# Data Synchronization at High Frequencies

Xinbing Kong,[a] Cheng Liu,[b] Bin Wu,[c*]
[a]Southeast University; [b]Wuhan University; [c]University of Science and Technology of China
*Corresponding author
**Contact:** xinbingkong@126.com (XK), chengliu_eco@whu.edu.cn (CL), bin.w@ustc.edu.cn (BW)



**Abstract.** Asynchronous trading in high-frequency financial markets introduces significant biases into econometric analysis, distorting risk estimates and leading to suboptimal portfolio decisions. Existing synchronization methods, such as the previous-tick approach, suffer from information loss and create artificial price staleness. We introduce a novel framework that recasts the data synchronization challenge as a constrained matrix completion problem. Our approach recovers the potential matrix of high-frequency price increments by minimizing its nuclear norm—capturing the underlying low-rank factor structure—subject to a large-scale linear system derived from observed, asynchronous price changes. Theoretically, we prove the existence and uniqueness of our estimator and establish its convergence rate. A key theoretical insight is that our method accurately and robustly leverages information from both frequently and infrequently traded assets, overcoming a critical difficulty of efficiency loss in traditional methods. Empirically, using extensive simulations and a large panel of S&P 500 stocks, we demonstrate that our method substantially outperforms established benchmarks. It not only achieves significantly lower synchronization errors, but also corrects the bias in systematic risk estimates (i.e., eigenvalues) and the estimate of betas caused by stale prices. Crucially, portfolios constructed using our synchronized data yield consistently and economically significant higher out-of-sample Sharpe ratios. Our framework provides a powerful tool for uncovering the true dynamics of asset prices, with direct implications for high-frequency risk management, algorithmic trading, and econometric inference.

**Key words:** Finance; High-Frequency Data; Asynchronous Trading; Nuclear Norm Minimization; Price Staleness.

## 1. Introduction

Asynchronicity is a stylized feature of high-frequency data, rooted in the natural mechanisms of data generation. In financial markets, for instance, assets trade at distinct, irregular time instances due to factors such as price staleness, market friction, varying liquidity, and the differential speed of information flow. Price staleness, a near-universal phenomenon, means that efficient prices update at heterogeneous rates across assets, leading to non-homogeneously spaced observations. When a price does not update, the observed price is merely a repetition of the last recorded tick, creating a challenge for time-series analysis. Beyond finance, asynchronicity is also pervasive in fields like ecology, clinical medicine, and environmental science, often arising from mixed-frequency sampling schemes.

The presence of asynchronicity poses significant challenges to econometric and statistical analysis, distorting estimates and leading to flawed conclusions in applications such as covariance matrix estimation, multivariate analysis, portfolio allocation, and beta estimation. The consequences can

be profound. For example, Hollstein et al. (2020) demonstrate that the empirical validity of the Conditional Capital Asset Pricing Model (CAPM) hinges critically on the accurate estimation of betas from high-frequency data. Their work suggests that measurement errors, exacerbated by asynchronous trading, can lead to the premature rejection of foundational economic theories. The core statistical challenges are twofold. First, simple imputation methods like the nearest-tick or previous-tick approach introduce biases in multivariate estimation (Hayashi and Yoshida 2005). Second, in high-dimensional settings, conventional data alignment via subsampling discards a vast amount of non-synchronized data, resulting in a significant loss of efficiency. While numerous approaches have been developed to mitigate these issues (c.f., Chen et al. 2020; Fan et al. (2016); Kong and Liu 2018; Pelger 2019; Shin et al. (2023); Kong et al. 2023; Cui et al. 2024), they often struggle to balance bias reduction with the full utilization of available information.

There are two streams of works that delve into synchronizing the high-frequency data to facilitate the subsequent applications. The first stream synchronizes the data by dropping some original data. Three main methods in this stream are the Previous Tick method (Zhang 2011), the Refresh Time Scheme (Barndorff-Nielsen et al. 2011), and the Generalized Synchronization procedure (Aït-Sahalia et al. 2010). All three methods first select the synchronized sampling time points and then, for each asset and each selected time point, choose one observation that is most close to each synchronized sampling time point from all the original observations of that asset. The drawback of those methods is the drop of a large proportion of data, especially when the number of assets is large and some assets are sparse. The second stream is considering the original data set as a set with missing values and imputing data relying on a parametric state space model and EM algorithm, see Liu and Tang (2014) and Shephard and Xiu (2017). Howerver, in high-dimensional settings, the EM algorithm is computationally time consuming.

In this paper, we introduce a novel method for data synchronization. We formulate the problem by minimizing the nuclear norm of the potential increment matrix that are ideally synchronized and well structured, under a large system of linear constraints of increments over non-synchronous durations, see (2) below. The usage of the nuclear norm is inspired by the well-known low-rank plus noise construction of the data matrix in many applications. The common factor component plus the idiosyncratic error term (discrete or continuous) is a concrete example in finance. The linear constraints over the durations serve as a natural extraction of the data generation mechanism that realizes the data asynchronicity.

Solving this constrained low-rank optimization problem effectively removes idiosyncratic noise and recovers the "signal" component of the potential increment matrix from complex, disorganized data. Our method differs from the tick-sampling approaches by utilizing all data points, thereby avoiding the efficiency loss and potential biases from using overlapping increments in subsequent inference,

such as covolatility estimation. It also differs from EM-based methods by being computationally stable and theoretically grounded in large-dimensional diffusion systems. At its core, our approach is a relaxed rank minimization problem, straightforward to implement and efficiently solved using the Alternating Direction Method of Multipliers (ADMM).

Our method is also different from the price matrix completion via sampling projection operator though our data synchronization approach can be deemed as a potential increment matrix completion via solving large random linear systems. Indeed, synchronizing the data discretely sampled from a large dimensional diffusion process by price matrix completion via the projection operator (projecting a parameter matrix onto the family of matrices that is supported only on sampled entries leaving missing entries set as zeros) meets difficulty since the idiosyncratic error process is intrinsically non-stationary as a general semi-martingale with stochastic volatility. This non-stationarity makes the noisy part as strong as the signal part leading to the so-called spurious factors, see Zhang et al. (2018) and Onatski and Wang (2024). Taking difference of the semi-martingale to relieve the non-stationarity and simultaneously do the sampling projection is however not applicable because of the data asynchronicity. But all durations and the increments over the durations are observable, and they are simply sum of potential (ideally synchronized but latent) increments, which forms a linear system. This is how come our procedure.

Rank minimization under linear constraints has been studied in optimization and operations research. For instance, Recht et al. (2010) proves that the nuclear norm minimization has a unique solution when the linear operator satisfies the nearly isometry condition. However, their framework is largely deterministic, assuming a noise-free data matrix and a linear operator with randomness properties that are independent of the data. In our setting, both the linear operator (related to the random trading durations) and the data-generating process are stochastic. We further allow for a stochastic idiosyncratic noise process to contaminate the low-rank signal. This introduces significant complexity, as the potential correlation between the diffusion process components and between the linear operator and the potential increment matrix makes the derivation of concentration inequalities highly non-trivial. Consequently, no statistical theory has previously existed for the solution's existence and convergence properties in the context of asynchronous, noise-contaminated, large-panel high-frequency data. This paper aims to fill this theoretical gap.

Our work makes several contributions. To the best of our knowledge, we are the first to prove, under a high-dimensional high-frequency asymptotic regime, that nuclear norm minimization under these stochastic linear constraints yields a unique solution equal to the true low-rank matrix with high probability. This is achieved by establishing the restricted isometry property via a novel concentration inequality for the self-normalized Frobenius norm of a large realized covariance matrix. We are also the first to provide a statistical convergence rate for the completed potential increment matrix. An

interesting finding is that the statistical accuracy depends on the sample sizes of both sparse and dense series, in contrast to many tick-based approaches that suffer from efficiency loss.

Empirically, our contributions are equally significant. Through extensive simulations and analysis of a large panel of S&P 500 stocks, we have demonstrated the clear superiority of our method. It not only yields substantially more accurate estimates of the underlying return and covariance matrices but also provides a more realistic depiction of systematic risk by correcting the bias in eigenvalues caused by stale prices. Most importantly, we have shown that this statistical superiority translates directly into economic value: portfolios constructed using our synchronized data generate consistently higher out-of-sample Sharpe ratios. Our analysis of spot beta dynamics during market turmoil further underscores the reliability of our approach, producing stable and economically intuitive risk profiles where traditional methods fail.

