Journal of Econometrics 216 (2020) 86–105

Contents lists available at ScienceDirect

Journal of Econometrics

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

High-frequency factor models and regressions☆

Yacine Aït-Sahalia [a,*], Ilze Kalnina [b], Dacheng Xiu [c]

[a] Department of Economics and Bendheim Center for Finance, Princeton University, JRR Building, Princeton, NJ 08544, United States of America

[b] Department of Economics, Poole College of Management, North Carolina State University, Nelson Hall, Box 8110, Raleigh, NC 27965, United States of America

[c] Booth School of Business, University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637, United States of America

ARTICLE INFO

Article history:
Available online 7 February 2020

JEL classification:
C13
C14
C55
C58
G01

Keywords:
Factor model
Time-varying betas
Fama–French factors
Idiosyncratic risk
Big data

ABSTRACT

We consider a nonparametric time series regression model. Our framework allows precise estimation of betas without the usual assumption of betas being piecewise constant. This property makes our framework particularly suitable to study individual stocks. We provide an inference framework for all components of the model, including idiosyncratic volatility and idiosyncratic jumps. Our empirical analysis investigates the largest dataset in the high-frequency literature. First, we use all traded stocks from NYSE, AMEX, and NASDAQ stock markets for 1996–2017 to construct the five Fama–French factors and the momentum factor at the 5-minute frequency. Second, we document the key empirical properties across all the stocks and the new factors, and apply the nonparametric time series regression model with the new high-frequency Fama–French factors. We find that this factor model is effective in explaining the systematic component of the risk of individual stocks. In addition, we provide evidence that idiosyncratic jumps are related to idiosyncratic events such as earnings disappointments.
© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Time series regressions of asset returns on Fama–French factors are commonly used in financial economics. This regression model is often estimated using rolling windows of short intervals of time such as one month, due to concerns over the time variation in the factor betas, and other potential forms of nonstationarity of the model. However, when relatively low-frequency data such as daily data is used, the monthly estimates can be noisy. The problem of noisy estimates due to a small number of time series observations can be somewhat alleviated by some additional averaging, for instance by forming portfolios of stocks with similarly estimated factor betas.

An alternative is to use high-frequency data. Because the factor betas are effectively covariance-like quantities, they benefit from the additional data collected at high frequency. Using relatively short time windows, we can decompose the idiosyncratic risk, which is usually measured as the variation of the stock returns once the impact of the common

☆ The paper previously circulated under the title "The Idiosyncratic Volatility Puzzle: a Reassessment at High Frequency." We are grateful to two referees and the Editor for very helpful comments. We thank seminar and conference participants at Duke University, the Stevanovich Center for Financial Mathematics at the University of Chicago, Measuring and Modeling Financial Risk with High Frequency Data in Florence, Econometric Study Group Conference in Bristol, Financial Statistics Conference in Chicago, Canadian Econometric Study Group Conference, Conference on High-Frequency Financial Data in Montreal, Time Series and Financial Econometrics Conference in Montreal, Princeton-QUT-SJTU-SMU econometrics conference, and Financial Econometrics Conference in Toulouse.

* Corresponding author.
E-mail addresses: yacine@princeton.edu (Y. Aït-Sahalia), ikalnin@ncsu.edu (I. Kalnina), dacheng.xiu@chicagobooth.edu (D. Xiu).

https://doi.org/10.1016/j.jeconom.2020.01.007
0304-4076/© 2020 Elsevier B.V. All rights reserved.

factors has been removed, into its continuous and jump components. Our model can be thought of as a nonparametric continuous-time regression model. The framework we employ allows for general time variation in factor betas, and so is particularly suitable to study individual stocks. We provide the asymptotic theory for the estimators of time-varying betas and each of the idiosyncratic risk components.

Empirically, we perform a much larger scale analysis than what has been done so far in the high-frequency literature. We use all traded stocks from NYSE, AMEX, and NASDAQ stock markets during the period 1996–2017 to construct high frequency versions of the Fama–French factors (see Fama and French, 1992, 1993, 2015) and the momentum factor (see Carhart, 1997) at the 5-min frequency. We document the key empirical properties across all the stocks and the new factors. We then investigate the potential of the new high-frequency factors in explaining the individual stock returns. Finally, we study the behavior of the individual stock betas and the components of the idiosyncratic risk measures.

An important preliminary step in our empirical analysis is the study of the appropriate sampling frequency for individual stocks. Given the large number of stocks in the sample (5005 on average), the liquidity naturally varies widely across stocks; it also changes over time. For every stock and every month, we select a sampling frequency where the stock is liquid enough and its return does not have significant market microstructure effects. For this purpose, we use the Hausman test of Aït-Sahalia and Xiu (2019a) to check for the absence of statistically significant market microstructure noise at a given frequency, and also take into account the number of zero returns at any potential frequency. We choose between 5-min, 10-min, 30-min, and daily frequencies. We document the variation across stocks and over time of the selected frequencies. Our procedure results in a clear increase towards higher frequencies over time, with the steepest increases following the decimalization in 2001. In the last ten years, some intraday frequency is appropriate for the majority of stocks.

Our further empirical findings can be summarized as follows. The estimated high-frequency betas are similar to the standard betas calculated at daily frequency, but the former are more precisely estimated and are more stable across time, especially towards the end of the sample. The additional high-frequency factors are helpful in explaining additional time-series variation in stock returns compared to the high-frequency market factor alone. The idiosyncratic risk estimates at high-frequency and daily frequency are also comparable, though again the former seem to be more accurate towards the end of our sample. Finally, we decompose the idiosyncratic risk into an idiosyncratic volatility and idiosyncratic jumps contributions, and we find that earnings surprises increase idiosyncratic jumps. Moreover, earnings disappointments have a larger effect on the idiosyncratic jumps than do positive earnings surprises. From the perspective of estimating the continuous component of the model using discrete data, jumps can be thought of as outliers in the same spirit as Box and Tiao (1968) and Chang et al. (1988).

The literature on nonparametric regressions at high-frequency is closely related. A realized beta estimator, constructed as the ratio of realized covariance to realized variance, was proposed in Barndorff-Nielsen and Shephard (2004) and Andersen et al. (2005). These papers do not allow for jumps, and the implicit regression model has constant betas over the time interval considered, such as a week or a month. When estimating the model on moving windows, the assumption is effectively that of piecewise constant betas. When jumps are also allowed, realized beta still estimates a meaningful quantity, provided the continuous and jump components of factors are constrained to have the same effect on the stock return. Todorov and Bollerslev (2010) maintain the piecewise constant betas assumption, but allow the continuous and jump betas to differ. Bollerslev et al. (2016) consider a closely related model. A regression relationship between jumps alone is considered by Li et al. (2017) who also assume piecewise constant betas.

By contrast, we allow for general time variation in betas. Nonparametric time series regression models with time-varying betas have been previously considered by Mykland and Zhang (2006) and Reiß et al. (2015). These papers assume there are no jumps and the regressor is scalar. Jacod and Rosenbaum (2013) develop a general inference framework that is useful to study the continuous components of more general regressions with time-varying betas, see also Li et al. (2016), Kalnina and Tewou (2017) and Kalnina and Xiu (2017). Neither these nor other results in the literature allow to conduct inference on the idiosyncratic jump risk, which we develop.

Our framework is more general than nonparametric jump-diffusions (see, e.g., Aït-Sahalia et al., 2009; Ang and Kristensen, 2012, and Bandi and Phillips, 2003), since we work with Itô semimartingales. We make no restrictions on the continuous or jump leverage effects, we allow for jumps in levels and volatilities of our processes, and we allow for general time-variations in factor betas. As a result, our betas may well depend on firm characteristics or macroeconomic variables, a specification that is popular in empirical finance, except we do not need to know what those variables are, as long as they satisfy some weak regularity assumptions.

The paper is organized as follows. Section 2 presents the continuous-time regression model and lists our assumptions. Section 3 outlines our identification strategy and presents the estimators. Section 4 presents their asymptotic properties. Section 5 provides Monte Carlo simulation evidence. Section 6 gives the details on the construction of the high-frequency factors, and presents the empirical results. Section 7 concludes. The appendix contains the proofs.

2. The model

We consider the following nonparametric time series regression model,

$$Y_t = Y_0 + \int_0^t \beta_{s-}^\intercal dX_s^c + \sum_{0\le s\le t} \tilde{\beta}_{s-}^\intercal \Delta X_s + Z_t, \tag{1}$$

where Y is the dependent process, X is a d-dimensional multivariate covariate process, and Z is the residual process. In the above, X^c denotes the continuous component of X, and ΔX_s denotes its jump (if any) at time s. β_t and $\tilde{\beta}_t$ are the factor loadings with respect to the continuous and the jump parts of X.

This model, cast in continuous time, is analogous by the discrete-time factor model with observable factors that is widely used in macroeconomics and empirical finance. The continuous-time model is a natural choice for modeling high-frequency data, particularly for intraday transaction prices. In contrast to the standard factor model in empirical asset pricing, the drift or the "expected return" is absorbed into the residual process Z_t for convenience. Within a short period of time, such as a day or a week, the drift is economically negligible, econometrically not identifiable over a fixed time horizon, and does not affect the measurement of volatilities and risk which are the main objectives of the paper.

Time-variation in betas is an important component of our model. While an assumption of constant betas may be used relatively safely for portfolios, it is not supported by data for many individual stocks over periods such as one month (the period we use in our empirical application), see, e.g., Kalnina (2015) and Reiß et al. (2015).

We assume that X and Z are Itô semimartingales satisfying standard regularity assumptions. We make no restrictions on possible continuous and/or jump leverage effects, allow for jumps in returns and volatilities of X and Z, and allow for general time-variations in the factor betas:

Assumption 1. Suppose that the d-dimensional vector-valued process X follows a general Itô semimartingale with the representation[1]

$$X_t = X_0 + \int_0^t b_s\, ds + \int_0^t \sigma_s dW_s + \delta \star \mu_t, \tag{2}$$

where W is a d'-dimensional Brownian Motion, and μ is a Poisson random measure with compensator ν of the form $\nu(dt, du) = dt \otimes \lambda(du)$ for some σ-finite measure λ on $\mathbb{R}^d$. The stochastic volatility process σ takes values in $\mathbb{R}^{d\otimes d'}$. The spot covariance, denoted as $c_t = \sigma_t \sigma_t^\top$, follows

$$c_t = c_0 + \int_0^t \tilde{b}_s ds + \int_0^t \tilde{\sigma}_s dW_s + (\tilde{\delta} 1_{\{\|\tilde{\delta}\| \le 1\}}) \star (\mu - \nu)_t + (\tilde{\delta} 1_{\{\|\tilde{\delta}\| > 1\}}) \star \mu_t. \tag{3}$$

The drift terms b_t and $\tilde{b}_t$ are progressively measurable and locally bounded, the processes σ_t and $\tilde{\sigma}_t$ are càdlàg, and c_t is bounded away from 0. In addition, for some $r \in [0, 1)$, there is a sequence of stopping times (τ_m) increasing to ∞, and λ-integrable deterministic functions $\varkappa_m$ and $\tilde{\varkappa}_m$ such that $\|\delta(\omega, t, x)\|^r \wedge 1 \le \varkappa_m(x)$ and $\|\tilde{\delta}(\omega, t, x)\|^2 \wedge 1 \le \tilde{\varkappa}_m(x)$, for all (ω, t, x) with $t \le \tau_m(\omega)$.

Assumption 2. Suppose the stochastic process Y follows the continuous-time multiple regression model given by (1), in which β_t and $\tilde{\beta}_t$ are predictable and locally bounded, and Z_t is another Itô semimartingale:

$$Z_t = Z_0 + \int_0^t h_s ds + \int_0^t \gamma_s dB_s + \varepsilon \star \varphi_t. \tag{4}$$

In addition, h_s is progressively measurable and locally bounded, and the process γ_t is càdlàg. φ is a Poisson random measure with compensator ξ of the form $\xi(dt, du) = dt \otimes \pi(du)$. Finally, there is a sequence of stopping times (τ_m) increasing to ∞, and a π-integrable deterministic function ζ_m such that $\|\varepsilon(\omega, t, x)\|^r \wedge 1 \le \zeta_m(x)$, for all (ω, t, x) with $t \le \tau_n(\omega)$, where r is, without loss of generality, the same as in Assumption 1.

Similarly to the classical regression framework, we also need an exogeneity assumption for identification. We impose such an assumption on the continuous and jump components. Denote by A^c and A^j the continuous and jump components, respectively, of any process A. For example, $Z^j = \varepsilon \star \varphi$, and $Z^c = Z - Z^j$.

Assumption 3. For any $0 \le s \le t$, the following orthogonality equalities hold:

$$[Z_s^c, \int_0^s \beta_h^\top dX_h^c] = 0, \quad \text{and} \tag{5}$$

$$[Z_s^j, \sum_{0 \le h \le s} \tilde{\beta}_h^\top \Delta X_h] = 0, \tag{6}$$

where $[\cdot, \cdot]$ denotes the quadratic covariation.

For example, if Y_t is a log-price of a stock, then changes in Z_t are usually interpreted as the stock idiosyncratic returns relative to the factors X_t. The identification Assumption 3 allows us to identify the decomposition of the total risk of the stock (measured by $[Y_t, Y_t]$, the quadratic variation of Y) into two components: the systematic risk component (as

[1] See Aït-Sahalia and Jacod (2014) for more details. The use of W and μ in both Eqs. (2) and (3) is without loss of generality, see, e.g., Section 8.1.3 of Jacod and Protter (2012).

measured by the quadratic variation of stock price that is due to the common factors X_t), and the stock idiosyncratic risk (measured by $[Z_t, Z_t]$). To measure the relative importance of the systematic risk over a fixed time interval $[0, t]$, we use

$$R_t^2 = 1 - \frac{[Z_t, Z_t]}{[Y_t, Y_t]}, \tag{7}$$

the continuous-time counterpart of the coefficient of determination.

We are also interested in decomposing the total idiosyncratic risk into the idiosyncratic volatility and the idiosyncratic jump risk. In our model, this corresponds to the decomposition of the quadratic variation of Z_t into its continuous and jump components. In particular, we define the Idiosyncratic Volatility (IdV) and Idiosyncratic Jumps (IdJ) as follows:

$$IdV_t = \frac{1}{t}\int_0^t \gamma_s^2 ds, \quad \text{and} \quad IdJ_t = \frac{1}{t}\sum_{s\le t}(\Delta Z_s)^2, \tag{8}$$

where γ_s^2 is the spot variance of Z_s, and $\Delta Z_s = Z_s - Z_{s-}$ is the jump size of Z when there is one at time s. Clearly,

$$\frac{1}{t}[Z_t, Z_t] = IdV_t + IdJ_t.$$

In addition, we define the Integrated Beta ($I\beta$) as the average of the time-varying "spot beta" β_s:

$$I\beta_t = \frac{1}{t}\int_0^t \beta_s ds. \tag{9}$$

Note that $I\beta_t$ is a d-dimensional vector. We do not estimate the "jump beta" $\tilde{\beta}$ and note that additional assumptions would be required for its identification (e.g. a constant "jump beta" assumption). Otherwise, when $d = 1$, $\tilde{\beta}_s$ is identified only if X jumps at s. When $d > 1$, there is no time point s when all d components of $\tilde{\beta}_s$ are identified. While additional assumptions would be needed to identify $\tilde{\beta}_s$, we show below that the current set of assumptions is sufficient for the identification and estimation of the remaining quantities of interest.

3. Econometric strategy

The strategy for estimation is conceptually simple. Similarly to the standard factor model with observable factors, we use OLS-type regressions for the continuous components of the factor model. However, due to the time variation in β_s and γ_s, we have to run regressions using data from a moving window, and then aggregate the spot estimates to obtain estimates for $I\beta$ and IdV. On the other hand, the estimation of the jump idiosyncratic risk IdJ requires estimation of various jump times and sizes, which does not involve any regressions or moving windows.

3.1. Identification

The key intuition for the identification of the spot β and γ is from the classical regression framework. First, we can identify the spot (or instantaneous) β by comparing the spot "variance" of X and its spot "covariance" with Y, where we use the orthogonality of X and Z in (5). For any s in $[0, t]$,

$$\frac{d}{ds}[Y_s^c, X_s^c] = \beta_s^\intercal \frac{d}{ds}[X_s^c, X_s^c]. \tag{10}$$

As usual in modeling of high-frequency data, our "variance" and "covariance" are measured by the quadratic covariations.

In addition, we can decompose the spot total variation of Y^c into the variation explained by the factors and the residual variation:

$$\frac{d}{ds}[Y_s^c, Y_s^c] = \beta_s^\intercal \frac{d}{ds}[X_s^c, X_s^c]\beta_s + \frac{d}{ds}[Z_s^c, Z_s^c] = \beta_s^\intercal c_s \beta_s + \gamma_s^2. \tag{11}$$

Therefore, we can estimate γ_s and β_s using data from a short window that contains s. Once these are estimated for each moving window, we can aggregate them to obtain an estimator of $I\beta$ and IdV.

The following identity allows for the identification of IdJ,

$$\sum_{s\le t}(\Delta Z_s)^2 = \sum_{s\le t}(\Delta Y_s)^2 \cdot 1_{\{\Delta X_s = 0\}}. \tag{12}$$

Eq. (12) follows from the identification assumption in (6), which implies that X and Z do not jump together. Hence, IdJ can be estimated by combining the estimates of Y jump times and sizes with the estimates of X jump times.

Notice that our spot β and γ are identified from the continuous part of the quadratic variations, so that the specification of the jump part does not affect the identification or estimation of β and γ. Conversely, the specification of the continuous components does not affect the identification or estimation of IdJ. Note also that in this high-frequency setting, the drift components of Y and X do not enter any of the above identification arguments, as they are of smaller order compared to the Brownian components and jumps.

3.2. Estimation

We denote the distance between adjacent observations by Δ_n. Let $\Delta_i^n A = A_{i\Delta_n} - A_{(i-1)\Delta_n}$, for $1 \le i \le [t/\Delta_n]$ and any process A. To implement the time-localized regressions, we form a sequence of moving windows of k_n observations, so that the length of the interval covered by each window is $k_n\Delta_n$.

We collect truncated returns of Y and X sampled within the ith window into a vector $\mathcal{Y}_i$ and a matrix $\mathcal{X}_i$,

$$\mathcal{Y}_i = \begin{pmatrix} (\Delta_{i+1}^n Y)1_{\left\{|\Delta_{i+1}^n Y| \le u_n\right\}} \\ (\Delta_{i+2}^n Y)1_{\left\{|\Delta_{i+2}^n Y| \le u_n\right\}} \\ \vdots \\ (\Delta_{i+k_n}^n Y)1_{\left\{|\Delta_{i+k_n}^n Y| \le u_n\right\}} \end{pmatrix}_{k_n \times 1} \quad \text{and} \quad \mathcal{X}_i = \begin{pmatrix} \left(\Delta_{i+1}^n X\right)^\intercal 1_{\left\{\|\Delta_{i+1}^n X\| \le u_n\right\}} \\ \left(\Delta_{i+2}^n X\right)^\intercal 1_{\left\{\|\Delta_{i+2}^n X\| \le u_n\right\}} \\ \vdots \\ \left(\Delta_{i+k_n}^n X\right)^\intercal 1_{\left\{\|\Delta_{i+k_n}^n X\| \le u_n\right\}} \end{pmatrix}_{k_n \times d},$$

where u_n is the threshold selected to remove jumps. For simplicity, we use the same notation u_n for the truncation threshold of both X and Y, but they can differ as long as they are chosen to satisfy the restrictions below.

