003).

Obviously, we can only estimate the continuous and jump factors in case 1.

In order to assess the accuracy of the estimators we calculate the correlations of the estimator for the loadings and factors with the true values. If jumps are included, we also have correlations for the continuous and jump estimators. In addition for $t = T$ and $i = N/2$ we calculate the asymptotic distribution of the rescaled and normalized estimators:

$$CLT_C = \left(\frac{1}{N}\hat{V}_{T,i} + \frac{1}{M}\hat{W}_{T,i}\right)^{-1/2}\left(\hat{C}_{T,i} - C_{T,i}\right), \quad CLT_F = \sqrt{N}\hat{\Theta}_F^{-1/2}(\hat{F}(T) - H^{-1}F(T))$$
$$CLT_\Lambda = \sqrt{M}\hat{\Theta}_{\Lambda,i}^{-1/2}(\hat{\Lambda}_i - H^\top \Lambda_i).$$

Table 1 reports the mean and standard deviation of the correlation coefficients between $\hat{F}(T)$ and $F(T)$ and $\hat{\Lambda}_i$ and Λ_i based on 1000 simulations. In case 1 we also estimate the continuous and jump parts. The correlation coefficient can be considered as a measure of consistency. For the factor processes the correlation is based on the quadratic covariation between the true and the estimated processes. We run the simulations for four combinations of N and M: $N = 200$, $M = 250$, $N = 100$, $M = 100$, $N = 500$, $M = 50$ and $N = 50$, $M = 500$. The correlation coefficients in all cases are very close to one, indicating that our estimators are very precise. Note that we can only estimate the continuous and jump factors up to a finite variation part. However, when calculating the correlations, the drift term is negligible. For a small number of high-frequency observations M the continuous and the jump factors are estimated with a lower precision as the total factor. This is mainly due to an imprecision in the estimation of the jumps. In all cases the loadings can be estimated very precisely. The simpler the processes, the better the estimators work. For sufficiently large N and M, increasing M improves the estimator for the loadings, while increasing N leads to a better estimation of the factors. Overall, the finite sample properties for consistency are excellent.

Figs. 1–3 and Table 7 in the Appendix summarize the simulation results for the normalized estimators CLT_C, CLT_F and CLT_Λ. The asymptotic distribution theory suggests that they should be $N(0, 1)$ distributed. Table 7 lists the means and standard deviations based on 1000 simulations. For the toy model in case 3 the mean is close to 0 and the standard deviation almost 1, indicating that the distribution theory works. Fig. 3 depicts the histograms overlaid with a normal distribution. The asymptotic theory provides a very good approximation to the finite sample distributions. Adding stochastic volatility and weak cross-sectional correlation still provides a good approximation to a normal distribution. The common component estimator is closer to the asymptotic distribution than the factor or loading estimator.[29] We have set the length of the local

[29] Even in case 1 with the additional jumps the approximation works well. The common component estimator still performs the best. Without an additional finite sample correction the loading estimator in case 1 would have some large outliers. In more detail, the derivations for case 1 assume that the time increments are sufficiently small such that the two independent processes $F(t)$ and $e_i(t)$ do not jump during the same time increment. Whenever this happens the rescaled loadings statistic explodes. For very few of the 1000 simulations in case 1 we observe this problem and exclude these simulations.

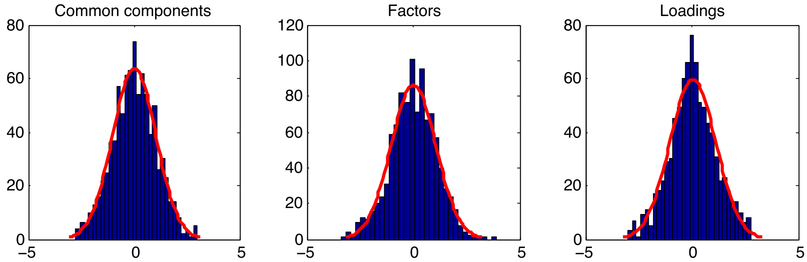

Fig. 1. Case 1 with $N = 200$ and $M = 250$. Histogram of standardized common components CLT_C, factors CLT_F and loadings CLT_Λ. The normal density function is superimposed on the histograms.

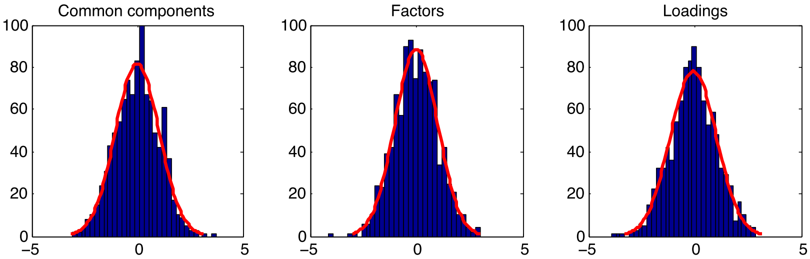

Fig. 2. Case 2 with $N = 200$ and $M = 250$. Histogram of standardized common components CLT_C, factors CLT_F and loadings CLT_Λ. The normal density function is superimposed on the histograms.