The present paper is organized as follows. In Section 2, we introduce our methodology and model set up in detail. Our main theoretical results are provided in Section 3. The Monte Carlo simulations are conducted in Section 4. Empirical studies and findings are given in Section 5. All technical proofs, robustness check and additional results are included in the Supplementary Appendix.

## 2. Methodology and Model Specification

### 2.1. Asynchronicity and Large-Scale Linear System

Motivated by the high-frequency data analysis, we assume that the asynchronous data is discretely sampled from a large-dimensional diffusion process defined on some filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t; 0 \le t \le T)$ as follows.

$$(dX_t)_{N\times 1} = (\mu_t)_{N\times 1}dt + (\sigma_t)_{N\times r}(dW_t)_{r\times 1} + (\sigma_t^*)_{N\times N}(dW_t^*)_{N\times 1}, \tag{1}$$

where $W_t$ is a standard multivariate Brownian motion, $\sigma_t$ is the spot volatility process for the "signal" component, $\sigma_t^*$ is a diagonal matrix of spot volatility process for the idiosyncratic diffusion process while the $W_t^*$ is the driving Brownian motion, and $T$ is a fixed time horizon. The first term in the right hand side of (1) represents the drift term, the second term is a low-rank common component to be reconstructed and the third one is an idiosyncratic "noise" term.

Data asynchronicity means that the coordinate processes $(X_{it}, i = 1, ..., N)$ of $X_t$ are separately generated in completely different time instances. Assume that $X_{it}$ is sampled at time instances $\{\tau_{i0}, \tau_{i1}, \tau_{i2}, ..., \tau_{ii_n}\}$ which are different across $i = 1, ..., N$. Let $\mathcal{T} = \{t_1, ..., t_n\} = \cup_{1\le i\le N}\{\tau_{i1}, ..., \tau_{ii_n}\}$ and $\tau_{i0} = 0$ for all $i = 1, ..., N$. Then $X_{it_j}$ can be thought of as missing value if $t_j \notin \{\tau_{i1}, ..., \tau_{ii_n}\}$. Let the potential increment matrix be $\Delta = [(\Delta_1, ..., \Delta_N)']_{N\times n}$ with $\Delta_{ij} = X_{it_j} - X_{it_{j-1}}$, and the potential matrix be $X = (X_{it_j})_{N\times n}$. Though $X_{\cdot t_j}$, the $j$-th column of $X$, has at least one observed data, $\Delta_{j\cdot}$, the $j$-th row of $\Delta$, may not. This makes the projection operator approach for matrix completion

based on $\Delta$ not applicable, see for example Candes and Recht (2012) and Chen et al. (2019). The same method applied to $X$ is of difficult also because the semi-martingale is generally nonstationary making the idiosyncratic "noise" part as strong as the "signal" part leading to spurious factors, see Zhang et al. (2018) and Onatski and Wang (2024).

Though the potential increment matrix $\Delta$ is far from fully observed, the increments over durations, a linear transform of $\Delta$, can be fully observed. Define a linear operator $\mathcal{A}$ as follows.

$$\mathcal{A}(\Delta) = \text{diag}\{A_1, ..., A_N\}\text{vec}(\Delta') =: A\text{vec}(\Delta'),$$

where $A_i = (a_{jk}^{(i)})_{n_i \times n}$ is a matrix of zeros and ones so that $A_i\Delta_i = b_i$ where $b_i = (X_{i\tau_{i1}} - X_{i\tau_{i0}}, ..., X_{i\tau_{in_i}} - X_{i\tau_{i(n_i-1)}})'$ which is the vector of observed increments of $X_{it}$ over durations, and $\text{vec}(\cdot)$ is the standard vectorization operator. A simple example of $A_i$ is as follows.

$$A_i = \begin{pmatrix} a_{11}^{(i)} & a_{12}^{(i)} & \cdots & a_{1n}^{(i)} \\ a_{21}^{(i)} & a_{22}^{(i)} & \cdots & a_{2n}^{(i)} \\ & \cdots & \cdots & \\ a_{n_i1}^{(i)} & a_{n_i2}^{(i)} & \cdots & a_{n_in}^{(i)} \end{pmatrix} = \begin{pmatrix} 1\ 1\ 0\ 0\ , \cdots , 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0\ , \cdots , 0\ 0\ 0\ 0 \\ & , \cdots , & \\ 0\ 0\ 0\ 1\ , \cdots , 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ , \cdots , 0\ 1\ 1\ 1 \end{pmatrix}.$$

The first row of $A_i$ amounts to saying that we can only observe the sum of the first two potential increments of $X_{it}$ but not any of them. In finance, this is caused by the trading mechanism so that the observation times are typically random and unequally spaced. Let $b = (b_1', ..., b_N')'$. The constraint for $\Delta$ is $\mathcal{A}(\Delta) = b$. An illustration of the asynchronicity for two assets is in Figure 1

Both in theory and applications, we assume that the signal term $\{\sigma_t dW_t | 0 \le t \le T\}$ is in some manifold of low-dimension. Suppose that for each sample path $\omega$, $\sigma_t(\omega) = (\sigma_0)_{N\times r}(\Sigma_t(\omega))_{r\times r}$ where $\sigma_0$ is a constant volatility level and the time variation of the volatility process $\sigma_t\sigma_t'$ is due to $\Sigma_t$. Typically, we assume that $\Sigma_t$ is always of full rank for any time $t$. Without loss of generality, we assume, as in Kong (2017, 2018), that $\sigma_t^*$ is a diagonal matrix with locally bounded entries of stochastic processes, and $W_t^*$ is a $N$-dimensional Brownian motion with correlation matrix $(\rho^*)_{N\times N}$. So the parameter space is

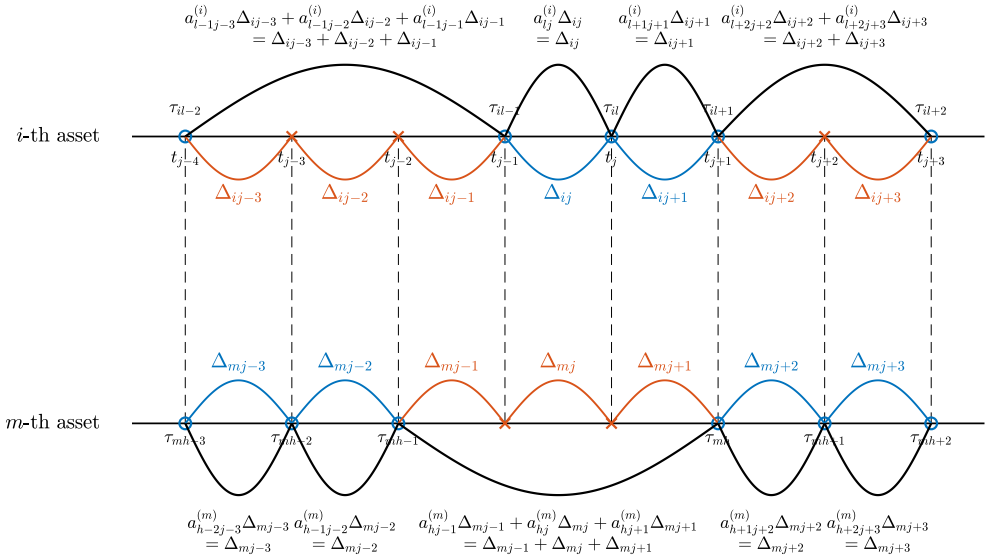
$$\Theta = \{(\sigma^0, \Sigma_t(\omega)dW_t(\omega)); (\sigma^0)'\sigma^0 = N^\alpha, \omega \in \Omega\},$$

where $\alpha$ is a constant in $(0, 1]$ that controls the strength of the low-rank signal component, $\Omega$ stands for the sample space. While the $\omega$ will be involved in the probability calculation, the deterministic parameter is $\sigma^0 \in \Sigma^0$.