Within the ith window, we first estimate the spot covariance matrix of X with the truncated realized variance over a short window, that is,

$$\widehat{c}_{i\Delta_n} = \frac{1}{k_n\Delta_n}(\mathcal{X}_i^\intercal \mathcal{X}_i) = \frac{1}{k_n\Delta_n}\sum_{j=1}^{k_n}\left(\Delta_{i+j}^n X\right)\left(\Delta_{i+j}^n X\right)^\intercal 1_{\left\{\|\Delta_{i+j}^n X\| \le u_n\right\}}. \tag{13}$$

This estimator is standard and taken directly from, e.g., Jacod and Protter (2012). Next, we construct an OLS estimator of $\widehat{\beta}_{i\Delta_n}$ based on Eq. (10),

$$\widehat{\beta}_{i\Delta_n} = (\mathcal{X}_i^\intercal \mathcal{X}_i)^{-1}\mathcal{X}_i^\intercal \mathcal{Y}_i. \tag{14}$$

Next, building on Eqs. (10) and (11), we obtain

$$\widehat{\gamma}^2_{i\Delta_n} = \frac{1}{k_n\Delta_n}\mathcal{Y}_i^\intercal \mathcal{Y}_i - \widehat{\beta}_{i\Delta_n}^\intercal \widehat{c}_{i\Delta_n}\widehat{\beta}_{i\Delta_n} = \frac{1}{k_n\Delta_n}\mathcal{Y}_i^\intercal\left(\mathrm{Id}_{k_n} - \mathcal{X}_i(\mathcal{X}_i^\intercal\mathcal{X}_i)^{-1}\mathcal{X}_i^\intercal\right)\mathcal{Y}_i,$$

where Id_{k_n} is a $k_n \times k_n$ identity matrix. These local estimates resemble closely the OLS regression estimates of the slope and the residual variance, except we use truncation to remove the discontinuous component.

We can now construct the following estimators for the Integrated Beta and the Idiosyncratic Volatility:

$$\widehat{I\beta}_t = \frac{1}{t}\sum_{i=0}^{[t/k_n\Delta_n]-1}\widehat{\beta}_{ik_n\Delta_n}k_n\Delta_n \quad \text{and} \quad \widehat{IdV}_t^{\text{naive}} = \frac{1}{t}\sum_{i=0}^{[t/k_n\Delta_n]-1}\widehat{\gamma}^2_{ik_n\Delta_n}k_n\Delta_n. \tag{15}$$

Although these estimators are constructed similarly, they have distinct asymptotic behavior. While the $\widehat{IdV}_t^{\text{naive}}$ estimator is consistent, it has a bias of larger order than $\Delta_n^{-1/2}$, which prevents the construction of confidence intervals:

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

$$k_n\left(\widehat{IdV}_t^{naive} - IdV_t\right) \xrightarrow{p} -\frac{d}{t}\int_0^t \gamma_s^2 ds.$$

Hence, we propose the following bias-corrected estimator for IdV,

$$\widehat{IdV}_t = \frac{1}{t}\left(1 + \frac{d}{k_n}\right)\sum_{i=0}^{[t/k_n\Delta_n]-1}\widehat{\gamma}^2_{ik_n\Delta_n}k_n\Delta_n. \tag{16}$$

Based on the identification identity in Eq. (12), we estimate IdJ as follows,

$$\widehat{IdJ}_t = \frac{1}{t}\sum_{i=1}^{[t/\Delta_n]}(\Delta_i^n Y)^2 \cdot 1_{\{\|\Delta_i^n X\| \le u_n,\, |\Delta_i^n Y| > u_n\}}. \tag{17}$$

The above estimator uses the Y returns in those time intervals where Z is estimated to jump; an interval is likely to contain a Z jump if Y return is relatively large (i.e., larger than u_n to be specified later), while at the same time each of the X returns is relatively small.

Hence, we can estimate the coefficient of determination in (7),

$$\widehat{R}_t^2 = 1 - \frac{\widehat{IdV}_t + \widehat{IdJ}_t}{RV(Y)_t}, \tag{18}$$

where $RV(Y)_t$ is the Realized Volatility of Y, $RV(Y)_t = \frac{1}{t}\sum_{i=1}^{[t/\Delta_n]}(\Delta_i^n Y)^2$.

We remark that the estimators in (15) and (16) use non-overlapping subsamples.[2] The same is true for the asymptotic variance estimators introduced in Section 4. Corresponding estimators based on overlapping subsamples can be constructed; one can show that they share the same asymptotic distribution as the non-overlapping estimators. However, the overlapping estimators exhibit finite-sample biases due to the end-of-sample effects. The latter biases are automatically avoided by the non-overlapping estimators.

We can make an analogy here between our framework and the classical multiple regression setting. Indeed, within each local window, our estimator resembles the standard estimator of the error variance, a result of the local linear structure of the stochastic processes. Due to the data being sampled from a fixed window of size T, our in-fill asymptotic setting allows for identification of realizations instead of expectations, which requires a substantial span of time. As a result, our setting allows for considerable dependence and heterogeneity in the underlying processes. This is in sharp contrast to the long-span or "large T" asymptotics in the standard time series regression, which typically imposes high-level conditions to limit dependence and heterogeneity.

4. The asymptotic distributions

This section provides the asymptotic distributions of our estimators, which are useful for constructing the confidence intervals.

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

$$\Delta_n^{-1/2}\left(\widehat{I\beta}_t - I\beta_t\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}_t^{\beta} \quad \text{and} \quad \Delta_n^{-1/2}\left(\widehat{IdV}_t - IdV_t\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}_t^{\gamma},$$

where $\mathcal{W}^{\beta}$ and $\mathcal{W}^{\gamma}$ are processes defined on the extension of the original space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$, which, conditionally on $\mathcal{F}$, are centered Gaussian with the covariance matrix and variance given by

$$\mathbb{E}(\mathcal{W}_t^{\beta}\mathcal{W}_t^{\beta\intercal}|\mathcal{F}) = \frac{1}{t^2}\int_0^t \gamma_s^2 c_s^{-1} ds \quad \text{and} \quad \mathbb{E}((\mathcal{W}_t^{\gamma})^2|\mathcal{F}) = \frac{2}{t^2}\int_0^t \gamma_s^4 ds.$$

The asymptotic variances can be estimated as follows,

$$\sum_{i=0}^{[t/k_n\Delta_n]-1} \widehat{\gamma}^2_{ik_n\Delta_n}(\widehat{c}_{ik_n\Delta_n})^{-1} k_n\Delta_n \xrightarrow{p} \int_0^t \gamma_s^2 c_s^{-1} ds, \quad \text{and} \quad 2\sum_{i=0}^{[t/k_n\Delta_n]-1} \left(\widehat{\gamma}^2_{ik_n\Delta_n}\right)^2 k_n\Delta_n \xrightarrow{p} 2\int_0^t \gamma_s^4 ds.$$

To state the next theorem, we introduce some additional notation. Let $\{T_q\}_{q=1,2,\ldots}$ be the sequence of the jump times of Z. Let η_t denote the spot volatility process of Y. Denote by κ_q a sequence of independent uniform random variables on $[0, 1]$, for $q = 1, 2, \ldots$, defined on the extension of the original space. Finally, denote by Ψ_{q-} and $\Psi_{q+}(q = 1, 2, \ldots)$ a sequence of independent standard normal random variables, also defined on the extension of the original space.

Theorem 3. *Suppose Assumptions 1–3 hold. Let $u_n \asymp \Delta_n^{\varpi}$ and $1/(4-2r) \leq \varpi < 1/2$. Then as $\Delta_n \to 0$, we have*

$$\Delta_n^{-1/2}\left(\widehat{IdJ}_t - IdJ\right) \xrightarrow{\mathcal{L}-s} \mathcal{W}_t^{J}$$

where

$$\mathcal{W}_t^{J} = \frac{2}{t}\sum_{q\geq 1: T_q\leq t} \Delta Y_{T_q}\left(\sqrt{\kappa_q}\eta_{T_q-}\Psi_{q-} + \sqrt{1-\kappa_q}\eta_{T_q}\Psi_{q+}\right).$$

Furthermore, if $(Y, X^{\intercal})^{\intercal}$ does not co-jump with its spot volatility process, then $\mathcal{W}_t^{J}$, conditionally on $\mathcal{F}$, is centered Gaussian with variance given by

$$\mathbb{E}((\mathcal{W}_t^{J})^2|\mathcal{F}) = \frac{4}{t^2}\sum_{q\geq 1: T_q\leq t}\left(\Delta Y_{T_q}\right)^2 \eta^2_{T_q}.$$

The above asymptotic variance can be estimated consistently by

$$\frac{4}{t^2}\sum_{i=0}^{[t/k_n\Delta_n]-1} \widehat{\eta}^2_{ik_n\Delta_n}\sum_{j=1}^{k_n}(\Delta^n_{ik_n+j}Y)^2 \cdot 1_{\left\{\|\Delta^n_{ik_n+j}X\|\leq u_n,\, |\Delta^n_{ik_n+j}Y|>u_n\right\}},$$

where $\widehat{\eta}^2_{i\Delta_n} = \frac{1}{k_n\Delta_n}(\mathcal{Y}_i^{\intercal}\mathcal{Y}_i)$, provided that $k_n \to \infty$ and $k_n\Delta_n \to 0$ as $\Delta_n \to 0$.

[2] Of course, $\widehat{IdJ}_t$ in (17) does not use subsamples at all.

[3] Here we use $\xrightarrow{\mathcal{L}-s}$ to denote stable convergence in law.

If $(Y, X^\intercal)^\intercal$ does co-jump with its spot volatility process, the limiting process $\mathcal{W}_t^J$ is not Gaussian conditionally on $\mathcal{F}$, and an asymptotic variance estimator is not helpful. In this case, the limiting distribution can be obtained by Monte Carlo simulation, see, e.g., Section 10.2.4 of Aït-Sahalia and Jacod (2014), who study inference for quadratic variation of a scalar process.

Remark 1. Inference on the total idiosyncratic risk, $IdV_t + IdJ_t$, follows from the above two theorems, plus the fact that the two standardized estimators are asymptotically independent. The independence of $\mathcal{W}_t^\gamma$ in Theorem 2 and $\mathcal{W}_t^J$ in Theorem 3 can be shown by a mix of the two proofs, for example, following the arguments of the proof of Theorem 5.4.2 of Jacod and Protter (2012).

5. Monte Carlo simulations

We now examine the finite sample performance of our estimators. For this purpose, we simulate many trajectories over one month for one stock with log-price Y_t following a three-factor model, which is a special case of (1).

The factors $X_t = (X_{1t}, X_{2t}, X_{3t})'$ follow the dynamics

$$\begin{pmatrix} dX_{1t} \\ dX_{2t} \\ dX_{3t} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} dt + \begin{pmatrix} \sigma_{1t} & 0 & 0 \\ 0 & \sigma_{2t} & 0 \\ 0 & 0 & \sigma_{3t} \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \begin{pmatrix} dW_{1t} \\ dW_{2t} \\ dW_{3t} \end{pmatrix} + \begin{pmatrix} J_{1t} \\ J_{2t} \\ J_{3t} \end{pmatrix} dN_t.$$

The factor volatilities are driven by Feller's square root process (a.k.a the Cox–Ingersoll–Ross model) with the same process N_t,

$$d\sigma_{it}^2 = \widetilde{\kappa}_i \left(\widetilde{\alpha}_i - \sigma_{it}^2\right) dt + \widetilde{\upsilon}_i \sigma_{it} d\widetilde{W}_{it} + \widetilde{J}_{it} dN_t, \quad i = 1, 2, 3.$$

For each $i = 1, 2, 3$, the size of jumps in X_i, J_{it}, has a double exponential distribution:

$$J_{it} \sim \begin{cases} \exp(g_i^+) & \text{with probability } q_i \\ - \exp(g_i^-) & \text{with probability } 1 - q_i \end{cases},$$

where q_i is the mixture probability. The size of each volatility jump, $\widetilde{J}_{it}$, is exponentially distributed with mean $\widetilde{g}_i$ for each $i = 1, 2, 3$. N_t is a Poisson process with λ jumps per year on average. The λ parameter is chosen so that on average, the discontinuous quadratic variation of Y equals the continuous quadratic variation.

The stock idiosyncratic returns are generated by the following jump-diffusion,

$$dZ_t = h_t dt + \gamma_t d\overline{W}_t + \bar{J}_t d\overline{N}_t,$$

where the idiosyncratic jump size $\bar{J}_t$ is another double exponential:

$$\bar{J}_t \sim \begin{cases} \exp(\overline{g}^+) & \text{with probability } \overline{q} \\ - \exp(\overline{g}^-) & \text{with probability } 1 - \overline{q} \end{cases}.$$

In the above, $\overline{N}_t$ is another Poisson process with the expected number of jumps per year also equal to λ, and is independent of N_t.

We simulate the betas as follows,

$$d\beta_{it} = \kappa_i (\alpha_i - \beta_{it}) dt + \upsilon_i dB_{it}, \quad i = 1, 2, 3.$$

In the above, W_t, $\widetilde{W}_t$, $\overline{W}_t$, and B_t are mutually independent standard Brownian motions. The values of all the above parameters are listed in Table 1. We simulate 1000 paths of this process over one month.

Table 2 presents the sample bias and the sample standard deviation of the estimation error, as well as the sample RMSE. Three different choices of the subsample size and three different sampling frequencies are considered (5, 10, and 30 min). The left panel corresponds to the subsample sizes used in the empirical application. When interpreting these results, it is useful to notice that in our model, the true betas and idiosyncratic jumps are random and vary across simulations. We set the truncation parameter as $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$, which is a standard choice in the literature, see, for example, Aït-Sahalia and Xiu (2019b).[4] Δ_n is the distance between the observations, and BV_t is the annualized bipower variation of the corresponding day. As a robustness check, Table 3 presents the same results with $u_n = 4\Delta_n^{0.47}\sqrt{BV_t}$. The simulation results support our theoretical consistency results: all estimators become more precise as more observations are available.

Fig. 1 shows the finite sample distributions of our standardized estimators, as well as the standardized naive IdV estimator. For comparison, Fig. 1 also superimposes the asymptotic distribution. We can see that the finite sample distributions of our estimators are close to their asymptotic distributions, and the bias correction for IdV estimation is effective.

[4] An alternative would be to use a theoretically optimal threshold recently proposed by Figueroa-López and Mancini (2019) for the estimation of univariate integrated variance.

Table 1
Monte Carlo simulations: parameter values.

Factor dynamics		Beta dynamics		Idiosyncratic dynamics	
b	(0.05, 0.03, 0.02)	κ	(2, 2, 2)	γ	0.35
σ_0^2	(0.12, 0.09, 0.04)	α	(0.15, 0.10, −0.10)	h	0
$\widetilde{\kappa}$	(3, 4, 5)	υ	(0.03, 0.03, 0.03)	$\overline{g}^+$	$\gamma \times 14\sqrt{\Delta_n}$
$\widetilde{g}$	(0.004, 0.005, 0.004)			$\overline{g}^-$	$\gamma \times 14\sqrt{\Delta_n}$
g_i^+	$\sigma_{i0} \times 7\sqrt{\Delta_n}$			$\overline{q}$	0.5
g_i^-	$\sigma_{i0} \times 7\sqrt{\Delta_n}$				
λ	67				
$\widetilde{\alpha}$	(0.09, 0.04, 0.06)				
$\widetilde{\upsilon}$	(0.3, 0.4, 0.3)				
q	(0.5, 0.5, 0.5)				
$(\rho_{12}, \rho_{13}, \rho_{23})$	(0.05, 0.10, 0.15)				

Note: This table reports the parameter values used in the data generating process of our Monte Carlo simulations.

Table 2
Simulation results.

5 min sampling frequency									
	$k_n = 78$			$k_n = 91$			$k_n = 117$		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
$\widehat{I\beta}_t^{(1)}$	−0.15	2.64	2.65	−0.16	2.64	2.65	−0.15	2.65	2.65
$\widehat{I\beta}_t^{(2)}$	−0.05	3.13	3.13	−0.05	3.13	3.13	−0.05	3.11	3.11
$\widehat{I\beta}_t^{(3)}$	0.26	4.18	4.19	0.31	4.16	4.17	0.24	4.14	4.14
$\widehat{IdV}_t$	0.06	0.46	0.47	0.07	0.46	0.47	0.08	0.46	0.47
$\widehat{Id\overline{J}}_t$	−0.06	0.78	0.78	−0.06	0.78	0.78	−0.06	0.78	0.78
10 min sampling frequency									
	$k_n = 39$			$k_n = 63$			$k_n = 91$		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
$\widehat{I\beta}_t^{(1)}$	−0.10	3.83	3.83	−0.14	3.75	3.75	−0.14	3.67	3.67
$\widehat{I\beta}_t^{(2)}$	0.06	4.67	4.67	0.02	4.47	4.47	0.03	4.49	4.49
$\widehat{I\beta}_t^{(3)}$	−0.17	6.46	6.46	−0.20	6.23	6.23	−0.18	6.25	6.26
$\widehat{IdV}_t$	0.03	0.67	0.67	0.07	0.66	0.66	0.09	0.65	0.66
$\widehat{Id\overline{J}}_t$	0.00	0.98	0.98	0.00	0.98	0.98	0.00	0.98	0.98
30 min sampling frequency									
	$k_n = 13$			$k_n = 21$			$k_n = 39$		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
$\widehat{I\beta}_t^{(1)}$	−0.17	7.56	7.56	0.01	6.97	6.97	0.04	6.47	6.47
$\widehat{I\beta}_t^{(2)}$	0.07	9.67	9.68	0.34	8.91	8.92	0.25	8.37	8.38
$\widehat{I\beta}_t^{(3)}$	0.18	13.02	13.02	−0.04	12.21	12.21	0.07	11.39	11.39
$\widehat{IdV}_t$	−0.60	1.22	1.36	−0.17	1.20	1.21	0.02	1.20	1.20
$\widehat{Id\overline{J}}_t$	−0.03	1.88	1.88	−0.03	1.88	1.88	−0.03	1.88	1.88

Note: Columns "Bias", "Stdev", and "RMSE" contain the sample Bias, the sample standard deviation, and the sample root mean squared error across simulations of the centered estimator, e.g., of the quantity $\widehat{IdV}_t - IdV_t$. All these quantities are multiplied by 100. k_n is the number of observations in a subsample. $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$.

6. Empirical results

6.1. Construction of intraday equity factors

We reconstruct the five Fama–French factors (see Fama and French, 1993 and Fama and French, 2015) and the momentum factor (see Carhart, 1997) at the 5-min frequency from January 1, 1996 to December 31, 2017. The construction takes a few steps and requires a combination of three databases. We describe the details below.

Since the daily portfolio constituents for the five Fama–French factors are not publicly available, we start by replicating these factors at the daily frequency to obtain the constituents (the website of Kenneth French contains the factors but not their constituents). We download daily adjusted returns and firm fundamentals from the merged Center for Research

Table 3
Simulation Results.