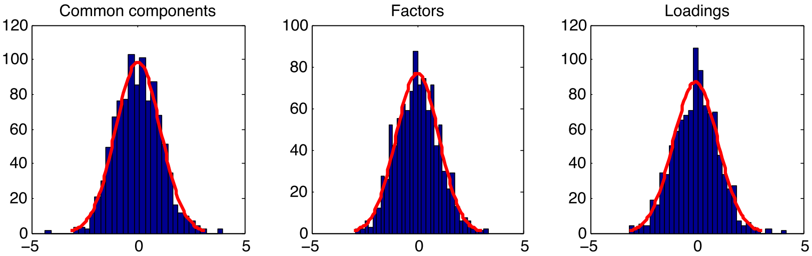

Fig. 3. Case 3 with $N = 200$ and $M = 250$. Histogram of standardized common components CLT_C, factors CLT_F and loadings CLT_Λ. The normal density function is superimposed on the histograms.

window in the covariance estimation of the loadings estimator to $k = \sqrt{M}$. The estimator for the covariance of the factors assumes cross-sectional independence, which is violated in the simulation example as well as Assumption 5. Nevertheless in the simulations the normalized statistics approximate a normal distribution very well. Overall, the finite sample properties for the asymptotic distribution work well.

5.2. Number of factors

In this subsection we analyze the finite sample performance of our diagnostic criterion for the number of factors and show that it performs as well or better than the most popular estimators in the literature. One of the main motivations for developing our estimator is that the assumptions needed for the Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013) estimator cannot be extended to the general processes that we need to consider.[30]

In the first part of this section we work with a variation of the toy model such that we can apply all four estimators and compare them:

$$X(t) = \Lambda W_F(t) + \theta A W_e(t),$$

where all the Brownian motions are independent and the $N \times N$ matrix A models the cross-sectional dependence, while θ captures the signal-to-noise ratio. The matrix A is a Toeplitz matrix with parameters $(1, a, a, a, a^2)$, i.e. it is a symmetric matrix with diagonal element 1 and the first four off-diagonals having the elements a, a, a and a^2. A dominant factor is modeled with $\sigma_{dominant} > 1$. Note that after rescaling this is the same model that is also considered in Bai and Ng, Onatski and Ahn and Horenstein. Hence, these results obviously extend to the long horizon framework. In the following simulations we always consider three factors, i.e. $K = 3$. We simulate four scenarios:

[30] In particular all three estimators assume essentially that the residuals can be written in the form BEA, where B is a $T \times T$ matrix capturing serial correlation, A is a $N \times N$ matrix modeling the cross-sectional correlation and E is a $T \times N$ matrix of i.i.d. random variables with finite fourth moments. Such a formulation rules out jumps and a complex stochastic volatility structure.

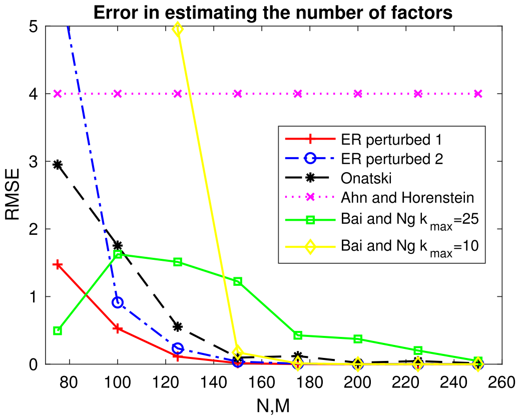

Fig. 4. Root-mean squared error in scenario 1 for $N = M$.

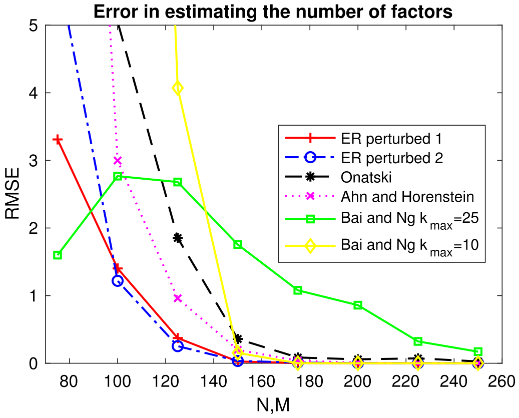

Fig. 5. Root-mean squared error in scenario 2 for $N = M$.

1. Scenario 1: Dominant factor, large noise-to-signal ratio, cross-sectional correlation: $\sigma_{dominant} = \sqrt{10}, \theta = 6$ and $a = 0.5$.
2. Scenario 2: No dominant factor, large noise-to-signal ratio, cross-sectional correlation: $\sigma_{dominant} = 1, \theta = 6$ and $a = 0.5$.
3. Scenario 3: No dominant factor, small noise-to-signal ratio, cross-sectional correlation: $\sigma_{dominant} = 1, \theta = 1$ and $a = 0.5$.
4. Scenario 4: Toy model: $\sigma_{dominant} = 1, \theta = 1$ and $a = 0$.