When the idiosyncratic error term is vanishing, the problem to solve is to recover the potential increment matrix $\Delta = \Pi$. One proposal is to find a low-rank matrix so that the following optimization has a solution.

$$\widehat{\Pi} = \arg\min_{\Pi} \|\Pi\|_* \ \text{ subject to } \ \mathcal{A}(\Pi) = b, \tag{2}$$

**Figure 1 Asynchronous observations of two assets.**




*Note.* This figure illustrates the problem of asynchronous trading for two assets ($i$-th and $m$-th) relative to a synchronous time grid, $\mathcal{T}$. The red line in the figure frames the range of returns that are in $\mathcal{T}$ and unobservable (potential), the blue line frames the range of returns that are in $\mathcal{T}$ and observable, and the black line frames the range of all observable returns. The equations represent the linear constraints.

where $\|\cdot\|_*$ stands for the nuclear norm of some matrix. When the idiosyncratic error term is present, (2) only identifies the low-rank signal part of the potential increment matrix ignoring completing the increments inside $b$ that includes idiosyncratic errors. To simultaneously recover the low-rank component $\Pi$ and impute the missing increments in $\Delta$, we consider the following optimization problem.

$$(\widehat{\Pi}, \widehat{\Delta}) = \arg\min_{\Pi,\Delta} \|\Pi\|_* \quad \text{subject to} \quad \mathcal{A}(\Delta) = b \;\&\; \Delta = \Pi + \Pi^*. \tag{3}$$

In the context of the model (1), $\Pi$ and $\Pi^*$ in (3) correspond to the increment matrices contributed by $\sigma_t dW_t$ and $\mu_t dt + \sigma_t^* dW_t^*$, respectively.

## 2.2. Computational Issues

To solve the optimization problem (3), we propose to apply the ADMM algorithm, which is simple but powerful. The ADMM decomposes a large global optimization problem into small local subproblems, such that it can be efficient when both the dimension and the sample size of the data are large, see Scheinberg et al. (2010) and Boyd et al. (2011) for details.

**2.2.1. Scaled ADMM Algorithm** Since $\Delta = \Pi + \Pi^*$ and $\Pi^*$ can be considered as the residuals, the equivalent Lagrangian formula is

$$\min_{(\Delta,\Pi)} \frac{1}{2}\|\mathcal{A}(\Delta) - b\|_F^2 + \frac{\mu}{2}\|\Delta - \Pi\|_F^2 + \lambda\|\Pi\|_*,$$

which is further equivalent to the following relaxed form

$$\begin{aligned} \min_{(\Delta,\Pi,Z_\Delta,Z_\Pi)} \ & \frac{1}{2}\|\mathcal{A}(\Delta)-b\|_F^2+\frac{\mu}{2}\|Z_\Delta-\Pi\|_F^2+\lambda\|Z_\Pi\|_* \\ & \text{subject to} \quad \Delta=Z_\Delta,\ \Pi=Z_\Pi, \end{aligned} \tag{4}$$

where $Z_\Delta$ and $Z_\Pi$ are auxiliary variables.

To solve (4), we borrow the idea from Scheinberg et al. (2010) and propose a scaled version of the ADMM algorithm, which relies on the following augmented Lagrangian:

$$\begin{aligned} \widetilde{\mathcal{L}}(\Delta,\Pi,Z_\Delta,Z_\Pi,U_\Delta,U_\Pi) =& \frac{1}{2}\|\mathcal{A}(\Delta)-b\|_F^2+\frac{\mu}{2}\|Z_\Delta-\Pi\|_F^2+\lambda\|Z_\Pi\|_* \\ & +\frac{\eta}{2}\|Z_\Delta-\Delta+U_\Delta\|_F^2+\frac{\eta}{2}\|Z_\Pi-\Pi+U_\Pi\|_F^2, \end{aligned} \tag{5}$$

where $\eta>0$ is a penalty parameter, $U_\Delta$ and $U_\Pi$ are scaled dual variables corresponding to the constraints in (4). We then solve problem (5) by updating the unknown terms one by one. Let $(\Delta^k,\Pi^k,Z_\Delta^k,Z_\Pi^k,U_\Delta^k,U_\Pi^k)$ be the solution at step $k$, for $k=0,1,2,\cdots$. We first update $\Delta$ according to

$$\begin{aligned} \Delta^{k+1} &= \arg\min_{\Delta}\ \widetilde{\mathcal{L}}(\Delta,\Pi^k,Z_\Delta^k,Z_\Pi^k,U_\Delta^k,U_\Pi^k) \\ &= \arg\min_{\Delta}\ \{\frac{1}{2}\|\mathcal{A}(\Delta)-b\|_F^2+\frac{\eta}{2}\|Z_\Delta^k-\Delta+U_\Delta^k\|_F^2\} \\ &= \arg\min_{\Delta}\ \{\frac{1}{2}\|\mathcal{A}(\Delta)-b\|_F^2+\frac{\eta}{2}\mathrm{vec}((Z_\Delta^k-\Delta+U_\Delta^k)')'\mathrm{vec}((Z_\Delta^k-\Delta+U_\Delta^k)')\}. \end{aligned}$$

Denote $\widetilde{\Delta}=\mathrm{vec}(\Delta')$, we rewrite above result as

$$\begin{aligned} \widetilde{\Delta}^{k+1} &= \arg\min_{\widetilde{\Delta}}\ \left\{\frac{1}{2}\|A\widetilde{\Delta}-b\|_F^2+\frac{\eta}{2}\left[\mathrm{vec}((Z_\Delta^k+U_\Delta^k)')'-\widetilde{\Delta}'\right]\left[\mathrm{vec}((Z_\Delta^k+U_\Delta^k)')-\widetilde{\Delta}\right]\right\} \\ &= (A'A+\eta I)^{-1}\left[A'b+\eta\mathrm{vec}((Z_\Delta^k+U_\Delta^k)')\right], \end{aligned}$$

where the last equation can be obtained by taking the derivative of $\widetilde{\Delta}$ for the above objective function and solving the equation

$$A'A\widetilde{\Delta}-Ab'+\eta\widetilde{\Delta}-\eta\mathrm{vec}(Z_\Delta^k+U_\Delta^k)=0.$$

We then have

$$\Delta^{k+1}=\mathrm{vec}^{-1}\left((A'A+\eta I)^{-1}\left[A'b+\eta\mathrm{vec}((Z_\Delta^k+U_\Delta^k)')\right]\right)',$$

where $\mathrm{vec}^{-1}$ is the inverse vectorization (or matricization) operator. The closed form of $\Pi^{k+1}$ is

$$\Pi^{k+1}=\arg\min_{\Pi}\ \widetilde{\mathcal{L}}(\Delta^{k+1},\Pi,Z_\Delta^k,Z_\Pi^k,U_\Delta^k,U_\Pi^k)=\frac{1}{\mu+\eta}(\mu Z_\Delta^k+\eta Z_\Pi^k+\eta U_\Pi^k),$$

and similarly, for $Z_\Delta^{k+1}$ and $Z_\Pi^{k+1}$,

$$Z_\Delta^{k+1}=\frac{1}{\mu+\eta}(\mu\Pi^{k+1}+\eta\Delta^{k+1}-\eta U_\Delta^k),\quad Z_\Pi^{k+1}=\mathrm{shrink}(\Pi^{k+1}-U_\Pi^k,\frac{\lambda}{\eta}).$$

To get $Z_\Pi^{k+1}$, we have used equation (8) in Gandy et al. (2011), where for a scalar $\psi>0$, $\mathrm{shrink}(\Psi,\psi)$ is an operator that gives a soft-threshold to each singular value of the matrix $\Psi$ such that $\mathrm{shrink}(\Psi,\psi)=$

**Algorithm 1** Estimation of $\Delta$ and $\Pi$ via scaled ADMM

**Input:** Observations of log-prices: $X = \{X_{i,\tau_{i0}}, X_{i,\tau_{i1}}, \ldots, X_{i,\tau_{iin}}; i = 1, \ldots, N\}$. Initial estimates: $\Pi^0, Z_\Delta^0, Z_\Pi^0, U_\Delta^0, U_\Pi^0$ (typically initialized to be zero). Tuning parameters: $\mu, \lambda, \eta$.

**Output:** Estimated matrices $\widehat{\Delta}$ and $\widehat{\Pi}$.

Construct matrices $A_1, \ldots, A_N$ and vector $b$ from the observations of $X$.

Set iteration counter $k \leftarrow 0$.