5 min sampling frequency									
	k_n= 78			k_n= 91			k_n= 117		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
$\widehat{I\beta}_t^{(1)}$	−0.14	2.63	2.64	−0.16	2.63	2.64	−0.15	2.64	2.64
$\widehat{I\beta}_t^{(2)}$	−0.03	3.15	3.15	−0.04	3.16	3.16	−0.02	3.13	3.13
$\widehat{I\beta}_t^{(3)}$	0.26	4.13	4.14	0.31	4.13	4.14	0.25	4.10	4.10
$\widehat{Id\mathbb{V}}_t$	0.20	0.50	0.54	0.20	0.50	0.54	0.21	0.50	0.54
$\widehat{Id\mathbb{J}}_t$	−0.13	0.77	0.78	−0.13	0.77	0.78	−0.13	0.77	0.78
10 min sampling frequency									
	k_n= 39			k_n= 63			k_n= 91		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
$\widehat{I\beta}_t^{(1)}$	−0.06	3.89	3.89	−0.08	3.82	3.82	−0.08	3.72	3.72
$\widehat{I\beta}_t^{(2)}$	0.06	4.62	4.62	0.01	4.44	4.44	0.02	4.47	4.47
$\widehat{I\beta}_t^{(3)}$	−0.21	6.51	6.51	−0.23	6.31	6.31	−0.23	6.31	6.31
$\widehat{Id\mathbb{V}}_t$	0.17	0.74	0.76	0.22	0.73	0.76	0.24	0.73	0.77
$\widehat{Id\mathbb{J}}_t$	−0.18	1.02	1.03	−0.18	1.02	1.03	−0.18	1.02	1.03
30 min sampling frequency									
	k_n= 13			k_n= 21			k_n= 39		
	Bias	Stdev	RMSE	Bias	Stdev	RMSE	Bias	Stdev	RMSE
$\widehat{I\beta}_t^{(1)}$	−0.13	7.59	7.59	0.06	6.97	6.97	0.07	6.46	6.46
$\widehat{I\beta}_t^{(2)}$	0.03	9.89	9.89	0.30	8.90	8.90	0.16	8.32	8.33
$\widehat{I\beta}_t^{(3)}$	0.15	12.84	12.84	−0.08	12.05	12.05	0.01	11.25	11.25
$\widehat{Id\mathbb{V}}_t$	−0.24	1.74	1.75	0.20	1.75	1.76	0.40	1.76	1.80
$\widehat{Id\mathbb{J}}_t$	−0.41	2.27	2.30	−0.41	2.27	2.30	−0.41	2.27	2.30

Note: Columns "bias", "stdev", and "RMSE" contain the sample bias, the sample standard deviation, and the sample root mean squared error across simulations of the centered estimator, e.g., of the quantity $\widehat{Id\mathbb{V}}_t - Id\mathbb{V}_t$. All these quantities are multiplied by 100. k_n is the number of observations in a subsample. $u_n = 4\Delta_n^{0.47}\sqrt{\mathbb{BV}_t}$.

in Security Prices (CRSP) database and Compustat database. We restrict the universe of stocks to those listed on NYSE, NASDAQ, and AMEX, and apply the same stock filters as those imposed by Fama and French (1993) based on data availability and concerns of sampling bias or survival bias. We then build six value-weighted portfolios by sorting on market equity (ME) and the ratio of book equity to market equity (BE/ME), where BE/ME for June of year t is the book equity for the last fiscal year end in $t-1$ divided by ME for December of $t-1$. Small (S) or big (B) portfolios are classified by ME according to their comparison with the median size of NYSE, whereas the book-to-market groups are divided based on the breakpoints for the bottom 30 percent (L), middle 40 percent (M), and top 30 percent (H) of ranked values of BE/ME. The value and size factors (spread) are given by:

$$HML = \frac{1}{2}(SH + BH) - \frac{1}{2}(SL + BL), \quad and \quad SMB = \frac{1}{3}(SH + SM + SL) - \frac{1}{3}(BH + BM + BL).$$

Similarly, for the profitability and investment factors introduced in Fama and French (2015), we form six portfolios each, based on the intersections of two size portfolios (S and B) and three profitability (OP) portfolios, denoted by robust (R), neutral (N), and weak (W), respectively, as well as three investment (Inv) portfolios, denoted by conservative (C), neutral (N), and aggressive (A), respectively. OP for June of year t is annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses divided by the book equity for the last fiscal year end in $t-1$. Inv is the change in total assets from the fiscal year ending in year $t-2$ to the fiscal year ending in $t-1$, divided by $t-2$ total assets. The OP and Inv breakpoints are the 30th and 70th NYSE percentiles. The profitability and investment factors are given by:

$$RMW = \frac{1}{2}(SR + BR) - \frac{1}{2}(SW + BW), \quad and \quad CMA = \frac{1}{2}(SC + BC) - \frac{1}{2}(SA + BA).$$

To construct the momentum factor, we form six value-weighted portfolios based on ME and prior (2–12) months' returns, following the procedure outlined on Ken French's website. We use up (U), flat (F), and down (D) to denote momentum categories based on the 30th and 70th NYSE percentiles of prior (2–12) months' returns. The final *MOM* factor is the

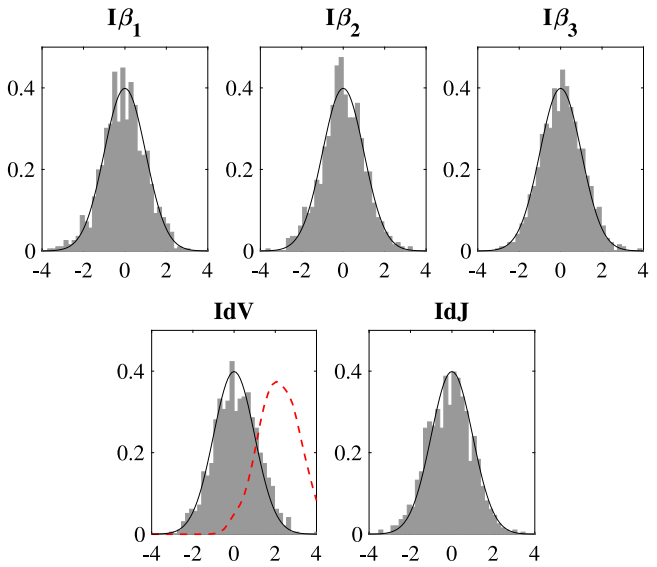

Fig. 1. Simulation results: Standardized estimates. Note: We plot the finite sample distributions of the standardized statistics (gray histograms), and we superimpose the standard normal law (black solid line). The finite sample distribution of the $\widehat{IdV}_t^{naive}$ estimator appears in red dashed lines. The sampling frequency is every 5 min and the subsamples are one day long, i.e., $k_n = 78$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

average return on the two high prior return portfolios minus the average return on the two low prior return portfolios, given by:

$$MOM = \frac{1}{2}(SU + BU) - \frac{1}{2}(SD + BD).$$

Finally, the market factor (MKT) is simply the return of a value-weighted portfolio of all stocks in this pool.

The replication at daily frequency provides daily portfolio constituents and their weights in these $6\times4 = 24$ component portfolios. Using these weights, we can obtain buy-and-hold portfolio returns at intraday intervals using transaction-level data on individual stocks. Note that all returns are simple returns, and that the portfolio weights are also adjusted at an intraday frequency (because high-frequency price changes lead to high-frequency market capitalization changes). In particular, on day s and at time interval i, the return of any of these portfolios is given by $PRet_{s,i}$:

$$PRet_{s,i} = \frac{\sum_{j=1}^{N_s} w_{s,i}^j \cdot Ret_{s,i}^j}{\sum_{j=1}^{N_s} w_{s,i}^j},$$

where N_s denotes the total number of stocks for this portfolio on day s, j denotes the jth stock of the portfolio, and $w_{s,i}^j$ is given by

$$w_{s,i}^j = w_s^j \cdot \prod_{l=0}^{i-1}(1 + Ret_{s,l}^j),$$

with w_s^j being the market capitalization based on the close price of stock j on day $s-1$. The return at time 0 on day s denotes the overnight return from day $s-1$ to s.

We collect trades and quotes data of all stocks from the New York Stock Exchange Trade and Quotes (NYSE TAQ) Database via Wharton Research Data Services (WRDS). The TAQ data require substantial cleaning due to idiosyncrasies of the market microstructure. We follow the cleaning procedure detailed in Da and Xiu (2017). First, we remove trades and quotes with condition codes Z, B, U, T, L, G, W, K, J, and the corresponding odd lot trades, which have an additional letter I, as well as those with non-empty suffix codes (preferred shares). Next, we identify the opening trades as those with condition codes O, Q, OI, or QI, closing trades with 6, M, 6I, or MI, and remove all trades beyond the window of opening and closing time points. We only keep trades with correction indicator 00 or 01. Then, we construct the national best bid and offer (NBBO) data using quotes from all exchanges at a 1-s frequency. We match trades with NBBOs by their recorded time points, and eliminate those trades that are outside the range of the corresponding NBBOs.

There are several obstacles when merging the cleaned TAQ data with the CRSP data. We use individual stock's CUSIP instead of ticker information, because CRSP tickers differ from those recorded in the TAQ data. There exists a small portion of firm-day pairs, for which CUSIPs are missing from the TAQ database. We handpick their CRSP tickers based on available

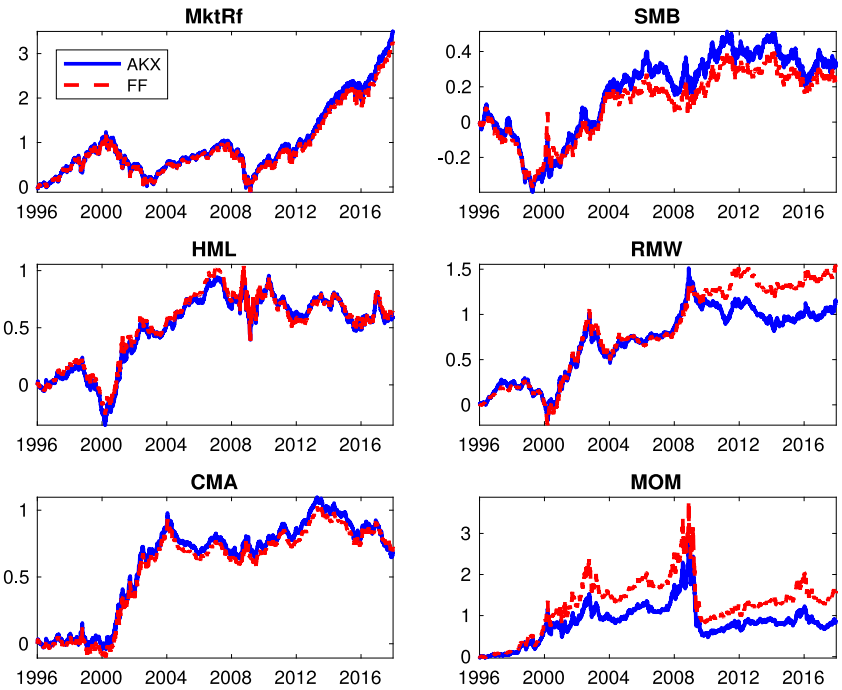

Fig. 2. Factors. Note: This figure compares the daily aggregated cumulative returns of the high-frequency factors, in blue solid lines (AKX), with those of the low-frequency factors, in red dashed lines (FF), downloaded from Kenneth French's website. The sampling period is January 1996–December 2017.

information from adjacent days. In addition, firms change their TAQ tickers from time to time due to mergers, acquisitions, or other reasons, so we instead use CRSP PERMINO's to index all stocks, which do not change over time. Finally, the intraday prices from TAQ are unadjusted for dividends or splits that occur overnight. To obtain the adjusted overnight returns, we use CRSP (unadjusted) open and close prices to obtain intraday open-to-close returns. Along with adjusted (close-to-close) returns from CRSP we can infer the (close-to-open) adjusted overnight returns. Because CRSP open and close prices are more carefully selected based on additional information other than the sequence of trades, we use them as the open and close prices for the TAQ data. This also ensures a perfect match between daily returns from CRSP and daily returns that aggregate intraday returns from the TAQ data.

In Fig. 2, we compare the daily cumulative returns of HML, SMB, RMW, CMA, MktRf, and MOM factors based on our high-frequency replication, with the daily returns downloaded from Kenneth French's website (MktRf denotes MKT in excess of the one-month T-bill rate). The levels of high- and low-frequency cumulative returns match almost perfectly for MktRf, HML, and CMA, though there are moderate differences for SMB, and sizeable differences for RMW and MOM. For RMW, the largest discrepancy appears to occur right after 2008, while for MOM it occurs around 2000. These two differences aside, the patterns of their cumulative returns remain alike shortly afterwards. In fact, the corresponding correlations of daily returns are very high: 0.9998 (MktRf), 0.9921 (SMB), 0.9932 (HML), 0.9566 (RMW), 0.9983 (CMA), and 0.9852 (MOM).

6.2. The choice of sampling frequency for the stock returns

Before running any high-frequency regressions for individual stocks, it is important to determine for each stock in each month a frequency at which the impact of the microstructure noise is negligible. We adopt the following procedure. We preselect a grid of four frequencies: 5-min, 10-min, 30-min, and daily, and assign each stock-month pair one of these frequencies. For a given stock-month pair, we assign the highest frequency, at which the prices satisfy the following two criteria. First, there should be at least 90% non-zero returns. Second, the hypothesis of no noise cannot be rejected. We select the daily frequency if at least one criterion fails at each of the three higher frequencies. To test the null hypothesis of no noise, we employ one of the equivalent Hausman tests proposed by Aït-Sahalia and Xiu (2019a) (the one based on first-order autocorrelations, T_n, in Section 3.1), which is more rigorous and convenient than reading off a frequency visually from a volatility signature plot. Despite running many statistical tests, we do not adjust for multiple testing because in our setting, a false positive leads to a more conservative choice of the sampling frequency. The 90% rule is arbitrary and rather conservative. It ensures a minimum number of non-zero returns for the Hausman test, which in turn ensures the test has reasonable power (zero returns are deleted before the application of the test).

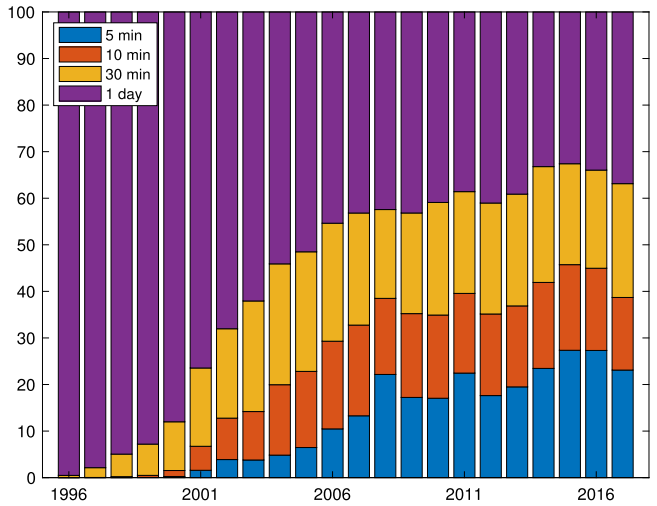

Fig. 3. Percentage of the selected sampling frequencies. Note: This figure compares the percentages of the selected sampling frequencies for combinations of stocks and months in each year from 1996 to 2017.

Fig. 3 plots the time series of the percentages of the selected sampling frequencies across all stock-month pairs each year. Around the beginning of our sample in 1996, there are less than 1% of all stock-month pairs using 30-min returns. No stock-month pairs meet the criteria for higher frequencies. In the following years, there is a remarkable trend indicating increased reliability of higher sampling frequencies based on our criteria. In 2017, over 20% of stock-months use 5-min returns, while the use of daily frequency drops to less than 40%.

6.3. A case study of IBM

We then estimate the factor model using high-frequency returns of each stock in each month from 1996 to 2017. We consider a six-factor model with Fama–French five factors plus the momentum factor.[5] The threshold is $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$ (see Section 5 for details); the block length is one day (except for the daily frequency assignment case, when it is a month). We also calculate the corresponding low-frequency benchmarks. In this section, as an example, we analyze the IBM stock in detail.

Fig. 4 plots the betas of IBM. There are several notable findings. First, the high-frequency betas are remarkably more stable than their low-frequency counterparts. The latter display much more variation from month to month.[6] Second, there are quite a few large outliers in the low-frequency estimates. For example, momentum beta has a large negative estimate with a magnitude greater than 10 in April 2003. Close inspection of these outliers shows that in all these scenarios, IBM has unusually bad earnings announcements that miss analysts' expectations. In contrast, the high-frequency beta estimates do not have such outliers because these large returns are identified as jumps, which do not contribute to the calculation of betas. Third, in the earlier period of the sample, the high-frequency estimates are noisier because of the use of low sampling frequencies selected by our procedure. Fig. 5 shows a time series of the selected frequencies. A relatively small number of months at the beginning of the sample use daily frequency, while starting with the decimalization in 2001, the selected frequency for IBM is almost always 5 min.

Fig. 6 compares the total risk of IBM with its idiosyncratic risk, which is further decomposed into a continuous component and a discontinuous component. The average percentage contribution of the idiosyncratic risk is 47.3%, so the six-factor model explains roughly half of the total variance of IBM returns. The idiosyncratic jump component accounts for, on average, 11.6% of the total idiosyncratic risk, although the percentage can reach as high as 75% occasionally. For example, three recent spikes in the idiosyncratic jump component occur in April 2014, July 2015, and October 2015. We

[5] We use high-frequency market return (MKT) directly instead of its excess return (MktRf).

[6] High- and low-frequency beta estimators can differ due to finite sample issues, but depending on the model, they can also have different population counterparts. For example, in our model, under the infill asymptotic scheme, the probability limit, say β^{LF}, of the standard (low-frequency) beta estimator is

$$\beta^{LF} = ([X_t, X_t])^{-1}[X_t, Y_t] = \left(\int_0^t c_s ds + \sum_{0\le s\le t} \Delta X_s \Delta X_s^\intercal\right)^{-1}\left(\int_0^t c_s \beta_s ds + \sum_{0\le s\le t} \Delta X_s \Delta X_s^\intercal \tilde{\beta}_s\right). \tag{19}$$

In the special case when β_t and $\tilde{\beta}_t$ are equal and time-invariant, we obtain $\beta^{LF} = \beta$, which also equals the integrated beta in (9). If $\beta \neq \tilde{\beta}$ and/or any of the two betas is time-varying, β^{LF} depends on all the quantities in (19).

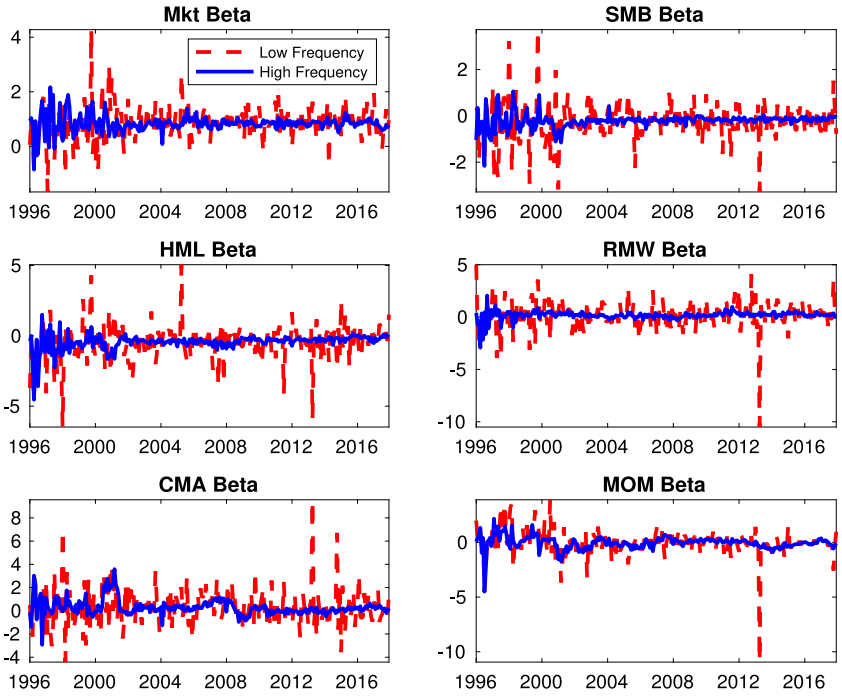

Fig. 4. The six-factor betas of the IBM stock. Note: This figure compares the time series of the high-frequency betas (blue solid line) with the low-frequency betas (red dashed line) for IBM estimated from the six-factor model for each month from 1996 to 2017. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

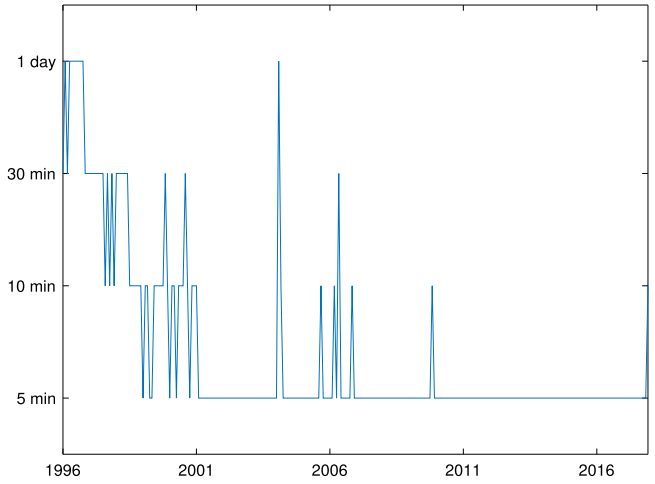

Fig. 5. Time series of the selected sampling frequencies for IBM. Note: This figure plots the monthly time series of the selected sampling frequencies for IBM from 1996 to 2017.

trace back the exact times of the jumps in high-frequency returns causing the large values of the monthly idiosyncratic jump measure. It turns out these jumps occur overnight following IBM's quarterly report release dates, and are therefore largely caused by IBM's earnings disappointments.