Our empirical studies suggest that in the data the first systematic factor is very dominant with a variance that is 10 times larger than those of the other weaker factors. Furthermore the idiosyncratic part seems to have a variance that is at least as large as the variance of the common components. Both findings indicate that scenario 1 is the most realistic case and an estimator of practical relevance should also work in this scenario.

Our perturbed eigenvalue ratio statistic has two choice parameters: the perturbation $g(N, M)$ and the cutoff γ. In the simulations we set the cutoff equal to $\gamma = 0.2$. For the perturbation we consider the two choices $g(N, M) = \sqrt{N} \cdot median\{\lambda_1, \dots, \lambda_N\}$ and $g(N, M) = \log(N) \cdot median\{\lambda_1, \dots, \lambda_N\}$. The first estimator is denoted by $ERP1$, while the second is $ERP2$. All our results are robust to these choice variables. For the Onatski (2010) estimator (labeled *Onatski*) we use the same parameters as in his paper. For the Ahn and Horenstein (2013) estimator (labeled *A&H*) we first demean the data in the cross-sectional and time dimension before applying principal component analysis as suggested in their paper. B&N denotes the BIC3 estimator of Bai and Ng (2002), which outperforms the other versions of the Bai and Ng estimators in simulations. For the last three estimators, we need to define an upper bound on the number of factors, which we set equal to $k_{max} = 25$. For the BIC3 estimator we also consider the case of $k_{max} = 10$.[31] For $ERP1$ and $ERP2$ we consider the whole spectrum. The figures and plots are based on 1000 simulations.[32]

Figs. 4–7 plot the root-mean squared error for the different estimators for a growing number $N = M$ and show that our estimators perform similarly or better than the other estimators. In the most relevant Scenario 1 depicted in Fig. 4 only the

[31] The BIC3 estimator requires an estimate of the average noise volatility that is based on the residuals after removing the first k_{max} principal components.

[32] Obviously there are more estimators in the literature, e.g. Harding (2013) and Hallin and Liska (2007). However, the simulation studies in their papers indicate that the Onatski and Ahn and Horenstein estimators dominate most other estimators.

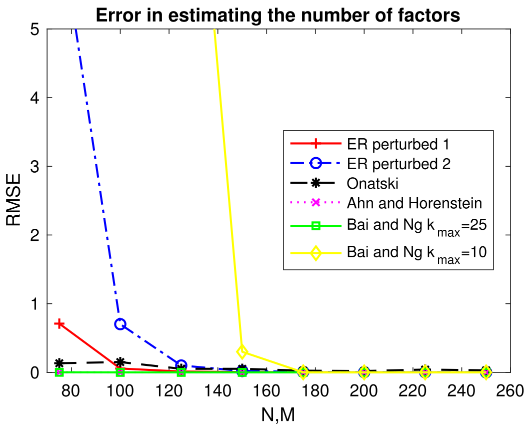

Fig. 6. Root-mean squared error in scenario 3 for $N = M$.

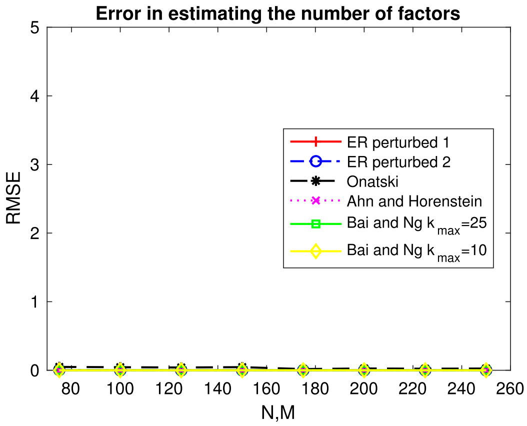

Fig. 7. Root-mean squared error in scenario 4 for $N = M$.

*ERP*1, *ERP*2 and *Onatski* estimators are reliable. This is because these three estimators focus on the residual spectrum and are not affected by dominant factors. Although we apply the demeaning as proposed in Ahn and Horenstein, their estimator misses the weaker factors. In scenario 1 *A&H* and *B&N* severely underestimate the number of factors, while the *ERP*1 and *ERP*2 estimators are the best. When reducing k_{max} for *B&N* it overestimates the number of factors.[33] In Fig. 5 we remove the dominant factor and the performance of *A&H* drastically improves. However *ERP*1 and *ERP*1 still show a comparable performance. In the less realistic Scenarios 3 and 4, all estimators are reliable and perform equally well. Note that for a sample size $N = M$ below 100, the estimators based on clustering arguments *ERP*1, *ERP*2 and *Onatski* can become unreliable.[34]

Figures 18 and 19 in the Supplementary Appendix show *ERP*1 applied to the benchmark model Case 1 from the last subsection. ERP1 can reliably estimate the number of continuous and jump factors. The plots also illustrate that even without choosing a cutoff threshold γ plotting the perturbed eigenvalue ratios is a very good first step for understanding the potential factor structure in the data.