**while** not converged **do**

► *Update primal variables*

$\Delta^{k+1} \leftarrow \text{vec}^{-1}\left((A'A + \eta I)^{-1}\left[A'b + \eta \text{vec}((Z_\Delta^k + U_\Delta^k)')\right]\right)'$

$\Pi^{k+1} \leftarrow \frac{1}{\mu+\eta}\left(\mu Z_\Delta^k + \eta Z_\Pi^k + \eta U_\Pi^k\right)$

► *Update auxiliary variables*

$Z_\Delta^{k+1} \leftarrow \frac{1}{\mu+\eta}\left(\mu \Pi^{k+1} + \eta \Delta^{k+1} - \eta U_\Delta^k\right)$

$Z_\Pi^{k+1} \leftarrow \text{shrink}\left(\Pi^{k+1} - U_\Pi^k, \frac{\lambda}{\eta}\right)$

► *Update dual variables*

$U_\Delta^{k+1} \leftarrow U_\Delta^k + Z_\Delta^{k+1} - \Delta^{k+1}$

$U_\Pi^{k+1} \leftarrow U_\Pi^k + Z_\Pi^{k+1} - \Pi^{k+1}$

$k \leftarrow k + 1$

**end while**

Set $\widehat{\Delta} \leftarrow \Delta^k$ and $\widehat{\Pi} \leftarrow \Pi^k$.

$U\,\text{diag}(\max(\rho_1 - \psi, 0), \cdots, \max(\rho_N - \psi, 0))V^*$ with $U\text{diag}(\rho_1, \cdots, \rho_N)V^*$ to be the singular value decomposition of $\Psi$. For convenience, we set

$$U_\Delta^{k+1} = U_\Delta^k + Z_\Delta^{k+1} - \Delta^{k+1}, \quad U_\Pi^{k+1} = U_\Pi^k + Z_\Pi^{k+1} - \Pi^{k+1}.$$

The computational steps are summarized in Algorithm 1.

Remark 1 (Computational Efficiency and Implementation). The main computational bottleneck in Algorithm 1 is the inversion of the matrix $(A'A + \eta I)^{-1}$, as $A'A$ can be a huge, non-diagonal matrix. To maintain computational efficiency, we leverage the Woodbury matrix identity:

$$(A'A + \eta I)^{-1} = \eta^{-1} I - \eta^{-1} I A' \left[I + A\eta^{-1} I A'\right]^{-1} A\eta^{-1} I = \eta^{-1} I - \eta^{-2} A' \left[I^{-1} + \eta^{-1} A A'\right]^{-1} A,$$

as calculating the inverse of the diagonal matrix $I + AA'/\eta$ is fast. In addition, since we have all the closed forms of $\Delta^{k+1}, \Pi^{k+1}, Z_\Delta^{k+1}, Z_\Pi^{k+1}, U_\Delta^{k+1}$, and $U_\Pi^{k+1}$, the Algorithm 1 is not time-consuming even if the sample size and dimension of $X$ are high. In practice, the algorithm is considered to have converged when the following condition is satisfied for a tolerance level of $\epsilon = 10^{-5}$:

$$\max\left\{\frac{\|\Delta^{k+1} - \Delta^k\|_F}{\max\left(1, \|\Delta^k\|_F, \|\Delta^{k+1}\|_F\right)}, \frac{\|\Pi^{k+1} - \Pi^k\|_F}{\max\left(1, \|\Pi^k\|_F, \|\Pi^{k+1}\|_F\right)}, \frac{\|\Delta^k - Z_\Delta^k\|_F}{\max\left(1, \|\Delta^k\|_F, \|Z_\Delta^k\|_F\right)}, \frac{\|\Pi^k - Z_\Pi^k\|_F}{\max\left(1, \|\Pi^k\|_F, \|Z_\Pi^k\|_F\right)}\right\} < \epsilon.$$

**2.2.2. Choices of Initial Estimates and Tuning Parameters** The above algorithm is not sensitive to the initial estimates of $\Pi$, $Z_\Delta$, $Z_\Pi$, $U_\Delta$, $U_\Pi$. We can set them to the zero matrices. First of all, to improve the convergence speed of the algorithm, we can use the previous tick method to get a full observed $X$, which means we let $X_{it_1}$ to be the first observation of the $i$th asset and $X_{it_j} = X_{it_{j-1}}$ if the $i$th asset has no observation at time $t_j$. We then set the initial values of $Z_\Delta$ to be the first-order difference of the initial estimate of $X$ and run the PCA to get the initial estimate of $\Pi$. The initial estimate of $U_\Delta$ and $U_\Pi$ can be set to zero matrices.

To select the optimal tuning parameter ($\mu$, $\lambda$, $\eta$), we propose a data-driven validation scheme based on artificial masking. The procedure aims to find the parameter that minimizes imputation error on a held-out portion of the data. Our scheme operates as follows. Given the input log-price matrix $\mathcal{P}$, a set of candidate masking probabilities $\{p_1, p_2, .., p_m\}$, and a number of repetitions $Q$, we sequentially do the following. (i) Mask generation: For each probability $p_i$ and each repetition $j \in \{1, ..., Q\}$, we generate a binary mask matrix $\mathcal{M}_{ij}$. This is done by first identifying all observable entries in $\mathcal{P}$ and then randomly selecting a fraction $p_i$ of them to be masked. In $\mathcal{M}_{ij}$, these masked positions are marked with 1, and all others with 0. (ii) Imputation: The Algorithm 1 is then applied onto the partially observed data (the entries of $\mathcal{P} \circ \mathcal{M}_{ij}$) to produce a completed matrix $\hat{\mathcal{P}}_{ij}(\mu, \lambda, \eta)$. (iii) Error calculation: We then calculate the imputation error between the imputed matrix $\hat{\mathcal{P}}_{ij}(\mu, \lambda, \eta)$ and the original matrix $\mathcal{P}$, but only on the set of entries that were artificially masked (corresponding position of $\mathcal{M}_{ij}$ is 1).

The optimal parameter is chosen as the one that yields the lowest average imputation error across all masks and repetitions. We evaluate the imputation accuracy using the following two error metrics:

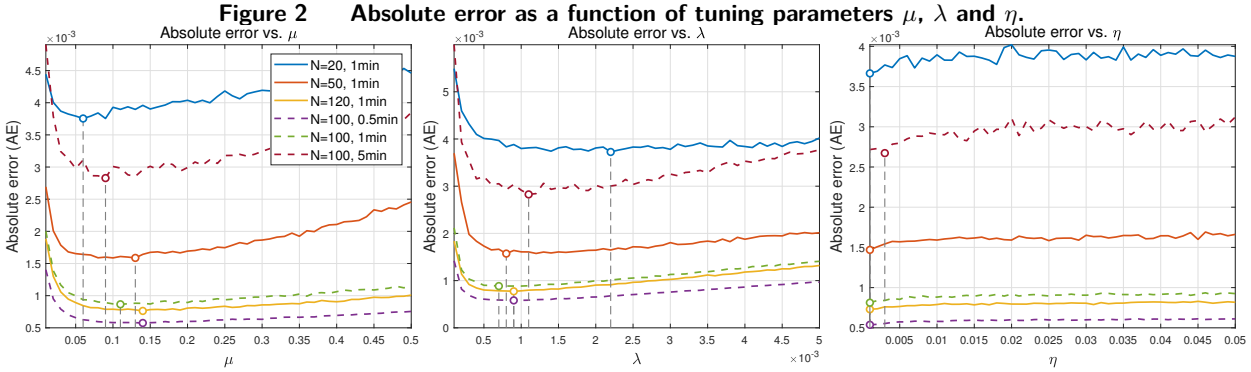
$$\begin{aligned} \mathcal{R}^{\text{absolute}}(\mu, \lambda, \eta) :=& \frac{1}{mQ} \sum_{i=1}^{m} \sum_{j=1}^{Q} \frac{\|(\hat{\mathcal{P}}_{ij}(\mu, \lambda, \eta) - \mathcal{P}) \circ \mathcal{M}_{ij}\|_F}{\|\mathcal{M}_{ij}\|_F}, \\ \mathcal{R}^{\text{relative}}(\mu, \lambda, \eta) :=& \frac{1}{mQ} \sum_{i=1}^{m} \sum_{j=1}^{Q} \frac{\|(\hat{\mathcal{P}}_{ij}(\mu, \lambda, \eta) - \mathcal{P}) \circ \mathcal{M}_{ij}\|_F}{\|\mathcal{P} \circ \mathcal{M}_{ij}\|_F}, \end{aligned} \tag{6}$$

where $\circ$ denotes the Hadamard product. The Frobenius norm $\|\cdot\|_F$ in the numerator is calculated only over the set of masked entries.

To demonstrate the effect of these tuning parameters, we conduct the validation procedure using the following candidate sets: $\mu \in \{0.01, 0.02, ..., 0.5\}$, $\lambda \in \{0.0001, 0.0002, ..., 0.005\}$, $\eta \in \{0.001, 0.002, ..., 0.05\}$. The set of masking probabilities is $\{0.1, 0.2, ..., 0.7\}$, and the number of repetitions is set to $Q = 1$. Figure 2 illustrates the impact of each tuning parameter on the absolute error. The corresponding results for the relative error are provided in the Supplementary Appendix.