6.4. The cross section of all stocks

The current section presents the summary across all stocks of the estimation results.

First, we provide cross-sectional quantiles of the beta estimates of all stocks in Figs. 7–12. As expected, the median estimates are close to one for the market betas and zero for HML, RMW, CMA, and MOM betas. For the size beta, the

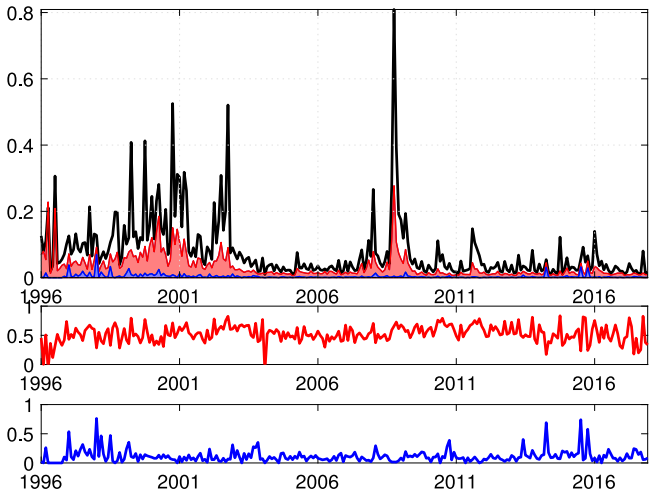

Fig. 6. Decomposition of the IBM's realized volatility. Note: The **top panel** compares the total risk of IBM with its total idiosyncratic risk, which is then further decomposed into the continuous component and the jump component. The total risk and the total idiosyncratic risk are measured by the annualized quadratic variations of IBM returns (thick black line) and its residual returns (thin red line), respectively. The jump component of the total idiosyncratic risk is shaded (in blue) at the bottom of the plot. The **middle panel** plots the fraction of total variation explained by the systematic component (R^2), whereas the **bottom panel** plots the fraction of the variation of the total idiosyncratic risk composed of idiosyncratic jumps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

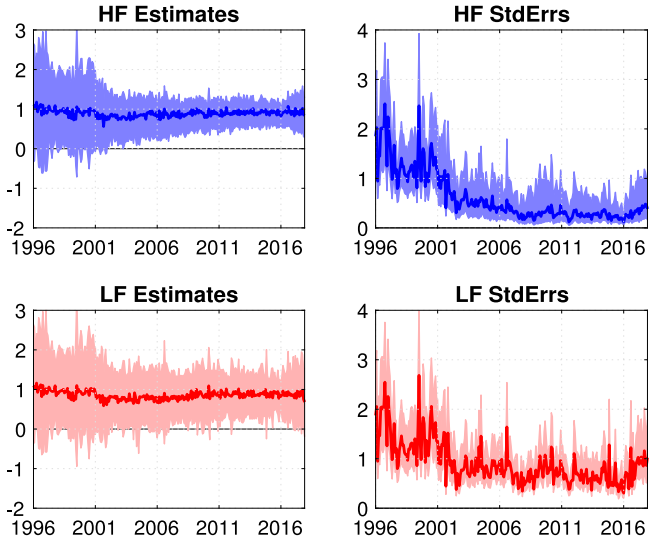

Fig. 7. Market beta. Note: In this figure, we compare the high-frequency and the low-frequency (i.e., daily frequency) estimates of the MKT beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

median is somewhat positive. This is because the size factor is value weighted, and the "median firm" is a small firm that loads positively on the SMB factor (see Fig. 9).

We find that both high- and low-frequency beta estimates demonstrate similar cross-sectional dispersion in the earlier part of the sample. However, the dispersion of high-frequency estimates is smaller towards the end of the sample. As the liquidity improves over time, more stocks can use higher frequency data, which improves precision and also helps to identify and remove jumps. Therefore, the relative advantages of high-frequency betas over the low-frequency alternative we demonstrate in the special case of IBM seem likely to hold in general. (When making these comparisons, one should keep in mind that the high- and low-frequency beta estimands only coincide under additional assumptions, see footnote 6.)

In addition, we plot in Fig. 13 the cross-sectional quantiles of the annualized idiosyncratic risk. For the high-frequency data, we obtain the total idiosyncratic risk as the sum of the idiosyncratic volatility and the idiosyncratic jumps. For the low-frequency data, we use the variance of the regression residuals as a measure of the idiosyncratic risk, which is

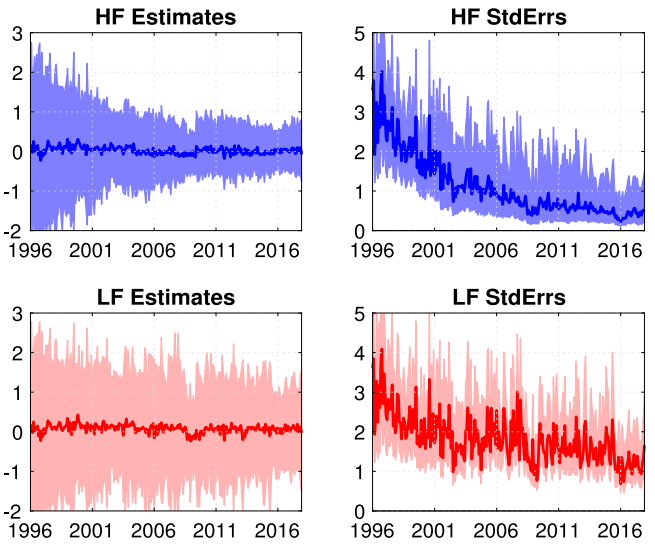

Fig. 8. HML beta. Note: In this figure, we compare the high-frequency and the low-frequency estimates of the HML beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

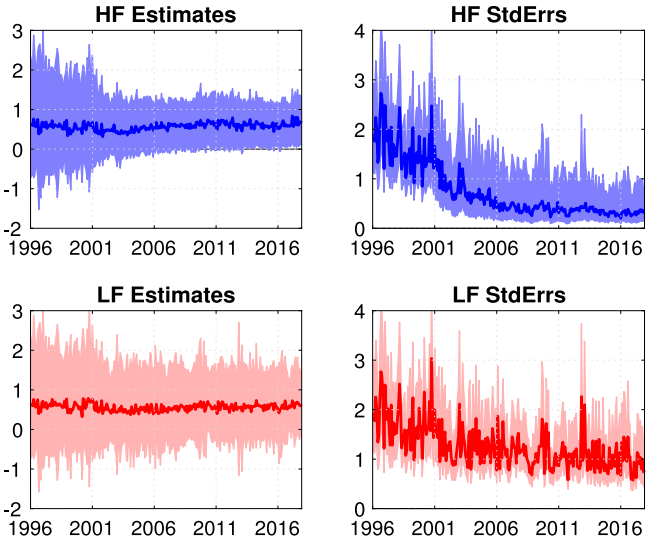

Fig. 9. SMB beta. Note: In this figure, we compare the high-frequency and the low-frequency estimates of the SMB beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

the standard estimator of the idiosyncratic risk in the empirical finance literature.[7] Fig. 13 suggests that the estimates are roughly comparable and have similar cross-sectional dispersion. Again, high-frequency estimates are relatively more accurate in the later sample (see Fig. 10).

Fig. 14 provides further details on the idiosyncratic jump component for the high-frequency data. The figure considers a smaller cross-section that excludes the stock-months with the daily frequency assignment, when it is difficult to identify the jumps. The top left plot shows that idiosyncratic jumps account for, on average, 10% of the total idiosyncratic risk, though the number increases to about 20% for roughly 25% of firms. In the top right panel, we report the same quantiles for idiosyncratic risks in the CAPM model. Not surprisingly, idiosyncratic jumps account for a slightly larger fraction of the total risk in the CAPM model since there is only one factor that can potentially jump. In addition, the bottom two plots

[7] In our model, when β_t and $\tilde{\beta}_t$ are equal and time-invariant, under the infill asymptotic scheme, the probability limit of this (annualized) standard idiosyncratic variance estimator is the same as for the estimators that we calculate on the high-frequency data.

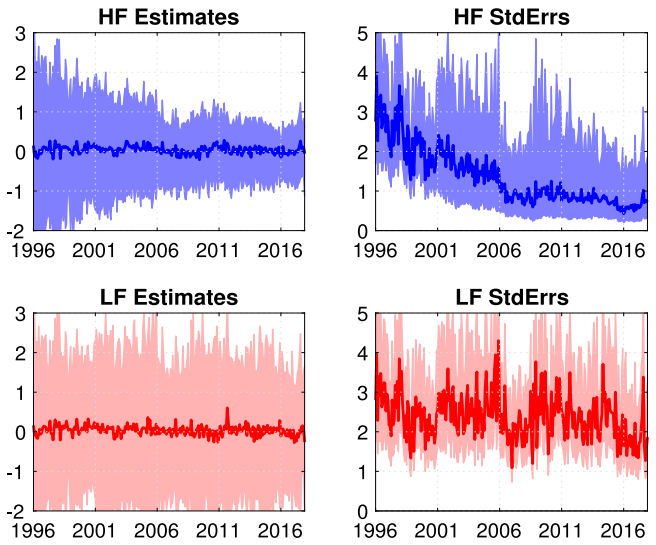

Fig. 10. CMA beta. Note: In this figure, we compare the high-frequency and the low-frequency estimates of the CMA beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

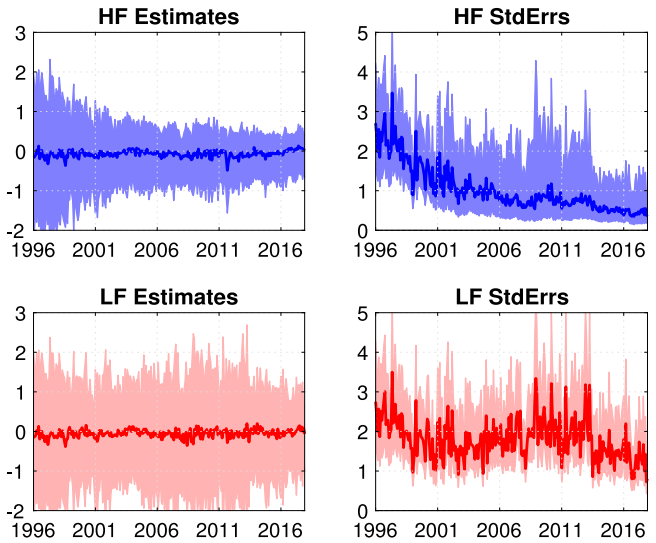

Fig. 11. RMW beta. Note: In this figure, we compare the high-frequency and the low-frequency estimates of the RMW beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

use a higher threshold $u_n = 4\Delta_n^{0.47}\sqrt{BV_t}$ (the default is $3\Delta_n^{0.47}\sqrt{BV_t}$), which produce lower idiosyncratic jumps ratios, as expected (see Fig. 11).

Fig. 15 plots the cross-sectional quantiles of R^2, the fraction of the total risk (quadratic variation) explained by the systematic component, in the six-factor model and the CAPM. The comparison clearly shows that the additional five factors are helpful in explaining a larger fraction of the time-series variation in stock returns compared to the CAPM. When we exclude the stock-month pairs using daily frequencies (the bottom graphs), R^2 increases in the earlier part of the sample, which is otherwise dominated by the daily frequency regressions with somewhat lower R^2.

Finally, we explore the relationship between the idiosyncratic jumps and earnings surprises in the cross-section of stocks. We define an earnings surprise as the percentage difference between realized earnings and analysts' expectations, which are available from the I/B/E/S dataset via WRDS. There are a few earnings measures available. We use the most prominent one, namely, earnings per share (EPS). Because I/B/E/S dataset is quarterly, when merging this database with our monthly high-frequency estimates, we only keep those months when the announcements are made. There are 248,756

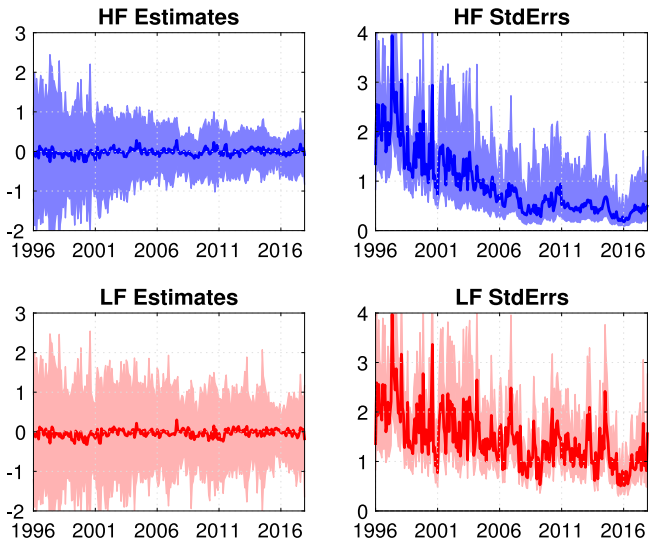

Fig. 12. MOM beta. Note: In this figure, we compare the high-frequency and the low-frequency estimates of the MOM beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

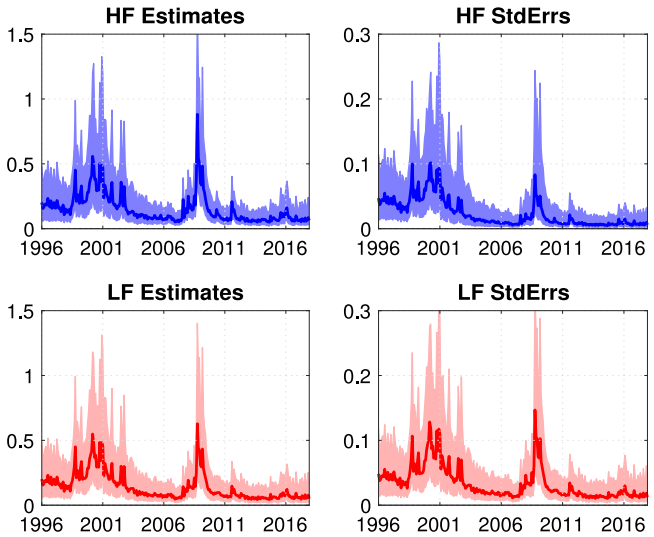

Fig. 13. Total idiosyncratic risk. Note: In this figure, we compare the high-frequency estimates of the total idiosyncratic risk (including both the continuous and jump components) with the low-frequency estimates of the idiosyncratic volatility in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

matched stock-month pairs. Among these, we only keep the stock-months with a positive idiosyncratic jump estimate. This leads to a total of 117,557 stock-month pairs which we use in the following regressions.

Because the dataset is rather heterogeneous with potentially many outliers, we regress the logarithm of idiosyncratic jumps onto the logarithm of the absolute earnings surprises. We consider several specifications of panel regressions with and without firm and month fixed effects. Table 4 provides the regression results for two different truncation thresholds. We see from Columns (a)–(d) that earnings surprises increase idiosyncratic jumps, a result that holds across several fixed effects specifications. Moreover, Columns (e)–(h) further suggest that earnings disappointments have a larger impact than positive earnings surprises, which appears to agree with the evidence in the case of IBM.

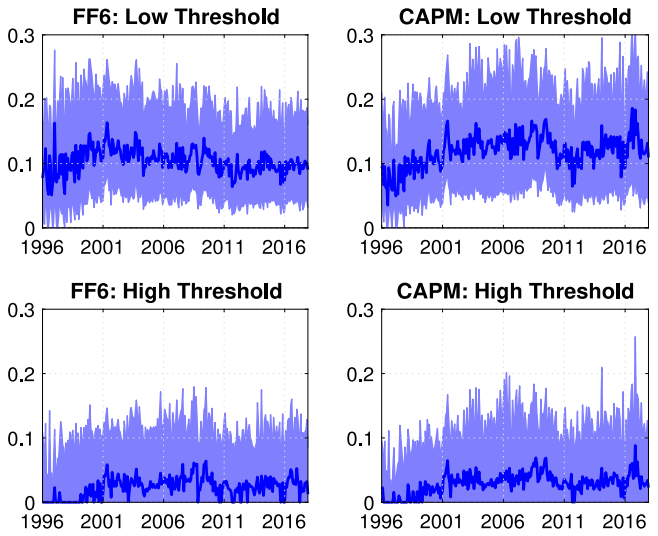

Fig. 14. Percentages of IdJ. Note: In this figure, we compare the ratios of the idiosyncratic jump component over the total idiosyncratic risk for the six-factor model and CAPM. The "low threshold" is $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$, whereas the "high threshold" is $u_n = 4\Delta_n^{0.47}\sqrt{BV_t}$. All plots of this figure use the smaller cross section, which excludes the stock-months for which the daily sampling frequency is selected. The solid blue line corresponds to the cross-sectional median; the bands are 25- and 75-percentiles of the considered cross section.

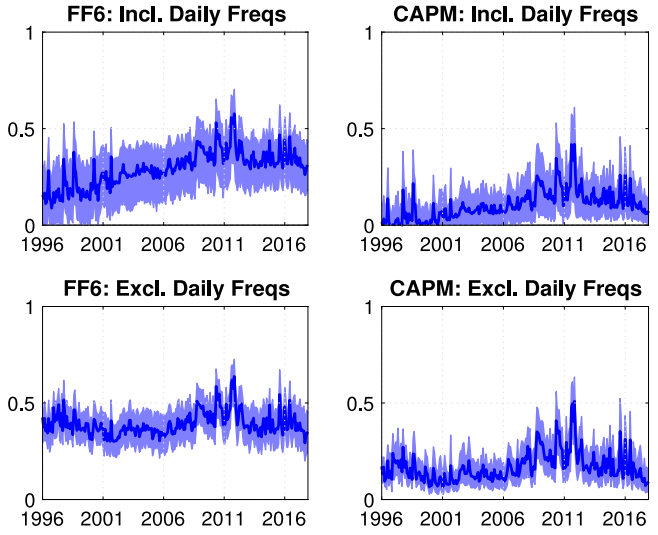

Fig. 15. Percentages of R^2. Note: In this figure, we compare the fraction of the variation in total risk explained by the systematic component (R^2) for the six-factor model and CAPM, using high-frequency data. The lower panel is based on a smaller cross section that excludes all stock-months for which the daily sampling frequency is selected for estimation. The solid blue line corresponds to the cross-sectional median; the bands are 25- and 75-percentiles of the considered cross section.

7. Conclusion

This paper shows how to identify and estimate, using high-frequency data, a nonparametric Fama–French factor model under broad assumptions on the data-generating process. We allow for general time-variation in the factor beta processes, which makes our framework particularly suitable to application to individual stocks. The definitions of the estimators are straightforward, but important technical difficulties associated with the classification of jumps into systematic and idiosyncratic components arise when deriving a limit theory.