6. Empirical application

6.1. Data

We use intraday log-prices from the Trade and Quote (TAQ) database for the time period from January 2003 to December 2012 for all the assets included in the S&P 500 index at any time between January 1993 and December 2012. In order to strike a balance between the competing interests of utilizing as much data as possible and minimizing the effect of microstructure

[33] Given a specific data set it is possible to find a k_{max} that improves the performance of *B&N*.

[34] Table 8 in the Appendix shows the summary statistics for all scenarios for $N = M = 125$ and confirms the above observations.

Table 2
(1) Fraction of increments identified as jumps for different thresholds. (2) Fraction of total quadratic variation explained by jumps for different thresholds. (3) Systematic jump correlation as measured by the fraction of the jump correlation explained by the first jump factor for different thresholds. (4) Systematic continuous correlation as measured by the fraction of the continuous correlation explained by the first four continuous factors.

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Percentage of increments identified as jumps										
a=3	0.011	0.011	0.011	0.010	0.010	0.009	0.008	0.008	0.007	0.008
a=4	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001
a=4.5	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.001
Variation explained by jumps										
a=3	0.19	0.19	0.19	0.16	0.21	0.16	0.16	0.15	0.12	0.15
a=4	0.07	0.07	0.07	0.05	0.10	0.06	0.06	0.06	0.03	0.05
a=4.5	0.05	0.04	0.05	0.04	0.08	0.04	0.05	0.05	0.02	0.04
Percentage of jump correlation explained by first 1 jump factor										
a=3	0.05	0.03	0.03	0.03	0.06	0.07	0.08	0.19	0.12	0.06
a=4	0.03	0.02	0.02	0.04	0.08	0.06	0.08	0.25	0.09	0.08
a=4.5	0.03	0.03	0.02	0.05	0.09	0.06	0.08	0.22	0.12	0.09
Percentage of continuous correlation explained by first 4 continuous factors										
	0.26	0.20	0.21	0.22	0.29	0.45	0.40	0.40	0.47	0.31

noise and asynchronous returns, we choose to use 5-minute prices.[35] More details about the data selection and cleaning procedures are in Appendix B. For each of the 10 years we have on average 250 trading days with 77 log-price increments per day. Within each year we have a cross-section N between 500 and 600 firms.[36] The exact number for each year is in Table 6 in the Appendix. After applying the cleaning procedure the intersection of the firms for the time period 2007–2012 is 498, while the intersection of all firms for the 10 years is 304. The yearly results use all the available firms in that year, while the analysis over longer horizons uses the cross-sectional intersection.

When identifying jumps, we face the tradeoff of finding all discontinuous movements against misclassifying high-volatility regimes as jumps. Therefore, the threshold should take into account changes in volatilities and intra-day volatility patterns. We use the *TOD* estimator of Bollerslev et al. (2013) for separating the continuous from the jump movements. Hence the threshold is set as $a \cdot 77^{-0.49}\hat{\sigma}_{j,i}$, where $\hat{\sigma}_{j,i}$ estimates the daily volatility of asset i at time j by combining an estimated Time-of-Day volatility pattern with a jump robust bipower variation estimator for that day. Intuitively we classify all increments as jumps that are beyond a standard deviations of a local estimator of the stochastic volatility. For our analysis we use $a = 3$, $a = 4$ and $a = 4.5$.

Table 2 lists the fraction of increments identified as jumps for different thresholds. Depending on the year for $a = 3$ more than 99% of the observations are classified as continuous, while less than 1% are jumps. In 2012, 99.2% of the movements are continuous and explain around 85% of the total quadratic variation, while the 0.8% jumps explain the remaining 15% of the total quadratic covariation. When increasing the threshold less movements are classified as jumps.[37] All the results for the continuous factors are extremely robust to this choice. However, the results for the jump factors are sensitive to the threshold. Therefore, we are very confident about the results for the continuous factors, while the jump factor results have to be interpreted with caution. If not noted otherwise, the threshold is set to $a = 3$ in the following.

As a first step Table 2 lists for each year the fraction of the total continuous variation explained by the first four continuous factors and the fraction of the jump variation explained by the first jump factor.[38] As expected systematic risk varies over time and is larger during the financial crisis. The systematic continuous risk with 4 factors accounts for around 40%–47% of the total correlation from 2008 to 2011, but explains only around 20%–31% in the other years.[39] A similar pattern holds for the jumps where the first jump factor explains up to 10 times more of the correlation in 2010 than in the years before the financial crisis.

[35] We have run robustness tests with 15 and 30 min data and the main results do not change.

[36] We do not extend our analysis to the time before 2003 as there are too many missing high-frequency observations for the large cross-section.