Figure 2 illustrates the sensitivity of the absolute error to the tuning parameters $\mu$, $\lambda$, and $\eta$. Several key observations guide our parameter selection strategy. First, the estimation error appears insensitive

**Figure 2 Absolute error as a function of tuning parameters $\mu$, $\lambda$ and $\eta$.**




*Note.* The absolute error is calculated according to (6). The vertical dotted line in each panel indicates the parameter value that minimizes the error.

to the specific value of $\eta$. Based on this observation and to enhance computational efficiency, we simplify the three-dimensional search by fixing the ratio between the two parameters, setting $\lambda/\eta = 0.1$. Second, the figure shows that the optimal values for $\mu$ and $\lambda$ are quite stable, consistently falling within a relatively narrow range. This suggests that the optimal parameter configuration is robust to the specific characteristics of the dataset (such as the number of assets or observation frequency, as implicitly varied across the panels). This stability allows us to streamline the tuning process in practice. Instead of searching over a wide grid, we can focus on more refined candidate sets, $\mathbb{M}$ and $\mathbb{L}$, for $\mu$ and $\lambda$, respectively. Formally, the optimal parameters $(\hat{\mu}, \hat{\lambda})$ can be determined by solving the following optimization problem:

$$(\hat{\mu}, \hat{\lambda}) = \underset{\mu \in \mathbb{M}, \lambda \in \mathbb{L}}{\arg\min}\, \mathcal{R}^{\text{absolute}}(\mu, \lambda, 10\lambda).$$

## 2.3. Technical Assumptions

To connect the matrix parameter $\Pi$ to the generative semi-martingale, one can relate $\Pi$ to $\sigma^0$ and $\Sigma_t$ so that the varying of $\Pi$ is caused by the smooth change of $\sigma^0$ and $\Sigma_t$ for every sample path $\omega \in \Omega$. That is the parameter $\Pi$ is simply a potential increment matrix of some regular semi-martingale for some volatility parameter given $\omega$. Thus we impose the technical conditions on the dynamics of the generative model (1) and the durations that realizes the operator $\mathcal{A}$.

The first assumption is a regular condition for the durations of the coordinate processes. Before stating it, we introduce one more notation. For a generic process $Y_t \in R^d$, let

$$d_{ik} := \sharp\{j; \tau_{i(l-1)} \le t_j \le t_{k-1} \le \tau_{il} \text{ for some } 1 \le l \le n_i\}$$

to be the number of potential increments before $\Delta_k^n Y := Y_{t_j} - Y_{t_{j-1}}$ in the observed interval $(\tau_{i(l-1)}, \tau_{il}]$ which contains the time point $t_{k-1}$. We make a convention that $d_{ik} := 0$ if the set is empty and the resulting sum $\sum_{l=1}^{0} \Delta_{k-l}^n Y = 0$.

ASSUMPTION 1. 1. *There exists positive random variables $R_j$'s independent of $\mathcal{F}$, such that $t_j - t_{j-1} \le R_j$, $\sum_{j=1}^n R_j \le C$, $\max_j R_j = o(1)$, and $(\sum_{j=1}^n R_j)^{3/4}/\max_j R_j \ge c$ for some $c > 0$.*

2. *The maximum diagonal element of the diagonal matrix $AA'$ satisfies*

$$P\left(\|AA'\|_\infty > 2\left(1+\sqrt{Nn/n^*}\right)\right) \le \exp\{-\gamma Nn\},$$

*where $n^* = \sum_{i=1}^N n_i$ and $\gamma$ is a positive constant.*

3. *There exists a sequence of numbers $\{D_{ik}\}$ that are independent of $\mathcal{F}$ and satisfy $d_{ik} \le D_{ik}$.*

Assumption 1-1 allows for random and unequally spaced sampling times. Though the upper bound $R_j$'s are assumed to be independent of the process $X_t$, the coordinate durations can be stopping times that are dependent on $X_t$. Assumption 1-2 regularizes the maximum number of potential increments missed in an observed duration. A simple example is that $\|AA'\|_\infty$ is bounded. In this case, the probability upper bound is zero for large enough $N$ and $n$ once $\sqrt{Nn/n^*} \to \infty$, and $t_j - t_{j-1}$'s are of order $1/n$ and all conditions in Assumption 1 are satisfied.

The next assumption assumes that the systematic and idiosyncratic spot volatility processes are locally bounded, and the correlation matrix $\rho^*$ is sparse in some sense.

ASSUMPTION 2. 1. *There exists a sequence of stopping times $s_m \to \infty$ and a sequence of numbers $c_m$, such that $\max_{i\le N}(|\sigma_{i(t\wedge s_m)}| + |\sigma^*_{i(t\wedge s_m)}|) \le c_m$.*

2. $\max_{i\le N}\sum_{j=1}^N |\rho^*_{ij}| \le C$.

3. $\|IV_1\|_F^2 := \|\sum_0^T \Sigma_t\Sigma_t' dt\|_F^2 \ge cr$ *for some* $c > 0$.

4. $|\sigma^*_{i(t+h)} - \sigma_{it}| \le Ch^{1/2-\epsilon}$ *for some constant* $C$ *and arbitrary constant* $\epsilon$.

## 3. Main Theoretical Results

The first theoretical result gives the nearly-isometry property for $\mathcal{A}$ operating on matrix generated from large-dimensional semi-martingales. Let $P_{R,D}$ be a probability measure conditional on $\{R_j, D_{ik}; i \le N, j \le n, k \le n\}$.

THEOREM 1. *Let the quantities $\overline{L}_1 = \sum_{k=1}^n R_k \sum_{i=1}^N \|\sigma^0(i,\cdot)\|^2(\sum_{l=1}^{D_{ik}} R_{k-l})\log(\sum_{l=1}^{D_{ik}} R_{k-l})$ and $\overline{L}_2 = \sum_{k=1}^n R_k \sum_{i=1}^N (\sum_{l=1}^{D_{ik}} R_{k-l})\log(\sum_{l=1}^{D_{ik}} R_{k-l})$, where $\sigma^0(i,\cdot)$ is the $i$-th row of $\sigma^0$. For some small $c > 0$,*

$$P_{R,D}\left(\left|\frac{\|\mathcal{A}(\Pi)\|_F^2 - \|\Pi\|_F^2}{\|\Pi\|_F^2}\right| > \delta\right) \le C\exp\left\{-c^2\delta^2 N^\alpha/\overline{L}_1\right\} + 2\exp\left\{-c^2r^2/\left(\sum_{j=1}^n R_j^2\right)^{1/2}\right\}, \quad (7)$$

*and*

$$P_{R,D}\left(\left|\frac{\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2}{\|\Delta\|_F^2}\right| > \delta\right)$$

$$\leq 2\exp\left\{-\frac{N^{2-\alpha}c^2\delta^2}{144Cr^2\overline{L}_1}\right\}+2\exp\left\{-\frac{N^2c^2\delta^2}{144C\overline{L}_2}\right\}+2N\exp\left\{-\frac{c}{24C\max_j R_j^{2-\epsilon}}\right\} \tag{8}$$
$$+2\exp\left\{-c^2N^{2-2\alpha}/\left(128C^2\sum_{j=1}^{n}R_j^2\right)\right\}+2\exp\left\{-\frac{Nc\epsilon^*}{6\sqrt{2C}\max_j R_j}\right\}.$$

REMARK 2 (INTERPRETATION OF $\overline{L}_1$ AND $\overline{L}_2$). The number $\overline{L}_2/N$ is an approximate measure of the average length of durations across all assets within the time window. As an example when the diagonal entries of $AA'$ are bounded so that $t_j - t_{j-1} = O(1/n)$, $\overline{L}_2$ is upper bounded by $\max_k \sum_{i=1}^{N} D_{ik}\log(n)/n \leq CN\log(n)/n$. In particular, if the vector $(D_{1k},\cdots,D_{Nk})'$ is sparse when the asynchronicity is rare, the upper bound could be of smaller order than $O(N\log(n)/n)$. Then the reciprocal $N/\overline{L}_2$ is approximately equal to the average number of observed durations across all assets, which is larger than the smallest sample size of the most sparsely sampled asset. This implies that our approach makes use of all the data points through the large-scale linear system, in constrast with the tick subsampling method listed in previous sections which depends much on the sample size of most illiquid assets, thus loss of efficiency in estimation and subsequent applications. The number $\overline{L}_1$ is a cross-sectionally weighted version of $\overline{L}_2$ and hence has a similar interpretation as $\overline{L}_2$.