Our empirical analysis uses a dataset that is larger than any dataset in the high-frequency literature. First, we use all traded stocks from NYSE, AMEX, and NASDAQ stock exchanges to reconstruct the five Fama–French factors and the momentum factor for 1996–2017 at the 5-min frequency. The construction requires merging of three databases: NYSE TAQ, CRSP, and Compustat. We provide the details of the construction and explain how the main practical

Table 4
Idiosyncratic jumps and earnings surprises.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
	Low threshold							
Intercept	−3.79 (−141.95)	−3.89 (−225.89)	−3.82 (−148.81)	−3.88 (−258.24)	−3.57 (−130.06)	−3.73 (−170.14)	−3.68 (−138.44)	−3.80 (−185.88)
log(surprise)	0.14 (18.57)	0.10 (17.86)	0.13 (17.95)	0.11 (21.09)	0.18 (21.77)	0.13 (17.44)	0.15 (19.29)	0.11 (16.56)
Positive					−0.45 (−15.45)	−0.31 (−12.78)	−0.27 (−10.64)	−0.15 (−6.62)
log(surprise) × positive					−0.09 (−10.39)	−0.05 (−7.15)	−0.05 (−6.42)	−0.01 (−2.06)
Firm effect		✓		✓		✓		✓
Month effect			✓	✓			✓	✓
R-Squared (%)	1.42	1.73	1.40	1.84	1.49	1.96	1.48	1.89
No. observations	117, 557	117, 557	117, 557	117, 557	117, 557	117, 557	117, 557	117, 557
	High threshold							
Intercept	−3.84 (−132.79)	−3.86 (−188.08)	−3.87 (−140.08)	−3.87 (−198.17)	−3.62 (−115.53)	−3.70 (−139.86)	−3.74 (−123.24)	−3.79 (−147.40)
log(surprise)	0.14 (17.05)	0.13 (18.78)	0.13 (15.88)	0.13 (19.06)	0.17 (18.46)	0.15 (16.81)	0.14 (15.54)	0.12 (14.56)
Positive					−0.43 (−12.84)	−0.30 (−9.98)	−0.25 (−8.23)	−0.14 (−4.98)
log(surprise) × positive					−0.08 (−7.81)	−0.04 (−4.19)	−0.04 (−3.90)	0.00 (0.14)
Firm effect		✓		✓		✓		✓
Month effect			✓	✓			✓	✓
R-squared (%)	1.31	1.61	1.32	1.72	1.32	1.75	1.31	1.69
No. observations	82, 210	82, 210	82, 210	82, 210	82, 210	82, 210	82, 210	82, 210

Note: This table reports 8 panel regression specifications. The dependent variable is always the logarithm of the idiosyncratic jumps. The covariates include the logarithm of the earnings surprise, defined as the absolute difference between the realized and expected earnings per share. " Positive" is a dummy variable that indicates whether the earnings surprise is positive or negative. We also include firm and month fixed effects in some of these regressions. t-statistics are provided in parentheses. The "low threshold" is $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$, whereas the "high threshold" is $u_n = 4\Delta_n^{0.47}\sqrt{BV_t}$.

challenges can be resolved. The resulting number of stocks is 5005 on average. We show how to automatically select an appropriate sampling frequency for each stock-month to account for the wide variation in liquidity across stocks and time. We document the key empirical properties across the stocks and the new factors, and apply the nonparametric factor model with the new high-frequency Fama–French factors. We find that this model is effective in explaining the systematic component of the risk of individual stocks. Finally, we decompose the idiosyncratic risk into the idiosyncratic volatility and the idiosyncratic jumps, and find that earnings surprises increase idiosyncratic jumps. Moreover, earnings disappointments have a larger effect on the idiosyncratic jumps than do the positive earnings surprises.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.01.007.

References

Aït-Sahalia, Y., Fan, J., Peng, H., 2009. Nonparametric transition-based tests for jump-diffusions. J. Amer. Statist. Assoc. 104, 1102–1116.
Aït-Sahalia, Y., Jacod, J., 2014. High Frequency Financial Econometrics. Princeton University Press.
Aït-Sahalia, Y., Xiu, D., 2019a. A Hausman test for the presence of market microstructure noise in high frequency data. J. Econometrics 211, 176–205.
Aït-Sahalia, Y., Xiu, D., 2019b. Principal component analysis of high frequency data. J. Amer. Statist. Assoc. 114, 287–303.
Andersen, T.G., Bollerslev, T., Diebold, F.X., Wu, J., 2005. A framework for exploring the macroeconomic determinants of systematic risk. Amer. Econ. Rev. 95, 398–404.
Ang, A., Kristensen, D., 2012. Testing conditional factor models. J. Financ. Econ. 106, 132–156.
Bandi, F.M., Phillips, P.C.B., 2003. Fully nonparametric estimation of scalar diffusion models. Econometrica 71, 241–283.
Barndorff-Nielsen, O.E., Shephard, N., 2004. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. Econometrica 72 (3), 885–925.
Bollerslev, T., Li, S.Z., Todorov, V., 2016. Roughing up beta: Continuous versus discontinuous betas and the cross section of expected stock returns. J. Financ. Econ. 120, 464–490.
Box, G.E.P., Tiao, G.C., 1968. A Bayesian approach to some outlier problems. Biometrika 55, 119–129.

Carhart, M.M., 1997. On persistence in mutual fund performance. J. Finance 52 (1), 57–82.
Chang, I., Tiao, G.C., Chen, C., 1988. Estimation of time series parameters in the presence of outliers. Technometrics 30, 193–204.
Da, R., Xiu, D., 2017. When Moving-Average Models Meet High-Frequency Data: Uniform Inference on Volatility. Tech. Rep., University of Chicago.
Fama, E.F., French, K.R., 1992. The cross-section of expected stock returns. J. Finance 47, 427–465.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.
Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.
Figueroa-López, J., Mancini, C., 2019. Optimum thresholding using mean and conditional mean squared error. J. Econometrics 208, 179–210.
Jacod, J., Protter, P., 2012. Discretization of Processes. Springer-Verlag.
Jacod, J., Rosenbaum, M., 2013. Quarticity and other functionals of volatility: Efficient estimation. Ann. Statist. 41, 1462–1484.
Kalnina, I., 2015. Inference for Nonparametric High-Frequency Estimators with an Application to Time Variation in Betas. Tech. Rep., University of Montreal.
Kalnina, I., Tewou, K., 2017. Cross-Sectional Dependence in Idiosyncratic Volatility. Tech. Rep., North Carolina State University.
Kalnina, I., Xiu, D., 2017. Nonparametric estimation of the leverage effect: A trade-off between robustness and efficiency. J. Amer. Statist. Assoc. 112, 384–396.
Li, J., Todorov, V., Tauchen, G., 2016. Inference theory on volatility functional dependencies. J. Econometrics 193, 17–34.
Li, J., Todorov, V., Tauchen, G., 2017. Jump regressions. Econometrica 85, 173–195.
Mykland, P.A., Zhang, L., 2006. ANOVA for diffusions and Itô processes. Ann. Statist. 34, 1931–1963.
Reiß, M., Todorov, V., Tauchen, G.E., 2015. Nonparametric test for a constant beta between Itô semi-martingales based on high-frequency data. Stochastic Process. Appl. 125 (8), 2955–2988.
Todorov, V., Bollerslev, T., 2010. Jumps and betas: A new framework for disentangling and estimating systematic risks. J. Econometrics 157, 220–235.

Appendix A Proofs

For convenience, let $(A)_{ij}$ denote the (i,j)th element of a matrix A, and $\mathrm{Tr}(A)$ denote its trace. In addition, we define e_i to be the $d \times 1$ unit vector with the i-th element equal to 1. Also, we define $\delta_{\{\text{statement}\}}$ as follows:

$$\delta_{\{\text{statement}\}} = \begin{cases} 1, & \text{if the statement is true} \\ 0, & \text{otherwise} \end{cases}.$$

Throughout, we denote by K a generic constant, which may change from line to line.

Appendix A.1 Proof of Theorem 1

We consider the vector $\mathcal{U} = (Y, X^\intercal)^\intercal$. Note that $\mathcal{U}$ is an Itô semimartingale, and its spot covariance matrix can be written as

$$u_s = \begin{pmatrix} \beta_s^\intercal c_s \beta_s + \gamma_s^2 & \beta_s^\intercal c_s \\ c_s \beta_s & c_s \end{pmatrix}.$$

Naturally, it can be estimated by:

$$\widehat{u}_{i\Delta_n} = \frac{1}{k_n \Delta_n} \sum_{j=1}^{k_n} (\Delta_{i+j}^n \mathcal{U})^\intercal (\Delta_{i+j}^n \mathcal{U}) 1_{\{||\Delta_{i+j}^n U|| \le u_n\}} = \frac{1}{k_n \Delta_n} \begin{pmatrix} \mathcal{Y}_i^\intercal \mathcal{Y}_i & \mathcal{Y}_i^\intercal \mathcal{X}_i \\ \mathcal{X}_i^\intercal \mathcal{Y}_i & \mathcal{X}_i^\intercal \mathcal{X}_i \end{pmatrix} := \begin{pmatrix} \widehat{u}_{i\Delta_n}^{11} & \widehat{u}_{i\Delta_n}^{12} \\ \widehat{u}_{i\Delta_n}^{21} & \widehat{u}_{i\Delta_n}^{22} \end{pmatrix}.$$

For any $(d+1)$-dimensional positive semidefinite matrix u, let

$$h(u) = (u^{22})^{-1} u^{21}, \quad \text{and} \quad g(u) = u^{11} - u^{12}(u^{22})^{-1} u^{21}, \tag{A.1}$$

where u^{22}, u^{21}, u^{21}, and u^{11} partition u as follows,

$$u = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix},$$

where u_{11} is of dimension 1×1, u_{12} is $1\times d$, u_{21} is $d\times 1$, and u_{22} is $d\times d$. With these definitions, we have

$$h(u_s)=\beta_s,\quad \text{and}\quad g(u_s)=\beta_s^\intercal c_s\beta_s+\gamma_s^2-\beta_s^\intercal c_s(c_s)^{-1}c_s\beta_s=\gamma_s^2.$$

As a result, our estimator can be represented as:

$$\widehat{I\beta}_t=\frac{k_n\Delta_n}{t}\sum_{i=0}^{[t/k_n\Delta_n]-1}h(\widehat{u}_{ik_n\Delta_n}),\quad \text{and}\quad \widehat{IdV}_t^{naive}=\frac{k_n\Delta_n}{t}\sum_{i=0}^{[t/k_n\Delta_n]-1}g(\widehat{u}_{ik_n\Delta_n}).$$

By the same argument as in the proof of Proposition 1 in Aït-Sahalia and Xiu (2019b), we obtain

$$k_n(\widehat{IdV}^{naive}-IdV)\xrightarrow{p}\frac{1}{2t}\int_0^t\sum_{j,k,l,m=1}^{d+1}\partial^2_{jk,lm}g(u_s)\Big((u_s)_{jl}(u_s)_{km}+(u_s)_{jm}(u_s)_{kl}\Big)ds.$$

For $1\le j,k,l,m\le d+1$, by taking derivatives in (A.1), we obtain:

$$\begin{aligned}
\partial_{jk}g(u_s)=&\delta_{\{j=1,k=1\}}-\delta_{\{j=1,k\ge 2\}}e_{k-1}^\intercal(u_s^{22})^{-1}u_s^{21}-u_s^{12}(u_s^{22})^{-1}\delta_{\{j\ge 2,k=1\}}e_{j-1}\\
&+u_s^{12}(u_s^{22})^{-1}\delta_{\{j\ge 2,k\ge 2\}}e_{j-1}e_{k-1}^\intercal(u_s^{22})^{-1}u_s^{21},\\
\partial^2_{jk,lm}g(u_s)=&+\delta_{\{j=1,k\ge 2,l\ge 2,m\ge 2\}}e_{k-1}^\intercal(u_s^{22})^{-1}e_{l-1}e_{m-1}^\intercal(u_s^{22})^{-1}u_s^{21}\\
&+\delta_{\{j\ge 2,k\ge 2,l=1,m\ge 2\}}e_{m-1}^\intercal(u_s^{22})^{-1}e_{j-1}e_{k-1}^\intercal(u_s^{22})^{-1}u_s^{21}\\
&-\delta_{\{j=1,k\ge 2,l\ge 2,m=1\}}e_{k-1}^\intercal(u_s^{22})^{-1}e_{l-1}\\
&-\delta_{\{j\ge 2,k=1,l=1,m\ge 2\}}e_{m-1}^\intercal(u_s^{22})^{-1}e_{j-1}\\
&-\delta_{\{j\ge 2,k\ge 2,l\ge 2,m\ge 2\}}(u_s^{12})(u_s^{22})^{-1}e_{l-1}e_{m-1}^\intercal(u_s^{22})^{-1}e_{j-1}e_{k-1}^\intercal(u_s^{22})^{-1}u_s^{21}\\
&-\delta_{\{j\ge 2,k\ge 2,l\ge 2,m\ge 2\}}(u_s^{12})(u_s^{22})^{-1}e_{j-1}e_{k-1}^\intercal(u_s^{22})^{-1}e_{l-1}e_{m-1}^\intercal(u_s^{22})^{-1}u_s^{21}\\
&+\delta_{\{j\ge 2,k\ge 2,l\ge 2,m=1\}}u_s^{12}(u_s^{22})^{-1}e_{j-1}e_{k-1}^\intercal(u_s^{22})^{-1}e_{l-1}\\
&+\delta_{\{j\ge 2,k=1,l\ge 2,m\ge 2\}}u_s^{12}(u_s^{22})^{-1}e_{l-1}e_{m-1}^\intercal(u_s^{22})^{-1}e_{j-1}.
\end{aligned}$$

Plugging it in, and by direct calculations, we have

$$\sum_{j,k,l,m=1}^{d+1}\partial^2_{jk,lm}g(u_s)\Big((u_s)_{jl}(u_s)_{km}+(u_s)_{jm}(u_s)_{kl}\Big)=2d\Big(u_s^{12}(u_s^{22})^{-1}u_s^{21}-u_s^{11}\Big)=-2d\gamma_s^2,$$

which concludes the proof.

Appendix A.2 Proof of Theorem 2

The proof needs a similar result to Theorem 2 in Aït-Sahalia and Xiu (2019b) (see also Theorem 3.2 of Jacod and Rosenbaum (2013) under slightly different conditions), which can be established following similar steps:

$$\Delta_n^{-1/2}\left\{k_n\Delta_n\sum_{i=0}^{[t/k_n\Delta_n]-1}\left(f(\widehat{u}_{ik_n\Delta_n})-\frac{1}{2k_n}\sum_{j,k,l,m=1}^{d+1}\partial^2_{jk,lm}f(\widehat{u}_{ik_n\Delta_n})\Big((\widehat{u}_{ik_n\Delta_n})_{jl}(\widehat{u}_{ik_n\Delta_n})_{km}\right.\right.$$
$$\left.\left.+(\widehat{u}_{ik_n\Delta_n})_{jm}(\widehat{u}_{ik_n\Delta_n})_{kl}\Big)\right)-\int_0^t f(u_s)ds\right\}\xrightarrow{\mathcal{L}-s}V_t,$$

where V_t is a conditionally Gaussian process with variance given by

$$\mathbb{E}(V_tV_t^{\intercal}|\mathcal{F})=\sum_{j,k,l,m=1}^{d+1}\int_0^t\partial_{jk}f(u_s)(\partial_{lm}f(u_s))^{\intercal}\Big((u_s)_{jl}(u_s)_{km}+(u_s)_{jm}(u_s)_{kl}\Big)ds,$$

and where f is any C^3 vector-valued function that satisfies certain polynomial growth condition.

Appendix A.2.1 Estimation of IdV

The asymptotic bias has been corrected as in the previous theorem. To calculate the asymptotic variance, we introduce

$$\partial g(u):=(\partial_{ij}g(u))_{(d+1)\times(d+1)}=\begin{pmatrix}1 & -u^{12}(u^{22})^{-1}\\ -(u^{22})^{-1}u^{21} & (u^{22})^{-1}u^{21}u^{12}(u^{22})^{-1}\end{pmatrix}.$$

Therefore, noting that $u_s=u_s^{\intercal}$ and $\partial g(u_s)=(\partial g(u_s))^{\intercal}$, we obtain

$$\mathbb{E}(V_t^2|\mathcal{F})=2\int_0^t\sum_{j,m=1}^{d+1}(\partial g(u_s)u_s)_{jm}(u_s\partial g(u_s))_{jm}ds=2\int_0^t\mathrm{Tr}\left((\partial g(u_s)u_s)^2\right)ds=2\int_0^t\gamma_s^4ds.$$

Appendix A.2.2 Estimation of $I\beta$

We start with the asymptotic bias term, for which we need to calculate the derivative of h defined in (A.1). For $1 \leq j, k, l, m \leq d+1$, we obtain:

$$\begin{aligned}
\partial_{jk} h(u_s) =& -\delta_{\{j\geq 2, k\geq 2\}} (u_s^{22})^{-1} e_{j-1} e_{k-1}^\intercal (u_s^{22})^{-1} u_s^{21} + \delta_{\{j\geq 2, k=1\}} (u_s^{22})^{-1} e_{j-1}, \\
\partial^2_{jk,lm} h(u_s) =& +\delta_{\{j\geq 2, k\geq 2, l\geq 2, m\geq 2\}} (u_s^{22})^{-1} e_{l-1} e_{m-1}^\intercal (u_s^{22})^{-1} e_{j-1} e_{k-1}^\intercal (u_s^{22})^{-1} u_s^{21} \\
&+\delta_{\{j\geq 2, k\geq 2, l\geq 2, m\geq 2\}} (u_s^{22})^{-1} e_{j-1} e_{k-1}^\intercal (u_s^{22})^{-1} e_{l-1} e_{m-1}^\intercal (u_s^{22})^{-1} u_s^{21} \\
&-\delta_{\{j\geq 2, k\geq 2, l\geq 2, m=1\}} (u_s^{22})^{-1} e_{j-1} e_{k-1}^\intercal (u_s^{22})^{-1} e_{l-1} \\
&-\delta_{\{j\geq 2, k=1, l\geq 2, m\geq 2\}} (u_s^{22})^{-1} e_{l-1} e_{m-1}^\intercal (u_s^{22})^{-1} e_{j-1}.
\end{aligned}$$

In addition, we notice that

$$\begin{aligned}
&\sum_{j,k,l,m=1}^{d+1} \delta_{\{j\geq 2, k\geq 2, l\geq 2, m\geq 2\}} (u_s^{22})^{-1} e_{l-1} e_{m-1}^\intercal (u_s^{22})^{-1} e_{j-1} e_{k-1}^\intercal (u_s^{22})^{-1} u_s^{21} \Big((u_s)_{jl} (u_s)_{km} + (u_s)_{jm} (u_s)_{kl} \Big) \\
=& \sum_{j,k,l,m=1}^{d} (u_s^{22})^{-1} e_l e_m^\intercal (u_s^{22})^{-1} e_j e_k^\intercal (u_s^{22})^{-1} u_s^{21} \Big((u_s^{22})_{jl} (u_s^{22})_{km} + (u_s^{22})_{jm} (u_s^{22})_{kl} \Big) \\
=& (d+1)(u_s^{22})^{-1} u_s^{21}, \qquad \text{(A.2)} \\
&\sum_{j,k,l,m=1}^{d+1} \delta_{\{j\geq 2, k\geq 2, l\geq 2, m=1\}} (u_s^{22})^{-1} e_{j-1} e_{k-1}^\intercal (u_s^{22})^{-1} e_{l-1} \Big((u_s)_{jl} (u_s)_{km} + (u_s)_{jm} (u_s)_{kl} \Big) \\
=& \sum_{j,k,l=1}^{d} (u_s^{22})^{-1} e_j e_k^\intercal (u_s^{22})^{-1} e_l \Big((u_s^{22})_{jl} (u_s^{21})_k + (u_s^{21})_j (u_s^{22})_{kl} \Big) \\
=& (d+1)(u_s^{22})^{-1} u_s^{21}. \qquad \text{(A.3)}
\end{aligned}$$

By switching m and k, j and l, we get the same equalities based on the remaining two terms of $\partial^2_{jk,lm} h(u_s)$, therefore

$$\sum_{j,k,l,m=1}^{d+1} \partial^2_{jk,lm} h(u_s) \Big((u_s)_{jl} (u_s)_{km} + (u_s)_{jm} (u_s)_{kl} \Big) = 0.$$

Hence, there is no asymptotic bias for $\widehat{I\beta}_t$. As to the asymptotic covariance, we have

$$
\begin{aligned}
&\mathbb{E}(V_t V_t^{\intercal}|\mathcal{F}) \\
&= \sum_{j,k,l,m=1}^{d+1} \int_0^t \partial_{jk} h(u_s)(\partial_{lm} h(u_s))^{\top} \Big((u_s)_{jl}(u_s)_{km} + (u_s)_{jm}(u_s)_{kl} \Big) ds, \\
&= \sum_{j,k,l,m=1}^{d+1} \int_0^t \Big(- \delta_{\{j\geq 2, k\geq 2\}} (u_s^{22})^{-1} e_{j-1} e_{k-1}^{\top} (u_s^{22})^{-1} u_s^{21} + \delta_{\{j\geq 2, k=1\}} (u_s^{22})^{-1} e_{j-1} \Big) \\
&\quad \Big(- \delta_{\{l\geq 2, m\geq 2\}} (u_s^{22})^{-1} e_{l-1} e_{m-1}^{\top} (u_s^{22})^{-1} u_s^{21} + \delta_{\{l\geq 2, m=1\}} (u_s^{22})^{-1} e_{l-1} \Big)^{\top} \Big((u_s)_{jl}(u_s)_{km} + (u_s)_{jm}(u_s)_{kl} \Big) ds \\
&= \int_0^t \Big\{ (u_s^{12})(u_s^{22})^{-1}(u_s^{21})(u_s^{22})^{-1} + 2(u_s^{22})^{-1}(u_s^{21})(u_s^{12})(u_s^{22})^{-1} + (u_s^{22})^{-1} u_s^{11} - (u_s^{12})(u_s^{22})^{-1}(u_s^{21})(u_s^{22})^{-1} \\
&\quad - 2(u_s^{22})^{-1}(u_s^{21})(u_s^{12})(u_s^{22})^{-1} - (u_s^{12})(u_s^{22})^{-1}(u_s^{21})(u_s^{22})^{-1} \Big\} ds \\
&= \int_0^t \gamma_s^2 c_s^{-1} ds.
\end{aligned}
$$

The asymptotic variance estimator can be justified by applying Theorem 9.4.1 from Jacod and Protter (2012). This concludes the proof.