[37] There is no consensus on the number of jumps in the literature. Christensen et al. (2014) use ultra high-frequency data and estimate that the jump variation accounts for about 1% of total variability. Most studies based on 5 min data find that the jump variation should be around 10%–20% of the total variation. Our analysis considers both cases.

[38] We have applied the factor estimation to the quadratic covariation and the quadratic correlation matrix, which corresponds to using the covariance or the correlation matrix in long-horizon factor modeling. For the second estimator we rescale each asset for the time period under consideration by the square-root of its quadratic covariation. Of course, the resulting eigenvectors need to be rescaled accordingly in order to obtain estimators for the loadings and factors. All our results are virtually identical for the covariation and the correlation approach, but the second approach seems to provide slightly more robust estimators for shorter time horizons. Hence, all results reported in this paper are based on the second approach.

[39] The percentage of correlation explained by the first four factors is calculated as the sum of the first four eigenvalues divided by the sum of all eigenvalues of the continuous quadratic correlation matrix.

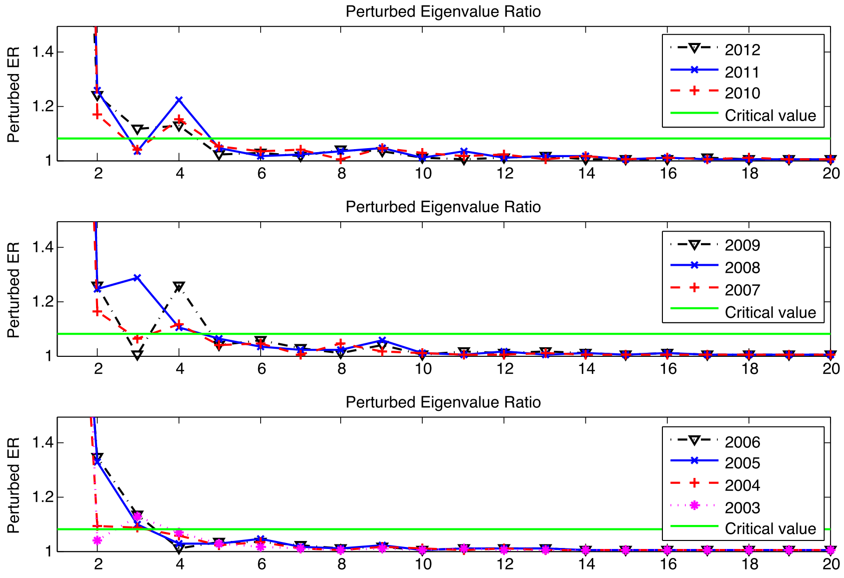

Fig. 8. Number of continuous factors.

6.2. *Continuous factors*

6.2.1. *Number of factors*

We estimate four continuous factors for each of the years from 2007 to 2012 and three continuous factors for the years 2003–2006. Fig. 8 shows the estimation results for the numbers of continuous factors. Starting from the right we are looking for a visible strong increase in the perturbed eigenvalue ratio[40]. Asymptotically any critical value larger than 1 should indicate the beginning of the systematic spectrum. However, for our finite sample we need to choose a critical value. In the plots we set the critical value equal to 1.08. Fortunately there are very visible humps at 4 for the years 2007–2012 and strong increases at 3 for the years 2003–2006, which can be detected for a wide range of critical values. Therefore, our diagnostic criterion strongly indicates that there are 4 continuous factors from 2007 to 2012 and three continuous factors from 2003 to 2006. As a robustness test in Figure 10 in the Appendix we also use an unperturbed eigenvalue ratio statistic. The results are very similar.

In Figure 9 in the Appendix we apply the same analysis without separating the data into continuous and jump components and obtain the same number of factors as in the continuous case. The perturbed eigenvalue ratios stop to cluster at the value 4 for 2007–2012 and at the value 3 for 2003–2006. This implies either that the continuous and jump factors are the same or that the continuous factors dominate the jump factors.

6.2.2. *Interpretation of factors*

The four stable continuous factors for 2007–2012 can be approximated very well by industry factors. The loading estimators can essentially be interpreted as portfolio weights for the factor construction. Simple eyeballing indicates that the first statistical factor seems to be an equally weighted market portfolio, a result which has already been confirmed in many studies. The loadings for the second to fourth statistical factors have a very particular pattern: Banks and insurance companies have very large loadings with the same sign, while firms related to oil and gas have large loadings with the opposite sign. Firms related to electricity seem to have their own pattern unrelated to the previous two. Motivated by these observations we construct four economic factors as (1) Market (equally weighted), (2) Oil and gas (40 equally weighted assets), (3) Banking and Insurance (60 equally weighted assets) and Electricity (24 equally weighted assets).[41]

The generalized correlations of the market, oil and finance factors with the first four largest statistical factors for 2007–2012 are very high as shown in the first analysis of Table 3. This indicates that three of the four statistical factors can almost perfectly be replicated by the three economic factors. This relationship is very stable over time. In Table 3 the top of the first column uses the factors and generalized correlations based on a 6 year horizon, while in the last six columns we estimate the yearly statistical factors and calculate their generalized correlations with the yearly market, oil and finance factors. The generalized correlations close to one indicate that at least three of the statistical factors do not change over time and are stable.