Before we prove a uniform result on $|\|\mathcal{A}(\Pi)\|_F^2 - \|\Pi\|_F^2|/\|\Pi\|_F^2$ and $|\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2|/\|\Delta\|_F^2$ over the set of matrices $\Pi$ and $\Delta$, we introduce some facts of the Grassmannian manifold. The set of all $d$-dimensional subspaces of $R^D$ is known as the Grassmannian manifold which is denoted by $\mathcal{B}(D,d)$. First of all, let $U$ be an arbitrary subspace of $N\times n$ matrices with dimension $r \leq N\wedge n$, then there exists a finite set $\Omega$ of at most $(12/\delta)^r$ points such that for every $\Pi\in U$ with $\|\Pi\|_F^2/N^\alpha \leq 1$, there exists a $Q\in\Omega$ such that $\|\Pi - Q\|_F/N^{\alpha/2} \leq \delta/4$. Define the natural distance between two subspaces by $\rho(T_1,T_2) := \|P_{T_1} - P_{T_2}\|$, where $T_1$ and $T_2$ are subspaces and $P_{T_i}$ is the orthogonal projection associated with each subspace. This equals to the sine of the largest principal angle between $T_1$ and $T_2$. As demonstrated by the work of Szarek on $\epsilon$-nets of the Grassmannian, c.f., Szarek (1983), the covering number of $\mathcal{B}(N,r)$ at resolution $\epsilon$ (i.e. the smallest number of subspaces $U_i$ such that for any subspace $U$, there is an $i$ with $\rho(U,U_i)\leq\epsilon$) is at most $(\frac{2C_0}{\epsilon})^{r(N-r)}$ where $C_0$ is a constant independent of $\epsilon$, $N$ and $r$.

By the union bound and Theorem 1, we have the following theorem.

THEOREM 2. *Suppose that Assumptions 1-2 hold. If*

$$r(N-r)\log\left(\sqrt{\frac{Nn}{n^*}}+1\right) = o\left\{\frac{N^\alpha}{\overline{L}_1}\right\} \text{ and } N^\alpha/\overline{L}_1 \to \infty,$$

*then for any* $0<\delta<1$,

$$\begin{aligned} &P_{R,D}\left(|\frac{\|\mathcal{A}(\Pi)\|_F^2-\|\Pi\|_F^2}{\|\Pi\|_F^2}|>\delta,\ \textit{for some } \sigma^0\in\mathcal{B}(N,r)\right)\\ &\leq C\exp\{-c^2\delta^2N^\alpha/\overline{L}_1\}+2\exp\left\{-c^2r^2/\left(\sum_{j=1}^n R_j^2\right)^{1/2}\right\}, \end{aligned} \tag{9}$$

*for* $c$ *small enough and* $C$ *large enough which might depend on* $\delta$.

*If*

$$r(N-r)\log\left(\sqrt{\frac{Nn}{n^*}}+1\right)=o\left\{\frac{N^{2-\alpha}}{\overline{L}_1}\right\}\ \textit{and}\ N^{2-\alpha}/\overline{L}_1\to\infty,$$

*then for any* $0<\delta<1$,

$$\begin{aligned} &P_{R,D}\left(|\frac{\|\mathcal{A}(\Delta)\|_F^2-\|\Delta\|_F^2}{\|\Delta\|_F^2}|>\delta,\ \textit{for some } \sigma^0\in\mathcal{B}(N,r)\right)\\ &\leq 2\exp\left\{-\frac{N^{2-\alpha}c^2\delta^2}{144Cr^2\overline{L}_1}\right\}+2\exp\left\{-\frac{N^2c^2\delta^2}{144C\overline{L}_2}\right\}+2\exp\left\{-\frac{Nc\epsilon^*}{6\sqrt{2C}\max_j R_j}\right\}\\ &+2\exp\left\{-c^2N^{2-2\alpha}/\left(128C^2\sum_{j=1}^n R_j^2\right)\right\}+2N\exp\left\{-\frac{c}{24C\max_j R_j^{2-\epsilon}}\right\}. \end{aligned} \tag{10}$$

*Let* $\Pi_0$ *be a matrix of rank less than or equal to* $r$ *and satisfy the constraints in (3). The problem (3) has a solution for* $\Pi$ *that equals to* $\Pi_0$ *with probability* $(P_{R,D})$ *at least*

$$1-C\exp\left\{-c^2\delta^2N^\alpha/\overline{L}_1\right\}-2\exp\left\{-c^2r^2/\left(\sum_{j=1}^n R_j^2\right)^{1/2}\right\}.$$

*Moreover, Let* $\Delta_0$ *be a matrix satisfying the constraints in (3). The problem (3) has a solution for* $\Delta$ *that equals to* $\Delta_0$ *with probability* $(P_{R,D})$ *at least*

$$\begin{aligned} &1-2\exp\left\{-\frac{N^{2-\alpha}c^2\delta^2}{144Cr^2\overline{L}_1}\right\}-2\exp\left\{-\frac{N^2c^2\delta^2}{144C\overline{L}_2}\right\}-2\exp\left\{-\frac{Nc\epsilon^*}{6\sqrt{2C}\max_j R_j}\right\}\\ &-2\exp\left\{-c^2N^{2-2\alpha}/\left(128C^2\sum_{j=1}^n R_j^2\right)\right\}-2N\exp\left\{-\frac{c}{24C\max_j R_j^{2-\epsilon}}\right\}. \end{aligned} \tag{11}$$

Remark 3 (Dependence on Data Frequency). Similar to Theorem 1, an interesting finding is that the uniform rate depends also on the average length of the observed durations which is different from the previous-tick approach in large-panel high-frequency data analysis literature. For the previous-tick method, the statistically efficiency rests on the length of the time lags of the most illiquid asset where data comes with the lowest frequency.

Theorem 2 implies that

$$\sup_{\sigma^0\in\mathcal{B}(N,r)}|\|\mathcal{A}(\Delta)\|_F^2-\|\Delta\|_F^2|=O_p\left(N\sqrt{\overline{L}_1/N^{2-\alpha}}\right)=:O_p(Na_{Nn}). \tag{12}$$

Let

$$G_n(\Pi,\Delta) := \frac{1}{2}(\|\mathcal{A}(\Delta)-b\|_F^2) + \frac{1}{2}\mu\|\Delta-\Pi\|_F^2 + \lambda\|\Pi\|_*,$$

and

$$G_{n0}(\Pi,\Delta) := \frac{1}{2}\|\Delta-\Delta_0\|_F^2 + \frac{1}{2}\mu\|\Delta-\Pi\|_F^2 + \lambda\|\Pi\|_*.$$

(12) shows that $G_n(\Pi,\Delta)$ and $G_{n0}(\Pi,\Delta)$ are uniformly close enough. The minimizers of $G_n(\Pi,\Delta)$ and $G_{n0}(\Pi,\Delta)$ are denoted by $(\widehat{\Pi},\widehat{\Delta})$ and $(\widehat{\Pi}_0,\widehat{\Delta}_0)$, respectively. The next theorem shows how close are the two minimizers.

THEOREM 3. *Under Assumptions 1-2,* $\|\widehat{\Pi}-\widehat{\Pi}_0\|_F/\sqrt{N} = O_p\left(\sqrt{\frac{a_{Nn}}{1+\lambda}}\right)$ *and* $\|\widehat{\Delta}-\widehat{\Delta}_0\|_F/\sqrt{N} = O_p\left(\sqrt{\frac{a_{Nn}}{1+\lambda}}\right)$.