Appendix A.3 Proof of Theorem 3

Define an auxiliary process $Y_t' = Y_0 + \int_0^t \beta_{s-}^{\intercal} dX_s^c + Z_t$. Let $A_i^n = \{\|\Delta_i^n X^c\| \leq u_n\}$ and $\Omega_t^n = \cap_{i\leq [t/\Delta_n]} A_i^n$. Note that by (2.1.33) and (2.1.34) of Jacod and Protter (2012), for any $q \geq 2$, $\mathbb{E}\|\Delta_i^n X^c\|^q \leq K\Delta_n^{q/2}$. By localization and Markov's inequality, we obtain

$$
\sum_{i=1}^{[t/\Delta_n]} \mathbb{P}\left(\|\Delta_i^n X^c\| > u_n\right) \leq u_n^{-q} \sum_{i=1}^{[t/\Delta_n]} \mathbb{E}\|\Delta_i^n X^c\|^q \leq K\Delta_n^{q/2-1-q\varpi} \to 0.
$$

The last step in the above holds because $0 < \varpi < 1/2$ ensures $q > 2/(1-2\varpi)$. This implies $\mathbb{P}(\Omega_t^n) \to 1$. On the set Ω_t^n, we have

$$
\sum_{i=1}^{[t/\Delta_n]} (\Delta_i^n Y')^2 \cdot 1_{\left\{\|\Delta_i^n X^c\| \leq u_n,\, |\Delta_i^n Y'| > u_n\right\}} = \sum_{i=1}^{[t/\Delta_n]} (\Delta_i^n Y')^2 \cdot 1_{\left\{|\Delta_i^n Y'| > u_n\right\}}.
$$

Moreover, because $\{T_q\}_{q=1,2,...}$ is a collection of jump times of Z and because Z and X do not co-jump, $\{T_q\}_{q=1,2,...}$ also exhausts the jump times of Y' and $\Delta Y'_{T_q} = \Delta Z_{T_q}$. By Theorem 13.1.1 of Jacod and Protter (2012), we have

$$\frac{1}{\sqrt{\Delta_n}} \left\{ \sum_{i=1}^{[t/\Delta_n]} (\Delta_i^n Y')^2 \cdot 1_{\left\{|\Delta_i^n Y'| > u_n\right\}} - \sum_{s \le t} (\Delta Z_s)^2 \right\} \xrightarrow{\mathcal{L}-s} t \cdot \mathcal{W}_t^J.$$

Therefore, it remains to show that

$$\frac{1}{\sqrt{\Delta_n}} \sum_{i=1}^{[t/\Delta_n]} \left\{ (\Delta_i^n Y)^2 \cdot 1_{\left\{\|\Delta_i^n X\| \le u_n,\, |\Delta_i^n Y| > u_n\right\}} - (\Delta_i^n Y')^2 \cdot 1_{\left\{|\Delta_i^n Y'| > u_n\right\}} \right\} = o_p(1). \tag{A.4}$$

To prove this, we need to derive a general inequality. Let $F_u(x,z) = (x+z)^2 1_{\{\|x+z\|>u,\, \|x\| \le u\}}$. We show below that for $u > 1$,

$$\begin{aligned} &\|F_u(x+y,z) - F_u(x,z)\| = \left\|(x+y+z)^2 1_{\{\|x+y+z\|>u,\, \|x+y\|\le u\}} - (x+z)^2 1_{\{\|x+z\|>u,\, \|x\|\le u\}}\right\| \\ \le & K\left(\|y\|^2 \wedge u^2 + \|y\| \wedge 1\right)\left(1 + \|x\|^2 + \|z\|^2\right) + K u^{-\frac{2}{1-2\varpi}} \|x\|^{2+\frac{2}{1-2\varpi}} + K u^{-\frac{2}{1-2\varpi}} \|z\|^2 \|x\|^{\frac{2}{1-2\varpi}} \\ & + K\|x\|^2 (\|y+z\| \wedge 1 + \|z\| \wedge 1) + K\|z\|^2 \wedge u^2. \end{aligned} \tag{A.5}$$

For convenience, we introduce a decomposition $Y_t = Y_0 + \tilde{X}'_t + \tilde{X}''_t + Z_t^j$, where

$$\tilde{X}'_t = \int_0^t \beta_{s-}^{\intercal} dX_s^c + Z_t^c, \quad \tilde{X}''_t = \sum_{s \le t} \tilde{\beta}_{s-}^{\intercal} \Delta X_s,$$

so that $Y'_t = Y_0 + \tilde{X}'_t + Z_t^j$. In addition, for any $i \le n$, we have

$$\Delta_i^n Y = \Delta_i^n \tilde{X}' + \Delta_i^n \tilde{X}'' + \Delta_i^n Z^j.$$

Applying (A.5) to

$$\eta_i^n := F_{u_n/\sqrt{\Delta_n}}\left(\frac{\Delta_i^n \tilde{X}' + \Delta_i^n \tilde{X}''}{\sqrt{\Delta_n}}, \frac{\Delta_i^n Z^j}{\sqrt{\Delta_n}}\right) - F_{u_n/\sqrt{\Delta_n}}\left(\frac{\Delta_i^n \tilde{X}'}{\sqrt{\Delta_n}}, \frac{\Delta_i^n Z^j}{\sqrt{\Delta_n}}\right),$$

we obtain

$$\begin{aligned}\|\eta_i^n\| \leq & K\left(v_n^2(V_i^n)^2 + W_i^n\right)\left(1 + (U_i^n)^2 + (Q_i^n)^2\right) + Kv_n^{-\frac{2}{1-2\varpi}}(U_i^n)^{2+\frac{2}{1-2\varpi}} + Kv_n^{-\frac{2}{1-2\varpi}}(U_i^n)^{\frac{2}{1-2\varpi}}(Q_i^n)^2 \\ & + K(U_i^n)^2(H_i^n + \overline{W}_i^n) + Kv_n^2(\overline{V}_i^n)^2,\end{aligned}$$

where $v_n = u_n/\sqrt{\Delta_n}$, which we can assume is greater than 1 for n sufficiently large, and

$$\begin{aligned}U_i^n &= \frac{\left\|\Delta_i^n \tilde{X}'\right\|}{\sqrt{\Delta_n}}, \quad V_i^n = \frac{\left\|\Delta_i^n \tilde{X}''\right\|}{\Delta_n^{\varpi}} \bigwedge 1, \quad W_i^n = \frac{\left\|\Delta_i^n \tilde{X}''\right\|}{\sqrt{\Delta_n}} \bigwedge 1, \quad Q_i^n = \frac{\left\|\Delta_i^n Z^j\right\|}{\sqrt{\Delta_n}}, \\ \overline{V}_i^n &= \frac{\left\|\Delta_i^n Z^j\right\|}{\Delta_n^{\varpi}} \bigwedge 1, \quad \overline{W}_i^n = \frac{\left\|\Delta_i^n Z^j\right\|}{\sqrt{\Delta_n}} \bigwedge 1, \quad H_i^n = \frac{\left\|\Delta_i^n Z^j + \Delta_i^n \tilde{X}''\right\|}{\sqrt{\Delta_n}} \bigwedge 1.\end{aligned}$$

By (2.1.33), (2.1.34), (2.1.44), and (2.1.47) in Jacod and Protter (2012), for $q > 0$, there exists a sequence $\phi_n \to 0$ such that

$$\begin{aligned}&\mathbb{E}\left((U_i^n)^q | \mathcal{F}_{(i-1)\Delta_n}\right) \leq K, \\ &\mathbb{E}\left((V_i^n)^q | \mathcal{F}_{(i-1)\Delta_n}\right) \leq \Delta_n^{(1-r\varpi)(1\wedge\frac{q}{r})}\phi_n, \quad \mathbb{E}\left((\overline{V}_i^n)^q | \mathcal{F}_{(i-1)\Delta_n}\right) \leq \Delta_n^{(1-r\varpi)(1\wedge\frac{q}{r})}\phi_n, \\ &\mathbb{E}\left((W_i^n)^q | \mathcal{F}_{(i-1)\Delta_n}\right) \leq \Delta_n^{(1-\frac{r}{2})(1\wedge\frac{q}{r})}\phi_n, \quad \mathbb{E}\left((\overline{W}_i^n)^q | \mathcal{F}_{(i-1)\Delta_n}\right) \leq \Delta_n^{(1-\frac{r}{2})(1\wedge\frac{q}{r})}\phi_n, \\ &\mathbb{E}\left((H_i^n)^q | \mathcal{F}_{(i-1)\Delta_n}\right) \leq \Delta_n^{(1-\frac{r}{2})(1\wedge\frac{q}{r})}\phi_n, \quad \mathbb{E}\left((Q_i^n)^2 | \mathcal{F}_{(i-1)\Delta_n}\right) \leq K.\end{aligned}$$

These estimates, along with $1/(4-2r) \leq \varpi < 1/2$, imply that $\mathbb{E}\|\eta_i^n\| \leq K\sqrt{\Delta_n}\phi_n$, which in turn implies that the left-hand side of (A.4) satisfies

$$\sqrt{\Delta_n}\sum_{i=1}^{[t/\Delta_n]} \eta_i^n \overset{u.c.p.}{\Longrightarrow} 0,$$

which concludes the proof of the asymptotic distribution. The consistency of the asymptotic variance estimator follows from a similar extension of Theorem 9.5.1 of Jacod and Protter (2012).

Appendix A.3.1 Proof of Equation (A.5)

There are several cases we need to distinguish:

1. $\|x+y+z\| > u, \|x+z\| > u$. There are four scenarios to consider.

a. $\|x+y\| \leq u, \|x\| \leq u$. Note that $1_{\{\|x+y\|\leq u, \|x\|\leq u\}} \leq 1_{\{\|y\|\leq 2u\}}$ and

$$(\|y\|^2 + \|y\|)1_{\{\|y\|\leq 2u\}} \leq K\left(u^2 1_{\{u\leq\|y\|\leq 2u\}} + \|y\|^2 1_{\{1\leq\|y\|<u\}} + \|y\| 1_{\{\|y\|<1\}}\right).$$

Therefore, the left-hand side of (A.5) satisfies:

$$\begin{aligned}\|(x+y+z)^2 - (x+z)^2\| \leq& \|y\|^2 + 2\|y\|(\|x\| + \|z\|) \leq K(\|y\|^2 + \|y\|)(1 + \|x\| + \|z\|) \\ \leq& K(\|y\|^2 \wedge u^2 + \|y\| \wedge 1)\left(1 + \|z\|^2 + \|x\|^2\right).\end{aligned}$$

b. $\|x+y\| \leq u$, $\|x\| > u$. The left-hand side is

$$\begin{aligned}\|x+y+z\|^2 \leq& K\|x+y\|^2 + K\|z\|^2 \leq K(u^2 + \|z\|^2)1_{\{\|x\|>u\}} \\ \leq& Ku^{-\frac{2}{1-2\varpi}}\|x\|^{2+\frac{2}{1-2\varpi}} + Ku^{-\frac{2}{1-2\varpi}}\|z\|^2\|x\|^{\frac{2}{1-2\varpi}}.\end{aligned}$$

c. $\|x+y\| > u$, $u/2 < \|x\| \leq u$. Similarly, the left-hand side is

$$\|x+z\|^2 \leq K(u^2 + \|z\|^2)1_{\{\|x\|>u\}} \leq Ku^{-\frac{2}{1-2\varpi}}\|x\|^{2+\frac{2}{1-2\varpi}} + Ku^{-\frac{2}{1-2\varpi}}\|z\|^2\|x\|^{\frac{2}{1-2\varpi}}.$$

d. $\|x+y\| > u$, $\|x\| < u/2$. This implies $\|y\| \geq u/2$. The left-hand side is

$$\|x+z\|^2 \leq K(\|x\|^2 + \|z\|^2)1_{\{\|y\|\geq u/2\}} \leq K(\|x\|^2 + \|z\|^2)(\|y\| \wedge 1).$$

2. $\|x+y+z\| > u, \|x+z\| \leq u$. In this case, only when $\|x+y\| \leq u$, the left-hand side is non-zero. We study two scenarios.

a. $\|x\| > u$. The left-hand side is

$$\begin{aligned}\|x+y+z\|^2 \leq& K\|x+y\|^2 + K\|z\|^2 \leq K(u^2 + \|z\|^2)1_{\{\|x\|>u\}} \\ \leq& Ku^{-\frac{2}{1-2\varpi}}\|x\|^{2+\frac{2}{1-2\varpi}} + Ku^{-\frac{2}{1-2\varpi}}\|z\|^2\|x\|^{\frac{2}{1-2\varpi}}.\end{aligned}$$

b. $\|x\| \leq u$. This leads to $\|y\| \leq 2u$ and $\|z\| \leq 2u$. Hence, the left-hand side satisfies:

$$\begin{aligned}\|x+y+z\|^2 \leq & K\|x\|^2 1_{\{\|x\|\leq u\}} + K\|z\|^2 1_{\{\|z\|\leq 2u\}} + K\|y\|^2 1_{\{\|y\|\leq 2u\}} \\ \leq & K\|x\|^2 1_{\{\|x\|>u/2\}} + K\|x\|^2 1_{\{\|y+z\|\geq u/2\}} + K\|z\|^2 \wedge u^2 + K\|y\|^2 \wedge u^2 \\ \leq & Ku^{-\frac{2}{1-2\varpi}}\|x\|^{2+\frac{2}{1-2\varpi}} + K\|x\|^2(\|y+z\| \wedge 1) + K\|z\|^2 \wedge u^2 + K\|y\|^2 \wedge u^2.\end{aligned}$$

3. $\|x+y+z\| \leq u, \|x+z\| > u$. In this case, the left-hand side is non-zero if $\|x\| \leq u$. Similarly, we consider two scenarios.

   a. $\|x+y\| > u$. The left-hand side becomes

$$\begin{aligned}\|x+z\|^2 \leq & K(\|x\|^2 + \|z\|^2)1_{\{\|x\|\geq u/2\}} + K(\|x\|^2 + \|z\|^2)1_{\{\|y\|\geq u/2\}} \\ \leq & Ku\left(^{-\frac{2}{1-2\varpi}}\|x\|^{2+\frac{2}{1-2\varpi}} + \|z\|^2\|x\|\right).\end{aligned}$$

   b. $\|x+y\| \leq u$. This leads to $\|z\| \leq 2u$. The left-hand side becomes

$$\begin{aligned}\|x+z\|^2 \leq & K\|x\|^2 1_{\{\|x\|>u/2\}} + K\|x\|^2 1_{\{\|z\|>u/2\}} + K\|z\|^2 1_{\{\|z\|\leq 2u\}} \\ \leq & Ku^{-\frac{2}{1-2\varpi}}\|x\|^{2+\frac{2}{1-2\varpi}} + K\|x\|^2(\|z\| \wedge 1) + K\|z\|^2 \wedge u^2.\end{aligned}$$

This concludes the proof.

Appendix B Tables and Figures

Factor Dynamics		Beta Dynamics		Idiosyncratic Dynamics	
b	$(0.05, 0.03, 0.02)$	κ	$(2, 2, 2)$	γ	0.35
σ_0^2	$(0.12, 0.09, 0.04)$	α	$(0.15, 0.10, -0.10)$	h	0
$\widetilde{\kappa}$	$(3, 4, 5)$	υ	$(0.03, 0.03, 0.03)$	g_i^+	$\sigma_{i0}\times 7\sqrt{\Delta_n}$
$\widetilde{g}$	$(0.004, 0.005, 0.004)$			g_i^-	$\sigma_{i0}\times 7\sqrt{\Delta_n}$
$\overline{g}^+$	$\gamma \times 14\sqrt{\Delta_n}$			$\overline{q}$	0.5
$\overline{g}^-$	$\gamma \times 14\sqrt{\Delta_n}$				
λ	67				
$\widetilde{\alpha}$	$(0.09, 0.04, 0.06)$				
$\widetilde{\upsilon}$	$(0.3, 0.4, 0.3)$				
q	$(0.5, 0.5, 0.5)$				
$(\rho_{12}, \rho_{13}, \rho_{23})$	$(0.05, 0.10, 0.15)$				

Table 1: Monte Carlo Simulations: Parameter Values

Note: This table reports the parameter values used in the data generating process of our Monte Carlo simulations.

	5 min sampling frequency								
	k_n= 78			k_n= 91			k_n= 117		
	bias	stdev	RMSE	bias	stdev	RMSE	bias	stdev	RMSE
$\widehat{I\beta_t^{(1)}}$	-0.15	2.64	2.65	-0.16	2.64	2.65	-0.15	2.65	2.65
$\widehat{I\beta_t^{(2)}}$	-0.05	3.13	3.13	-0.05	3.13	3.13	-0.05	3.11	3.11
$\widehat{I\beta_t^{(3)}}$	0.26	4.18	4.19	0.31	4.16	4.17	0.24	4.14	4.14
$\widehat{IdV}_t$	0.06	0.46	0.47	0.07	0.46	0.47	0.08	0.46	0.47
$\widehat{IdJ}_t$	-0.06	0.78	0.78	-0.06	0.78	0.78	-0.06	0.78	0.78
	10 min sampling frequency								
	k_n= 39			k_n= 63			k_n= 91		
	bias	stdev	RMSE	bias	stdev	RMSE	bias	stdev	RMSE
$\widehat{I\beta_t^{(1)}}$	-0.10	3.83	3.83	-0.14	3.75	3.75	-0.14	3.67	3.67
$\widehat{I\beta_t^{(2)}}$	0.06	4.67	4.67	0.02	4.47	4.47	0.03	4.49	4.49
$\widehat{I\beta_t^{(3)}}$	-0.17	6.46	6.46	-0.20	6.23	6.23	-0.18	6.25	6.26
$\widehat{IdV}_t$	0.03	0.67	0.67	0.07	0.66	0.66	0.09	0.65	0.66
$\widehat{IdJ}_t$	0.00	0.98	0.98	0.00	0.98	0.98	0.00	0.98	0.98
	30 min sampling frequency								
	k_n= 13			k_n= 21			k_n= 39		
	bias	stdev	RMSE	bias	stdev	RMSE	bias	stdev	RMSE
$\widehat{I\beta_t^{(1)}}$	-0.17	7.56	7.56	0.01	6.97	6.97	0.04	6.47	6.47
$\widehat{I\beta_t^{(2)}}$	0.07	9.67	9.68	0.34	8.91	8.92	0.25	8.37	8.38
$\widehat{I\beta_t^{(3)}}$	0.18	13.02	13.02	-0.04	12.21	12.21	0.07	11.39	11.39
$\widehat{IdV}_t$	-0.60	1.22	1.36	-0.17	1.20	1.21	0.02	1.20	1.20
$\widehat{IdJ}_t$	-0.03	1.88	1.88	-0.03	1.88	1.88	-0.03	1.88	1.88

Table 2: Simulation Results

Note: Columns "bias", "stdev", and "RMSE" contain the sample bias, the sample standard deviation, and the sample root mean squared error across simulations of the centered estimator, e.g., of the quantity $\widehat{IdV}_t - IdV_t$. All these quantities are multiplied by 100. k_n is the number of observations in a subsample. $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$.