[40] We use the median eigenvalue rescaled by $\sqrt{N}$ for the perturbation term g. We have conducted the same analysis for more perturbation functions with the same findings. The results are available upon request.

[41] The details are in Appendix B.

Table 3
Interpretation of statistical continuous factors. Generalized correlation of economic factors (market, oil, finance and electricity factors) with first four largest statistical factors for different time periods.

Generalized correlations of 4 continuous factors with market, oil and finance factors										
N=498	2007–2012				2007	2008	2009	2010	2011	2012
1. Gen. Corr.	**1.00**				1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	**0.98**				0.98	0.97	0.98	0.97	0.98	0.93
3. Gen. Corr.	**0.95**				0.91	0.95	0.94	0.93	0.97	0.87
Generalized correlations of 4 continuous factors with market, oil, finance and electricity factors										
N=498	2007–2012				2007	2008	2009	2010	2011	2012
1. Gen. Corr.	**1.00**				1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	**0.98**				0.98	0.97	0.99	0.97	0.98	0.93
3. Gen. Corr.	**0.95**				0.91	0.95	0.95	0.93	0.94	0.90
4. Gen. Corr.	**0.80**				0.87	0.78	0.75	0.75	0.80	0.76
Generalized correlations of 4 continuous factors with market, oil, finance and electricity factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.97	0.99	1.00	1.00	0.99	0.97	0.98	0.96	0.98	0.95
3. Gen. Corr.	0.57	0.75	0.77	0.89	0.85	0.92	0.95	0.92	0.93	0.83
4. Gen. Corr.	**0.10**	**0.23**	**0.16**	**0.35**	0.82	0.74	0.72	0.68	0.78	0.78
Generalized correlations of 4 continuous factors with market, oil and finance factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.97	0.99	1.00	1.00	0.99	0.97	0.98	0.96	0.97	0.94
3. Gen. Corr.	**0.46**	**0.49**	**0.47**	**0.49**	0.84	0.92	0.94	0.89	0.93	0.83
Generalized correlations of 4 continuous factors with market, oil and electricity factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.97	0.99	1.00	1.00	0.98	0.97	0.95	0.94	0.96	0.93
3. Gen. Corr.	**0.36**	**0.64**	**0.97**	**0.84**	0.83	0.76	0.73	0.69	0.78	0.78
Generalized correlations of 4 continuous factors with market, finance and electricity factors										
N=302	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1. Gen. Corr.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. Gen. Corr.	0.57	0.75	0.98	0.89	0.88	0.92	0.98	0.94	0.95	0.85
3. Gen. Corr.	**0.19**	**0.27**	**0.57**	**0.45**	0.83	0.74	0.73	0.72	0.78	0.78

Identifying the fourth continuous factor is challenging and the closest approximation seems to be an electricity factor. The second analysis in Table 3 shows the generalized correlations of the four continuous statistical factors for 2007–2012 with the four economic factors. The fourth generalized correlation essentially measures how well the additional electricity factor can explain the remaining statistical factor. The fourth yearly generalized correlation takes values between 0.75 and 0.87, which means that the electricity factor can help substantially to explain the statistical factors, but it is not sufficient to perfectly replicate them. The first column shows the result for the total six year time horizon while the last six columns list the yearly results. In conclusion it seems that the relationship between the four economic and statistical factors is stable over time.

The third analysis in Table 3 shows that as expected one factor disappears in the early four years. A fourth generalized correlation between 0.16 and 0.35 for 2003–2006 suggests strongly that the statistical factors and industry factors have at most three factors in common. The fourth, fifth and sixth analyses in Table 3 try to identify the disappearing factor. Looking at the fifth analysis it seems that dropping the finance factor for the time period 2003–2006 leads to the smallest reduction in generalized correlations, i.e. the three statistical factors for 2003–2006 are not well-explained by a finance factor. On the other hand this finance factor is crucial for explaining the statistical factors for 2007–2012.

As a statistical measure for the closeness between the continuous statistical and economic factors, we calculate the total generalized correlations and their confidence intervals. The left part of Table 4 lists the total generalized correlation for different time periods for three economic factors while the right table does the same for four economic factors. The standard deviations are very small leading to very tight confidence intervals with the exception of the years 2008 and 2009, where the confidence intervals are somewhat wider. Our total generalized correlation statistic confirms that the industry factors closely approximate the statistical factors.

6.3. *Jump factors*

There seems to be a lower number of jump factors, which do not coincide with the continuous factors. Only the jump market factor seems to be stable, while neither the number nor the structure of the other jump factors have the same stability as for the continuous counterpart. Figures 11, 12 and 13 estimate the number of jump factors for different thresholds. In most

Table 4
Total generalized correlations (=sum of squared generalized correlations) with standard deviations and confidence intervals for the four statistical factors with three economic factors (market, oil and finance) and four economic factors (additional electricity factor). Number of assets $N = 498$.