The minimizer of $G_{n0}(\Pi,\Delta)$ satisfies the following first order condition.

$$(1+\mu)\widehat{\Delta}_0 = \Delta_0 + \mu\widehat{\Pi}_0,\ \widehat{\Pi}_0 = U_0 D_{\lambda/\mu+} V_0', \tag{13}$$

where $D_{v+}$ always denotes a thresholded singular value matrix whose $j$-th singular value is $\max\{\lambda_j - v, 0\}$ with $\lambda_j$ being the $j$th largest singular value of $\widehat{\Delta}_0$ (the $v$ in the above equation is $\lambda/\mu$), and $U_0$ and $V_0$ are the left and right singular vectors of $\widehat{\Delta}_0$, respectively. Let $D$ be the the matrix of singular values of $\widehat{\Delta}_0$. Combining the two conditions in (13) leads to the following equation

$$U_0\left(D - \frac{\mu}{1+\mu}D_{\lambda/\mu+}\right)V_0' = \frac{1}{1+\mu}\Delta_0.$$

A solution to this equation is $D - \frac{\mu}{1+\mu}D_{\lambda/\mu+} = \frac{1}{1+\mu}D_*$ where $D_*$ is the matrix of singular values of $\Delta_0$. That being said,

$$\widehat{\Delta}_0 = U_*\{(D_*)_{\lambda+} + D_{**}\}V_*' \text{ and } \widehat{\Pi}_0 = U_*\{(D_*)_{\lambda+} + D_{**}\}_{\lambda/\mu+}V_*',$$

where $U_*$ and $V_*$ are the left and right singular vector matrices of $\Delta_0$, respectively, and

$$D_{**} = \frac{1}{1+\mu}\left\{D_* - \left[(D_*)_{\lambda+} + \lambda\begin{pmatrix} I_k & 0 \\ 0 & 0\end{pmatrix}\right]\right\},$$

with $k$ being the number of singular values of $D_*$ greater than or equal to $\lambda$. Let $J$ be the number of singular values of $D_*$ greater than or equal to $\lambda(1+1/\mu)$. $\widehat{\Delta}_0$ shrinks the first $k$ singular values by subtracting $\lambda$ from them and scaling down the remaining singular values with a factor of $1/(1+\mu)$. $\widehat{\Pi}_0$ shrinks the first $J$ singular values of $D_*$ by subtracting $\lambda(1+1/\mu)$ from them and truncate the remaining singular values to zero. The shrinkage of the largest singular values for $\widehat{\Pi}_0$ is more than that for $\widehat{\Delta}_0$ because the former is of low rank but the latter is not, and thus $J \geq k$. When $\mu = 0$, $\widehat{\Pi}_0 = U_*(D_*)_{\lambda+}V_*'$ and $\widehat{\Delta}_0 = \Delta_0$, which boils down to the solution to (2). Notice here that $\Delta_0$ as a true potential increment matrix is unknown and has to be numerically solved and approximated

by $\widehat{\Delta}$, but this is not achievable when setting $\mu = 0$ in the optimization. Also notice the difference between $(\widehat{\Pi}_0, \widehat{\Delta}_0)$ and $(\widehat{\Pi}, \widehat{\Delta})$. The latter is the computationally feasible version while the former is a theoretical approximation that is easy to analyze mathematically. Theorem 3 demonstrates that they are close in terms of the averaged Frobenius norm.

Let $\|\cdot\|$ be the operator norm of a matrix. Next, we analyze the closeness of $(\widehat{\Pi}, \widehat{\Delta})$ to $(\Pi_0, \Delta_0)$.

THEOREM 4. *Assume that* $\Pi_0 = U_{*r} diag\{\lambda_1^*, ..., \lambda_r^*\} V_{*r}'$ *where* $U_{*r}$ *and* $V_{*r}$ *are matrices consisting of* $r$ *left and right singular vectors of* $\Delta_0$ *corresponding to* $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_r^*$*, respectively. Under Assumptions 1-2,*

$$\|\widehat{\Delta} - \Delta_0\| \leq \lambda \vee \frac{\mu}{1+\mu}\lambda_{k+1}^* + \sqrt{N a_{Nn}/(1+\lambda)},$$
$$\|\widehat{\Delta} - \Delta_0\|_F \leq \lambda k \vee \frac{\mu}{1+\mu}\sum_{j=k+1}^{N\wedge n} \lambda_j^* + \sqrt{N a_{Nn}/(1+\lambda)},$$

*and*

$$\|\widehat{\Pi} - \Pi_0\| \leq \lambda(1+\frac{1}{\mu}) \vee \left[\left(\lambda_{r+1}^* - \lambda\left(1+\frac{1}{\mu}\right)\right) I(J>r) + \lambda_{J+1}^* I(J \leq r)\right] + \sqrt{N\frac{a_{Nn}}{1+\lambda}},$$
$$\|\widehat{\Pi} - \Pi_0\|_F \leq \left[r\lambda\left(1+\frac{1}{\mu}\right) + (J-r)\left(\lambda_{r+l}^* - \lambda\left(1+\frac{1}{\mu}\right)\right)\right] I(J>r)$$
$$+[J\lambda(1+1/\mu) + (r-J)\lambda_{J+1}^*] I(J \leq r) + \sqrt{N\frac{a_{Nn}}{1+\lambda}},$$

*with probability approaching one.*

Theorem 4 gives the convergence rates of the estimators $\widehat{\Delta}$ and $\widehat{\Pi}$. In each upper bound, the first term is a bias due to the shrinkage brought by the low-rank penalty, while the last term comes from the data synchronization or approximation error of $\|\mathcal{A}(\Delta)\|$ by $\|\Delta\|_F$. In estimating $\Delta_0$, setting $\lambda = \mu = 0$ is optimal to reduce the bias, but large value of $\lambda$ decreases the approximation error. In estimating $\Pi_0$, small value of $\mu$ and large value of $\lambda$ enlarge the bias term $\lambda(1+1/\mu)$ but decrease the bias term $\lambda_{r+1}^* - \lambda(1+1/\mu)$ when $J > r$. So there is a tradeoff between the bias terms, between the bias and approximation error, and between estimating $\Pi_0$ and $\Delta_0$.

To relieve the complexity in choosing the tuning parameter, we introduce a bias-corrected version of the estimators. Given $\widehat{\Delta}$ at hand, we do the SVD, and correct the singular value matrix by adding $\lambda$ onto its first $k$ (assume a priori known first) largest singular value and multipling $1+\mu$ onto the remaining singular values. We denote the resulting estimator by $\widetilde{\Delta}$. Given the $\widehat{\Pi}$, we do the SVD also, and correct the singular value matrix by adding $\lambda(1+1/\mu)$ onto its first $J$ (again known a priori first) singular values. We denote the resulting estimator by $\widetilde{\Pi}$. Now, we have the following asymptotic results for the de-biased estimators.

THEOREM 5. *Under Assumptions 1-2,*

$$\|\widetilde{\Delta}-\Delta_0\|/\sqrt{N}=O_p\left(\left(\lambda\vee\frac{\mu}{1+\mu}\lambda^*_{k+1}+1\right)\sqrt{a_{Nn}/(1+\lambda)}\right),$$
$$\|\widetilde{\Delta}-\Delta_0\|_F/\sqrt{N}=O_p\left(\left(\lambda\vee\frac{\mu}{1+\mu}\lambda^*_{k+1}+1\right)\sqrt{a_{Nn}/(1+\lambda)}\right),$$

*and*

$$\frac{\|\widetilde{\Pi}-\Pi_0\|}{\sqrt{N}}=O_p\left\{\left[\lambda\left(1+\frac{1}{\mu}\right)+1\right]\sqrt{\frac{a_{Nn}}{1+\lambda}}+\frac{\lambda^*_{J+1}I(J<r)+\lambda^*_{r+1}I(J>r)+0I(J=r)}{\sqrt{N}}\right\},$$
$$\frac{\|\widetilde{\Pi}-\Pi_0\|_F}{\sqrt{N}}=O_p\left\{\left[\lambda\left(1+\frac{1}{\mu}\right)+1\right]\sqrt{\frac{a_{Nn}}{1+\lambda}}+\frac{\sum_{l=J+1}^{r}\lambda^*_l I(J<r)+\sum_{l=r+1}^{J}\lambda^*_{r+1}I(J>r)+0I(J=r)}{\sqrt{N}}\right\}.$$

Theorem 5 shows that as in the typical case where $\sqrt{\frac{a_{Nn}}{1+\lambda}}=o(1)$, the de-biased estimator decreases the bias with a downward scaling of $\sqrt{\frac{a_{Nn}}{1+\lambda}}$ for $\Delta_0$. As for $\Pi$, a good choice is setting $J=r$ (or equivalently tuning $\lambda(1+1/\mu)$). In this case, the convergence rate for $\widetilde{\Pi}$ simplifies to $\lambda(1+1/\mu)\sqrt{\frac{a_{Nn}}{1+\lambda}}$ in both averaged operator norm or Frobenius norm, and the bias is also scaled down by a factor of $\sqrt{\frac{a_{Nn}}{1+\lambda}}$. To further understand the choice of $J=r$, we separate $\Delta$ into the sum of $\Delta_1+\Delta_2$ where $\Delta_1$ is the potential increment matrix attributed by low-rank diffusion $\sigma_t dW_t$ and $\Delta_2$ is that attributed by $\mu_t dt+\sigma^*_t dW^*_t$. In traditional high-dimensional factor analysis, $\|E(\Delta_2\Delta_2')\|$ is bounded by $C_{\text{noise}}$ while $\|E(\Delta_1\Delta_1')\|\geq crN^\alpha>C_{\text{noise}}$ for large $N$ and $n$ (recall the parameter space $\Theta$). Then choosing $\lambda(1+1/\mu)$ so that $J=r$ is selecting a threshold in between $C_{\text{noise}}$ and $crN^\alpha$ to differentiate the low-rank component and the idiosyncratic component. As an end of this section, it is worthy of notice that our results hold for weak factor cases where $\alpha<1$, that being said the strong factor condition is not necessary.