	5 min sampling frequency								
	k_n= 78			k_n= 91			k_n= 117		
	bias	stdev	RMSE	bias	stdev	RMSE	bias	stdev	RMSE
$\widehat{I\beta_t^{(1)}}$	-0.14	2.63	2.64	-0.16	2.63	2.64	-0.15	2.64	2.64
$\widehat{I\beta_t^{(2)}}$	-0.03	3.15	3.15	-0.04	3.16	3.16	-0.02	3.13	3.13
$\widehat{I\beta_t^{(3)}}$	0.26	4.13	4.14	0.31	4.13	4.14	0.25	4.10	4.10
$\widehat{IdV}_t$	0.20	0.50	0.54	0.20	0.50	0.54	0.21	0.50	0.54
$\widehat{IdJ}_t$	-0.13	0.77	0.78	-0.13	0.77	0.78	-0.13	0.77	0.78
	10 min sampling frequency								
	k_n= 39			k_n= 63			k_n= 91		
	bias	stdev	RMSE	bias	stdev	RMSE	bias	stdev	RMSE
$\widehat{I\beta_t^{(1)}}$	-0.06	3.89	3.89	-0.08	3.82	3.82	-0.08	3.72	3.72
$\widehat{I\beta_t^{(2)}}$	0.06	4.62	4.62	0.01	4.44	4.44	0.02	4.47	4.47
$\widehat{I\beta_t^{(3)}}$	-0.21	6.51	6.51	-0.23	6.31	6.31	-0.23	6.31	6.31
$\widehat{IdV}_t$	0.17	0.74	0.76	0.22	0.73	0.76	0.24	0.73	0.77
$\widehat{IdJ}_t$	-0.18	1.02	1.03	-0.18	1.02	1.03	-0.18	1.02	1.03
	30 min sampling frequency								
	k_n= 13			k_n= 21			k_n= 39		
	bias	stdev	RMSE	bias	stdev	RMSE	bias	stdev	RMSE
$\widehat{I\beta_t^{(1)}}$	-0.13	7.59	7.59	0.06	6.97	6.97	0.07	6.46	6.46
$\widehat{I\beta_t^{(2)}}$	0.03	9.89	9.89	0.30	8.90	8.90	0.16	8.32	8.33
$\widehat{I\beta_t^{(3)}}$	0.15	12.84	12.84	-0.08	12.05	12.05	0.01	11.25	11.25
$\widehat{IdV}_t$	-0.24	1.74	1.75	0.20	1.75	1.76	0.40	1.76	1.80
$\widehat{IdJ}_t$	-0.41	2.27	2.30	-0.41	2.27	2.30	-0.41	2.27	2.30

Table 3: Simulation Results

Note: Columns "bias", "stdev", and "RMSE" contain the sample bias, the sample standard deviation, and the sample root mean squared error across simulations of the centered estimator, e.g., of the quantity $\widehat{IdV}_t - IdV_t$. All these quantities are multiplied by 100. k_n is the number of observations in a subsample. $u_n = 4\Delta_n^{0.47}\sqrt{BV_t}$.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
	Low Threshold							
Intercept	-3.79	-3.89	-3.82	-3.88	-3.57	-3.73	-3.68	-3.80
	(-141.95)	(-225.89)	(-148.81)	(-258.24)	(-130.06)	(-170.14)	(-138.44)	(-185.88)
log(Surprise)	0.14	0.10	0.13	0.11	0.18	0.13	0.15	0.11
	(18.57)	(17.86)	(17.95)	(21.09)	(21.77)	(17.44)	(19.29)	(16.56)
Positive					-0.45	-0.31	-0.27	-0.15
					(-15.45)	(-12.78)	(-10.64)	(-6.62)
log(Surprise) × Positive					-0.09	-0.05	-0.05	-0.01
					(-10.39)	(-7.15)	(-6.42)	(-2.06)
Firm Effect		✓		✓		✓		✓
Month Effect			✓	✓			✓	✓
R-Squared (%)	1.42	1.73	1.40	1.84	1.49	1.96	1.48	1.89
No. Observations	117,557	117,557	117,557	117,557	117,557	117,557	117,557	117,557
	High Threshold							
Intercept	-3.84	-3.86	-3.87	-3.87	-3.62	-3.70	-3.74	-3.79
	(-132.79)	(-188.08)	(-140.08)	(-198.17)	(-115.53)	(-139.86)	(-123.24)	(-147.40)
log(Surprise)	0.14	0.13	0.13	0.13	0.17	0.15	0.14	0.12
	(17.05)	(18.78)	(15.88)	(19.06)	(18.46)	(16.81)	(15.54)	(14.56)
Positive					-0.43	-0.30	-0.25	-0.14
					(-12.84)	(-9.98)	(-8.23)	(-4.98)
log(Surprise) × Positive					-0.08	-0.04	-0.04	0.00
					(-7.81)	(-4.19)	(-3.90)	(0.14)
Firm Effect		✓		✓		✓		✓
Month Effect			✓	✓			✓	✓
R-Squared (%)	1.31	1.61	1.32	1.72	1.32	1.75	1.31	1.69
No. Observations	82,210	82,210	82,210	82,210	82,210	82,210	82,210	82,210

Table 4: Idiosyncratic Jumps and Earnings Surprises

Note: This table reports 8 panel regression specifications. The dependent variable is always the logarithm of the idiosyncratic jumps. The covariates include the logarithm of the earnings surprise, defined as the absolute difference between the realized and expected earnings per share. "Positive" is a dummy variable that indicates whether the earnings surprise is positive or negative. We also include firm and month fixed effects in some of these regressions. t-statistics are provided in parentheses. The "low threshold" is $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$, whereas the "high threshold" is $u_n = 4\Delta_n^{0.47}\sqrt{BV_t}$.

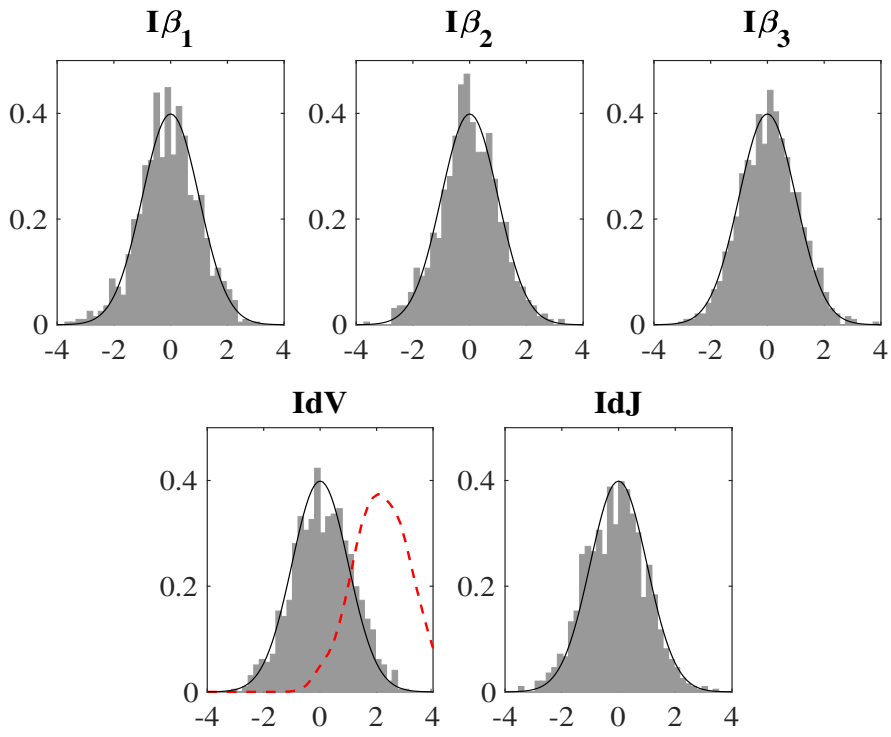

Figure 1: Simulation Results: Standardized Estimates

Note: We plot the finite sample distributions of the standardized statistics (gray histograms), and we superimpose the standard normal law (black solid line). The finite sample distribution of the $\widehat{IdV}_t^{naive}$ estimator appears in red dashed lines. The sampling frequency is every 5 minutes and the subsamples are one day long, i.e., $k_n = 78$.

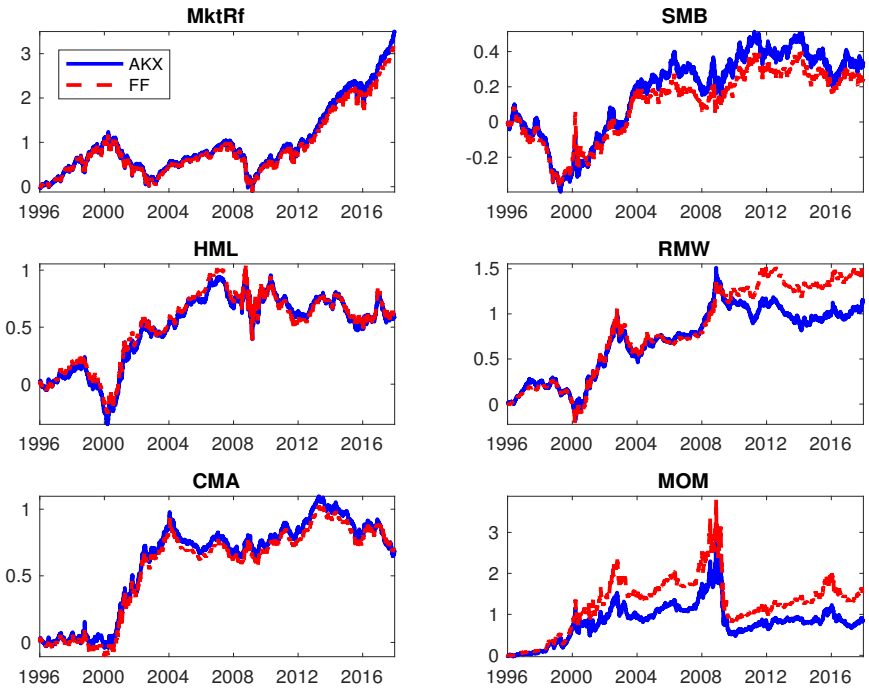

Figure 2: Factors

Note: This figure compares the daily aggregated cumulative returns of the high-frequency factors, in blue solid lines (AKX), with those of the low-frequency factors, in red dashed lines (FF), downloaded from Kenneth French's website. The sampling period is January 1996 - December 2017.

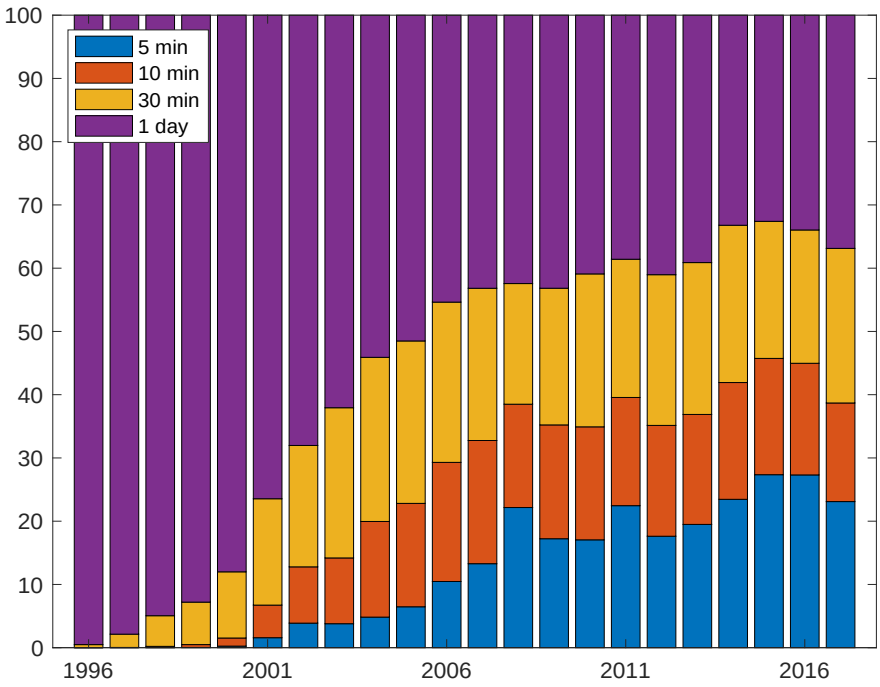

Figure 3: Percentage of the Selected Sampling Frequencies

Note: This figure compares the percentages of the selected sampling frequencies for combinations of stocks and months in each year from 1996 to 2017.

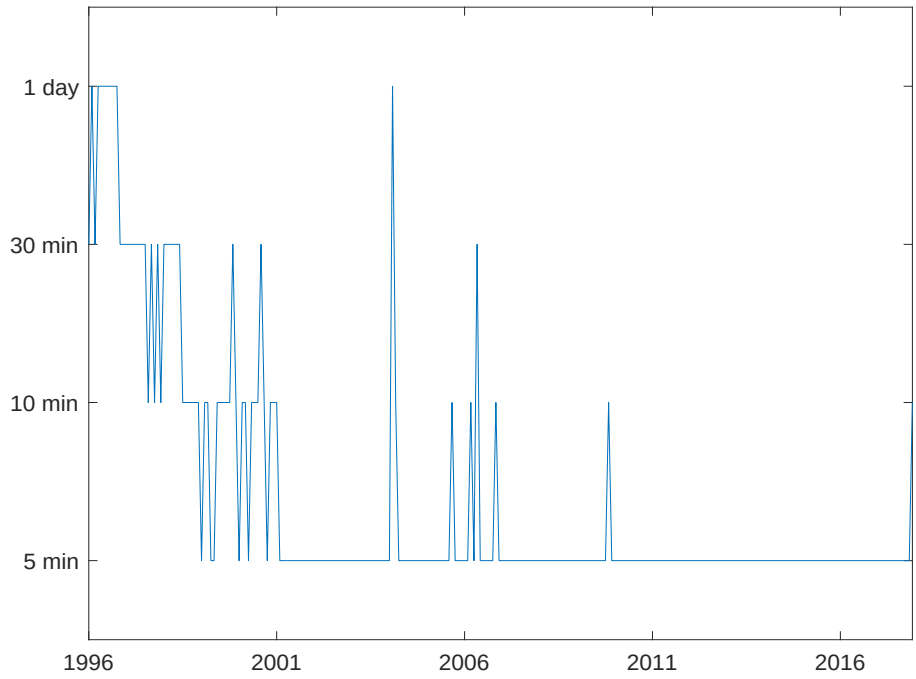

Figure 4: Time Series of the Selected Sampling Frequencies for IBM

Note: This figure plots the monthly time series of the selected sampling frequencies for IBM from 1996 to 2017.

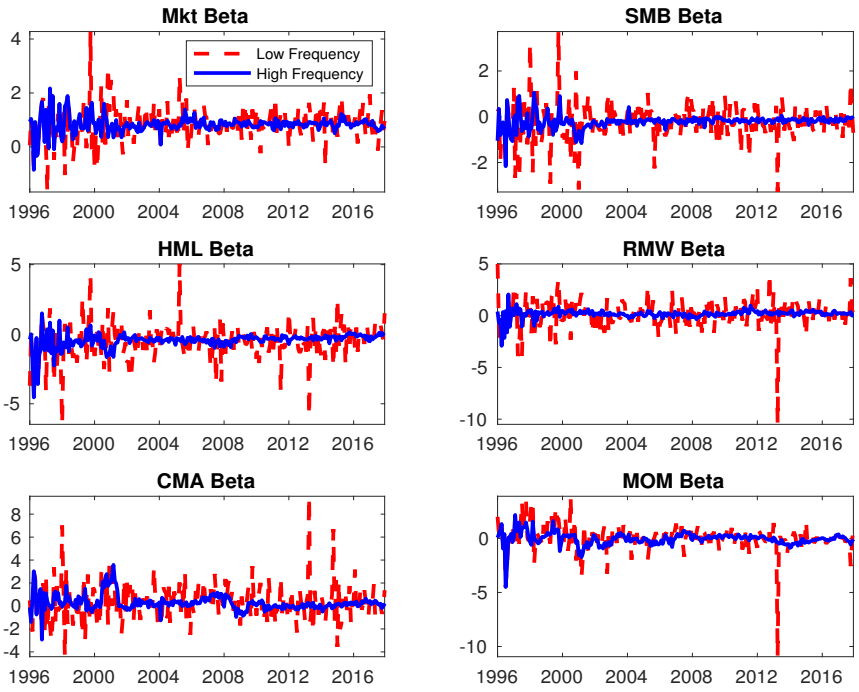

Figure 5: The Six-Factor Betas of the IBM Stock

Note: This figure compares the time series of the high-frequency betas (blue solid line) with the low-frequency betas (red dashed line) for IBM estimated from the six-factor model for each month from 1996 to 2017.

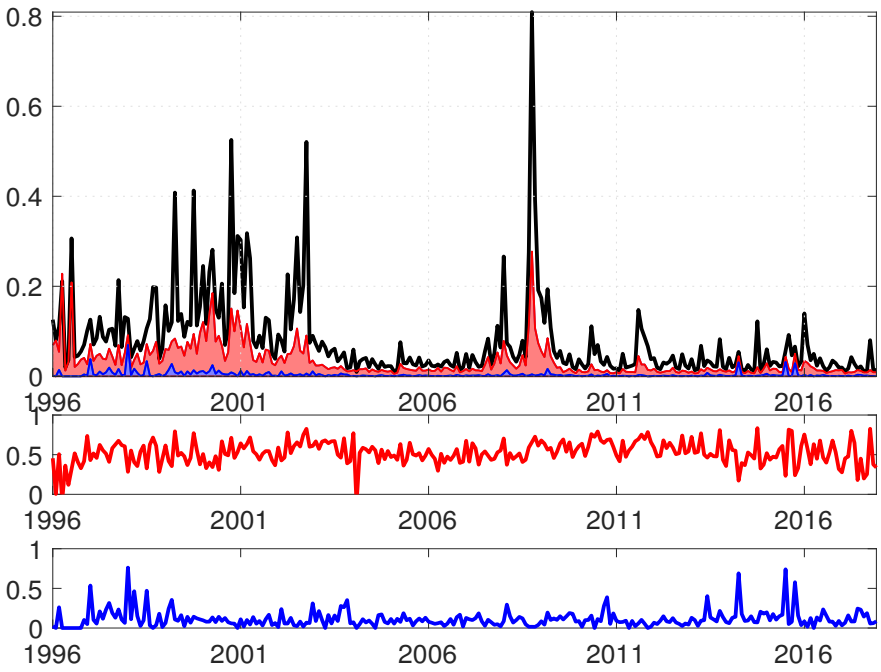

Figure 6: Decomposition of the IBM's Realized Volatility

Note: The **top panel** compares the total risk of IBM with its total idiosyncratic risk, which is then further decomposed into the continuous component and the jump component. The total risk and the total idiosyncratic risk are measured by the annualized quadratic variations of IBM returns (thick black line) and its residual returns (thin red line), respectively. The jump component of the total idiosyncratic risk is shaded (in blue) at the bottom of the plot. The **middle panel** plots the fraction of total variation explained by the systematic component (R^2), whereas the **bottom panel** plots the fraction of the variation of the total idiosyncratic risk composed of idiosyncratic jumps.