	4 statistical and 3 economic factors			4 statistical and 4 economic factors		
	$\hat{\hat{\rho}}$	SD	95% CI	$\hat{\hat{\rho}}$	SD	95% CI
2007–2012	2.72	0.001	(2.71, 2.72)	3.31	0.003	(3.30, 3.31)
2007	2.55	0.06	(2.42, 2.67)	3.21	0.01	(3.19, 3.22)
2008	2.66	0.08	(2.51, 2.81)	3.18	0.29	(2.62, 3.75)
2009	2.86	0.10	(2.67, 3.05)	3.42	0.15	(3.14, 3.71)
2010	2.80	0.04	(2.72, 2.88)	3.38	0.01	(3.37, 3.39)
2011	2.82	0.00	(2.82, 2.82)	3.47	0.06	(3.35, 3.58)
2012	2.62	0.03	(2.56, 2.68)	3.25	0.01	(3.24, 3.26)

Table 5
Generalized correlations of market, oil, finance and electricity jump factors with first 4 jump factors from 2007–2012 for N=498 and for different thresholds.

Generalized correlations of 4 economic jump with 4 statistical jump factors							
	2007–2012	2007	2008	2009	2010	2011	2012
a=3	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**	**1.00**
	0.85	**0.95**	0.62	**0.86**	**0.81**	**0.86**	**0.83**
	0.61	0.77	0.40	0.76	0.31	0.61	0.59
	0.21	0.10	0.22	0.50	0.10	0.20	0.28
a=4	**0.99**	**0.99**	**0.95**	**0.94**	**1.00**	**0.99**	**0.99**
	0.74	0.53	0.41	0.59	**0.90**	0.53	0.57
	0.31	0.35	0.29	0.44	0.39	0.35	0.42
	0.03	0.19	0.20	0.09	0.05	0.14	0.16
a=4.5	**0.99**	**0.99**	**0.91**	**0.91**	**1.00**	**0.98**	**0.99**
	0.75	0.54	0.41	0.56	**0.93**	0.55	0.75
	0.29	0.35	0.30	0.40	0.68	0.38	0.29
	0.05	0.18	0.22	0.04	0.08	0.03	0.05

years the estimator indicates only one jump factor. Under almost all specifications there seem to be at most four jump factors and hence we will restrict the following analysis to the first four largest jump factors.[42]

Table 5 confirms that the jump factors are different from the continuous factors. Here we estimate the generalized correlations of the first four statistical jump factors with the market, oil, finance and electricity jump factors for 2007–2012. We can show that the first statistical jump factor is essentially the equally weighted market jump factor which is responsible for the first generalized correlation to be equal to 1. However, the correlations between the other statistical factors and the industry factors are significantly lower.

7. Conclusion

This paper studies factor models in the setting of a large cross-section and many high-frequency observations under a fixed time horizon. We propose a principal component estimator based on the increments of the observed time series, which is a simple and feasible estimator. For this estimator we develop the asymptotic distribution theory. Using a simple truncation approach the same methodology allows us to estimate continuous and jump factors. Our results are obtained under general conditions for the stochastic processes and allow for cross-sectional and serial correlation in the residuals. We also propose a novel diagnostic criterion for the number of factors, that can also consistently estimate the number of continuous and jump factors. Furthermore, we provide the inferential theory for a new statistic that compares estimated statistical factors with observed economic factors. We apply the estimation approaches to 5 min high-frequency price data of S&P 500 firms from 2003 to 2012. We show that the continuous factor structure is very stable in some years, but there is also time variation in the number and structure of factors over longer horizons. For the time period 2007–2012 we estimate four continuous factors which can be approximated well by a market, oil, finance and electricity factor. From 2003 to 2006 one continuous systematic factor disappears. Systematic jump risk seems to be different from systematic continuous risk[43].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2018.09.004.

[42] Our estimator for identifying the jumps might erroneously classify high volatility time periods as jumps. Increasing the threshold in the estimator reduces this error, while we might misclassify small jumps as continuous movements. Increasing the threshold, reduces the stability in the jump factors up to the point where only a market jump factor remains. It is unclear if the stability in the jump factor structure for small jump thresholds is solely due to misclassified high volatility movements.

[43] The empirical companion paper Pelger (2018) studies the asset pricing implications of the high-frequency factors.

References

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

Aït-Sahalia, Y., Fan, J., Xiu, D., 2010. High-frequency estimates with noisy and asynchronous financial data. J. Amer. Statist. Assoc. 105, 1504–1516.

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

Aït-Sahalia, Y., Mykland, P.A., Zhang, L., 2005b. A tale of two time scales: determining integrated volatility with noisy high-frequency data. J. Amer. Statist. Assoc. 100, 1394–1411.

Aït-Sahalia, Y., Xiu, D., 2017a. Principal component analysis of high frequency data. J. Amer. Statist. Assoc..