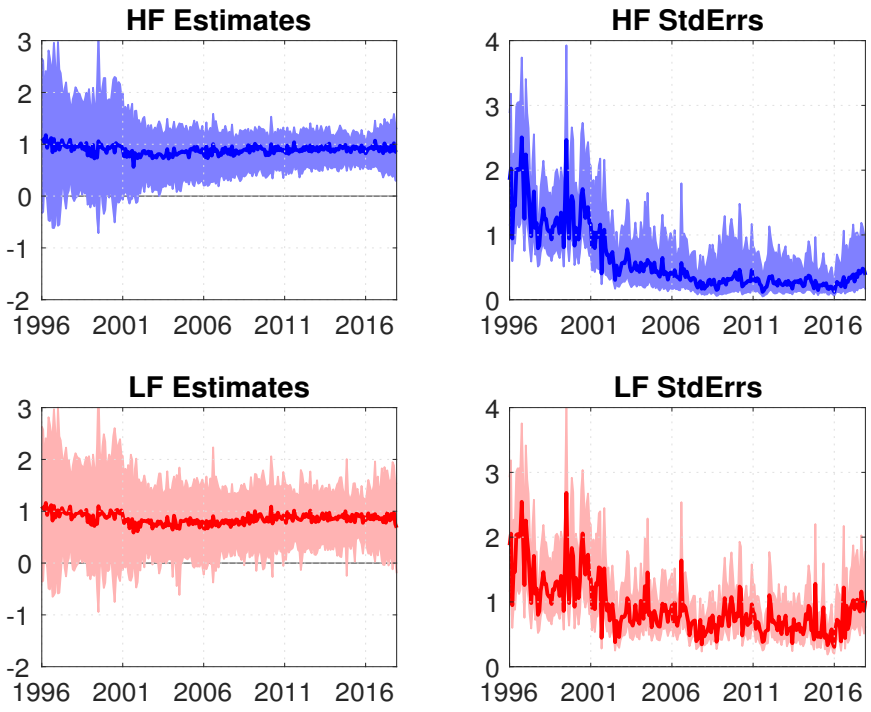

Figure 7: Market Beta

Note: In this figure, we compare the high-frequency and the low-frequency (i.e., daily frequency) estimates of the MKT beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

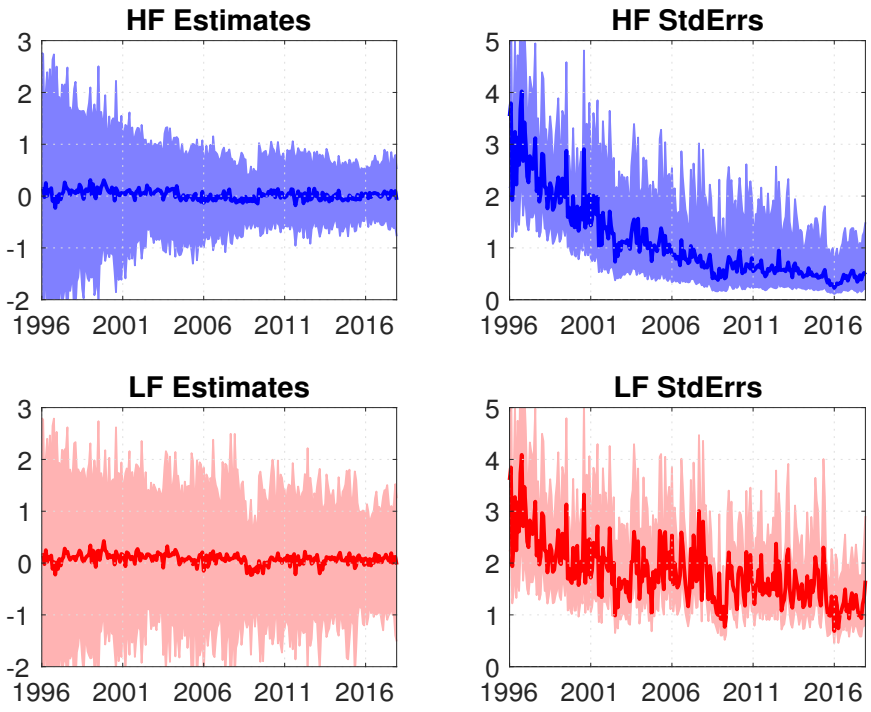

Figure 8: HML Beta

Note: In this figure, we compare the high-frequency and the low-frequency estimates of the HML beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

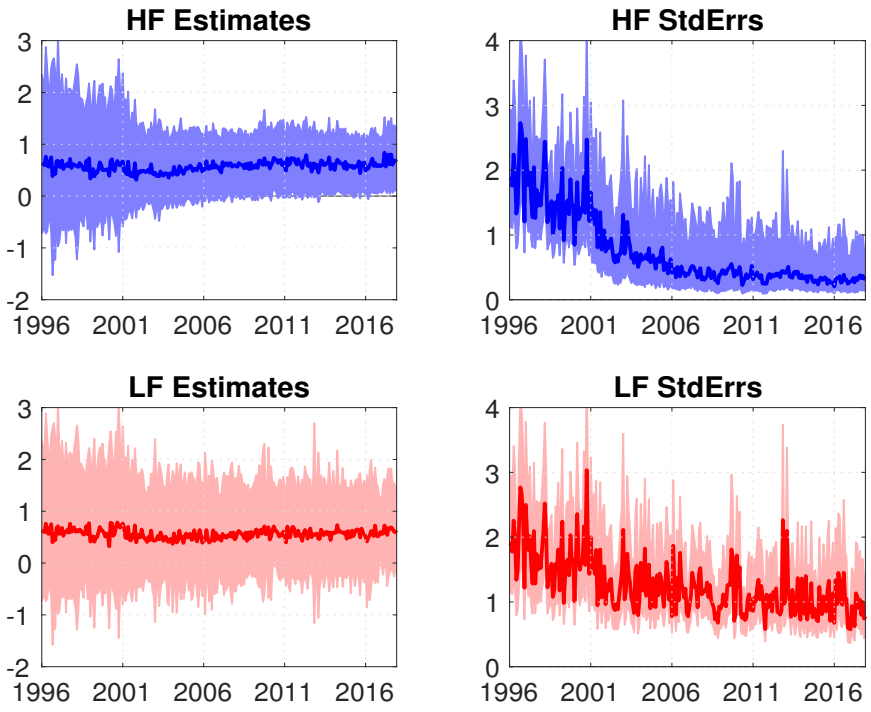

Figure 9: SMB Beta

Note: In this figure, we compare the high-frequency and the low-frequency estimates of the SMB beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

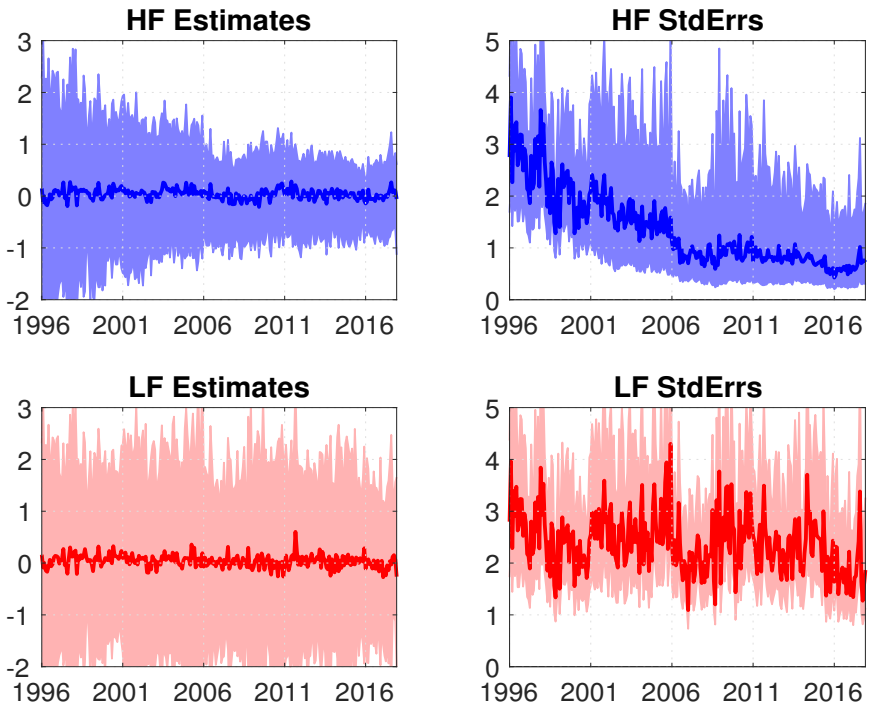

Figure 10: CMA Beta

Note: In this figure, we compare the high-frequency and the low-frequency estimates of the CMA beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

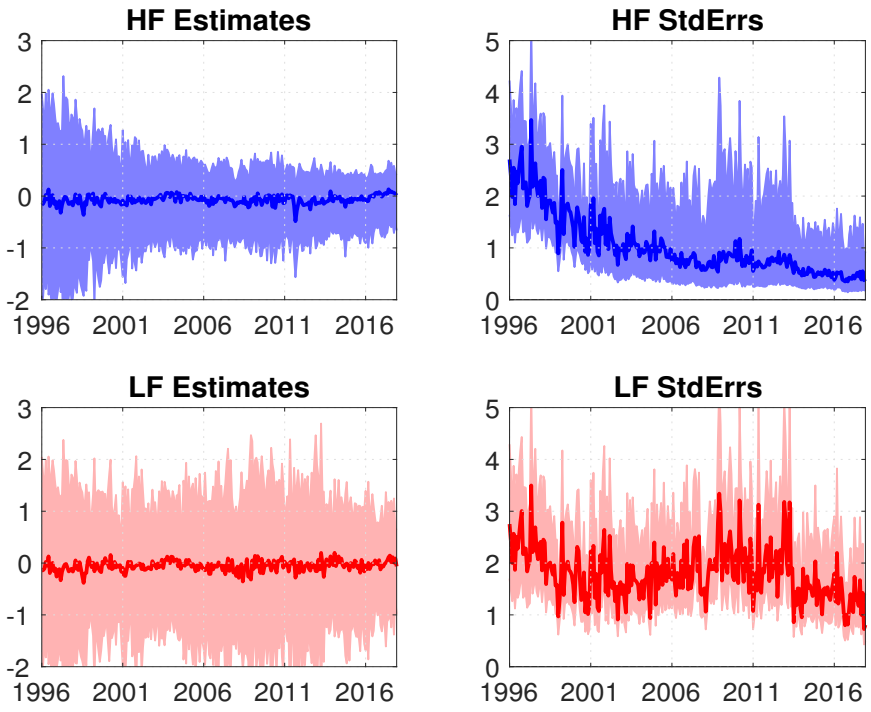

Figure 11: RMW Beta

Note: In this figure, we compare the high-frequency and the low-frequency estimates of the RMW beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

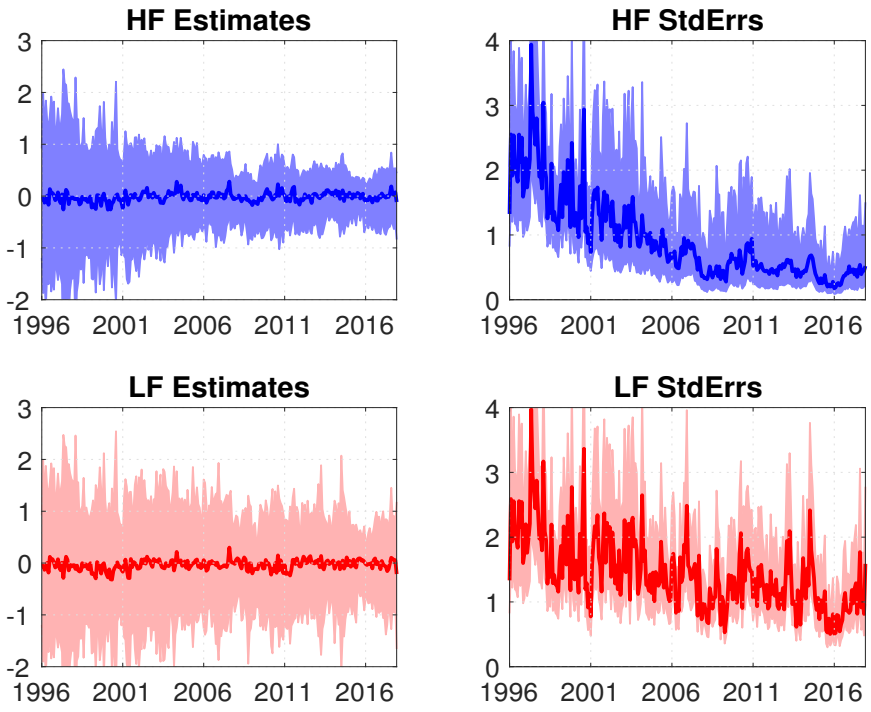

Figure 12: MOM Beta

Note: In this figure, we compare the high-frequency and the low-frequency estimates of the MOM beta in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

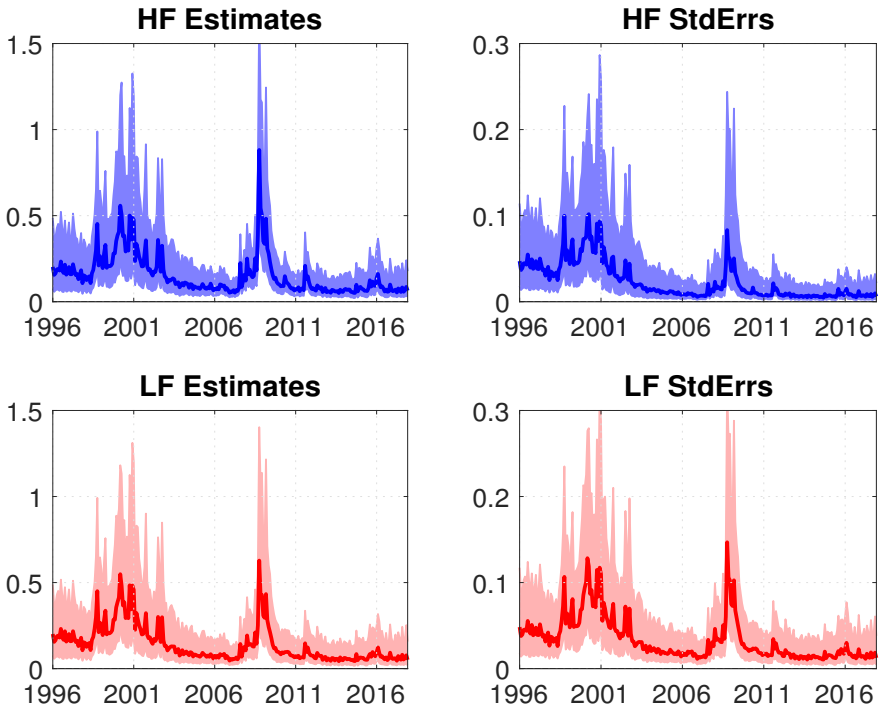

Figure 13: Total Idiosyncratic Risk

Note: In this figure, we compare the high-frequency estimates of the total idiosyncratic risk (including both the continuous and jump components) with the low-frequency estimates of the idiosyncratic volatility in the six-factor model. The solid bold lines correspond to the cross-sectional median; the bands are the cross-sectional 25- and 75-percentiles of the estimates. "StdErrs" denotes the estimated asymptotic standard errors.

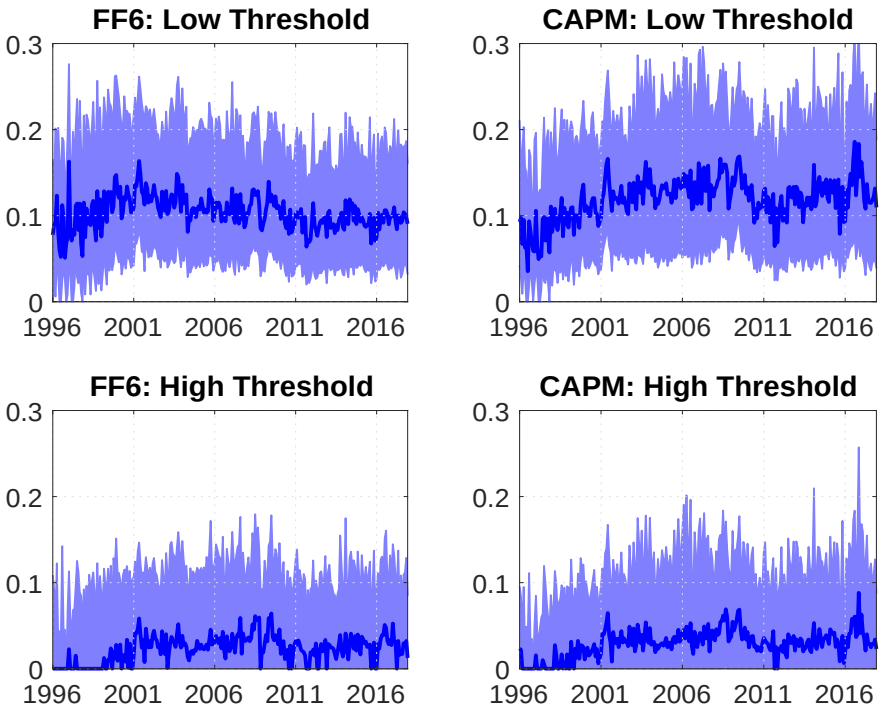

Figure 14: Percentages of IdJ

Note: In this figure, we compare the ratios of the idiosyncratic jump component over the total idiosyncratic risk for the six-factor model and CAPM. The "low threshold" is $u_n = 3\Delta_n^{0.47}\sqrt{BV_t}$, whereas the "high threshold" is $u_n = 4\Delta_n^{0.47}\sqrt{BV_t}$. All plots of this figure use the smaller cross section, which excludes the stock-months for which the daily sampling frequency is selected. The solid blue line corresponds to the cross-sectional median; the bands are 25- and 75-percentiles of the considered cross section.

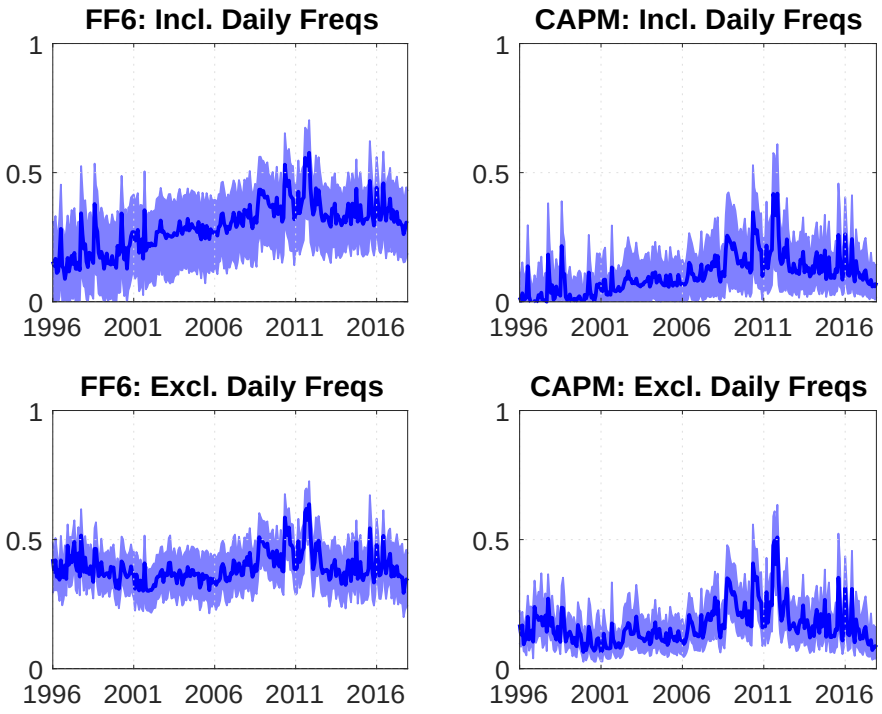

Figure 15: Percentages of R^2

Note: In this figure, we compare the fraction of the variation in total risk explained by the systematic component (R^2) for the six-factor model and CAPM, using high-frequency data. The lower panel is based on a smaller cross section that excludes all stock-months for which the daily sampling frequency is selected for estimation. The solid blue line corresponds to the cross-sectional median; the bands are 25- and 75-percentiles of the considered cross section.