Aït-Sahalia, Y., Xiu, D., 2017b. Principal component estimation of a large covariance matrix with high-frequency data. J. Econometrics 201, 384–399.

Amengual, D., Watson, M., 2007. Consistent estimation of the number of dynamic factors in a large n and t panel. J. Bus. Econom. Statist. 25 (1), 91–96.

Andersen, T.G., Benzoni, L., Lund, J., 2002. An empirical investigation of continuous-time equity return models. J. Finance 57 (4), 1239–1284.

Andersen, T., Bollerslev, T., Diebold, F.X., Labys, P., 2001. The distribution of realized exchange rate volatility. J. Amer. Statist. Assoc. 42, 42–55.

Back, K., 1991. Asset prices for general processes. J. Math. Econom. 20, 371–395.

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

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

Bai, J., Ng, S., 2006. Evaluating latent and observed factors in macroeconomics and finance. J. Econometrics (131), 507–537.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2008. Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. Econometrica 76, 1481–1536.

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

Barndorff-Nielsen, O., Shephard, N., 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. J. Roy. Statist. Soc. 253–280.

Bibinger, M., Winkelmann, L., 2014. Econometrics of co-jumps in high-frequency data with noise. J. Econometrics 184 (2), 361–378.

Bollerslev, T., Li, S.Z., Todorov, V., 2013. Jump tails, extreme dependencies and the distribution of stock returns.. J. Econometrics 172, 307–324.

Bollerslev, T., Li, S.Z., Todorov, V., 2016. Roughing up beta: Continuous vs. Discontinuous betas, and the cross section of expected stock returns. J. Financ. Econ. 120, 464–490.

Bollerslev, T., Todorov, V., 2010. Jumps and betas: A new theoretical framework for disentangling and estimating systematic risks. J. Econometrics 157, 220–235.

Chamberlain, G., 1988. Asset pricing in multiperiod securities markets. Econometrica 56, 1283–1300.

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

Christensen, K., Oomen, R.C.A., Podolskij, M., 2014. Fact or friction: Jumps at ultra high frequency. J. Financ. Econ. 114 (3), 576–599.

Duffie, D., Pan, J., Singleton, K.J., 2000. Transform analysis and asset pricing for affine jump-diffusions. Econometrica 68 (6), 1343–1376.

Fan, J., Furger, A., Xiu, D., 2014. Incorporating Global Industrial Classification Standard into Portfolio Allocation: A Simple Factor-Based Large Covariance Matrix Estimator with High Frequency Data. Working paper.

Fan, L., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. Roy. Statist. Soc. 75, 603–680.

Hallin, M., Liska, R., 2007. The generalized dynamic factor model: determining the Number of Factors. J. Amer. Statist. Assoc. 102, 603–617.

Hansen, P., Lunde, A., 2006. Realized variance and market microstructure noise. J. Bus. Econom. Statist. 24, 127–161.

Harding, M., 2013. Estimating the number of factors in large dimensional factor models. J. Econometrics.

Jacod, J., 2008. Asymptotic properties of realized power variations and related functionals of semimartingales. Stochastic Process. Appl. 118, 517–559.

Jacod, J., Li, Y., Mykland, P., 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.

Kapetanios, G., 2010. A testing procedure for determining the number of factors in approximate factor models. J. Bus. Econom. Statist. 28, 397–409.

Lee, S.S., Mykland, P.A., 2008. Jumps in financial markets: A new nonparametric test and jump dynamics. Rev. Financ. Stud. 21, 2535–2563.

Magnus, J.R., 1988. Linear Structures. Oxford University Press.

Mancini, C., 2009. Non parametric threshold estimation for models with stochastic diffusion coefficient and jumps. Scand. J. Stat. 42–52.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Rev. Econ. Stat. 92, 1004–1016.

Pelger, M., 2018. Understanding Systematic Risk: A High-Frequency Approach. Working paper.

Pelger, M., Xiong, R., 2018. State-Varying Factor Models of Large Dimensions. Working Paper.

Podolskij, M., Vetter, M., 2009. Bipower-type estimation in a noisy diffusion setting. Stochastic Process. Appl. 11, 2803–2831.

Ross, S.A., 1976. The arbitrage theory of capital asset pricing. J. Econom. Theory 13, 341–360.

Tao, M., Wang, Y., Chen, X., 2013a. Fast convergence rates in estimating large volatility matrices using high-frequency financial data. Econometric Theory 29, 838–856.

Tao, M., Wang, Y., Zhou, H.H., 2013b. Optimal sparse volatility matrix estimation for high dimensional Itô processes with measurement errors. Ann. Statist. 41, 1816–1864.

Wang, Y., Zhou, 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., 2011. Estimating covariation: Epps effect, microstructure noise. J. Econometrics 160, 33–47.

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.

Zheng, X., Li, Y., 2011. On the estimation of integrated covariance matrices of high dimensional diffusion processes. Ann. Statist. 39, 3121–3151.