## 4. Monte Carlo Simulation

In this section, we use Monte Carlo simulations to demonstrate the effectiveness of our methodology.

### 4.1. Simulation Settings

We adopt a stochastic volatility model without jumps, similar to the framework in Kong et al. (2023). The latent log-price process, $X_t$, and the latent factor process, $V_t$, are specified as follows:

$$dX_t=\beta_t dV_t+dZ_t\quad\text{and}\quad dV_t=\mu_v dt+\sigma_{v,t}\rho^{1/2}dW^v_t,$$

where $Z_t$ is an idiosyncratic error process and $W^v_t$ is a multivariate ($r$-dimensional) Brownian motion. The matrix $\rho=(\rho_{ij})_{r\times r}=\{\text{diag}(H)\}^{-1/2}HH'\{\text{diag}(H)\}^{-1/2}$, where $H=(h_{ij})_{r\times r}$ is a lower triangular matrix with elements $h_{ij}=0.6^{|i-j|}$ for $i\geq j$. The diagonal volatility matrix, $\sigma_{v,t}=\text{diag}(\sigma_{v,t1},...,\sigma_{v,tr})$, contains individual factor volatilities, each following a Heston-type square-root process:

$$d\sigma^2_{v,tl}=\kappa(\bar{\sigma}^2-\sigma^2_{v,tl})dt+s\sqrt{\sigma^2_{v,tl}}d\bar{W}_{v,tl},$$

where $\bar{W}_{v,t} = (\bar{W}_{v,t1}, ..., \bar{W}_{v,tr})'$ is a multivariate Brownian motion independent of $W_t^v$, with correlation matrix $I_r$. The initial value for each variance process, $\sigma^2_{v,0l}$, is drawn from a uniform distribution $U[0.8\bar{\sigma}^2, 1.2\bar{\sigma}^2]$. We set $r = 3$, $(\kappa, \bar{\sigma}^2, s) = (3, 0.3^2, 0.3)$ and $\mu_v = (0.05, 0.03, 0.02)'$.

The process $Z_t = \sqrt{\mu}\sigma_t^* dW_t^*$, where $\sigma_{tj}^{*2}$ is generated by the same procedure as $\sigma_{tl}^2$ with the same parameters, and $W_t^*$ is a $N$-dimensional Brownian motion with the correlation matrix $\rho^*$ being a block diagonal matrix with each block being $\left(\tilde{\rho}_{ij}^* = 0.6^{|i-j|}\right)_{10\times 10}$. This setting is similar to that in Ait-Sahalia and Xiu (2017). We let $\mu = 0.1$ such that the averaged standard errors of all elements of $Z_t$ is around 18% of that of $Y_t$.

The factor loading matrix $\beta_t = \widetilde{\beta}^0 M$ where $M = (\widetilde{\beta}^{0\prime}\widetilde{\beta}^0/p^\alpha)^{-1/2}$ such that $\beta_t'\beta_t/p^\alpha = I_p$, and $\widetilde{\beta}^0 = \left(\tilde{\beta}_j^0(l)\right)_{j=1,...,N}^{l=1,...,r}$ with $\tilde{\beta}_j^0(1) \sim U[0.25, 1.75]$ associated with the market factor, and $\tilde{\beta}_j^0(l) \sim N(0, 0.5^2)$ for $l = 2, ..., r$. In the base case, the factor strength $\alpha$ is set to 1.

Our simulation is configured with $N = 100$ assets over a time horizon of $T = 5$ trading days. Each day is discretized into a grid of $n = 390$ equally-spaced intervals, corresponding to a 1-minute frequency with a time step of $t_j - t_{j-1} = \delta_n = 1/390$. Asynchronous trading times for each asset are then generated from a standard Poisson process with an arrival intensity of $\lambda_{\text{asy}} = 1$. For our proposed algorithm, the tuning parameters are initialized to $\lambda = 0.001$, $\mu = 0.1$, and $\eta = 0.01$.

To assess the robustness of our method, we systematically vary key parameters of the simulation, keeping all other settings fixed according to the baseline configuration described above. We consider four scenarios:

(i) **Number of Assets**: We set the number of assets $N$ to 20, 50, 120.

(ii) **Observation Frequency**: We vary the length of the observation interval $\delta_n$, setting it to 1/780 (30-second), 1/390 (1-minute), and 1/78 (5-minute).

(iii) **Asynchronous Intensity**: We divide the $N = 100$ assets into two equally-sized groups and assign different Poisson arrival intensities, ($\lambda_{\text{asy},1}$, $\lambda_{\text{asy},2}$), to each. We test three configurations: uniform low intensity (0.5, 0.5), mixed intensity (0.5, 3), and uniform high intensity (3, 3).

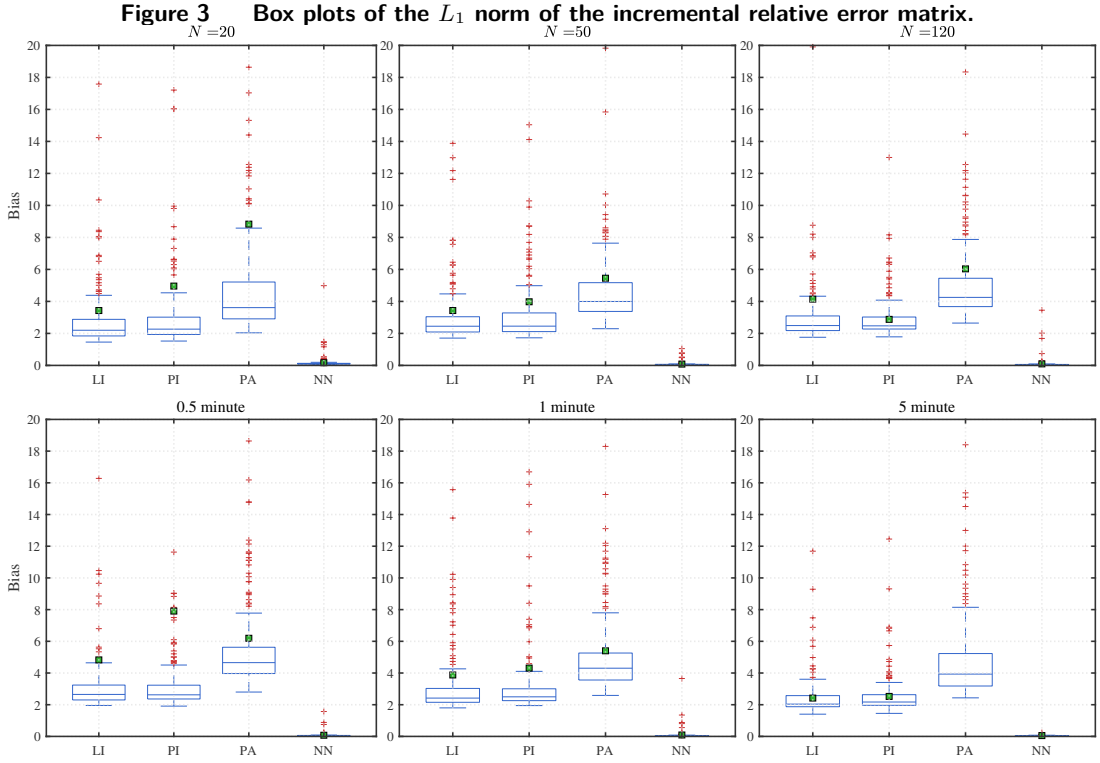
(iv) **Factor Strength**: We examine the impact of different factor strengths by setting $\alpha$ to 0.8 (strong), 0.5 (moderate), and 0.2 (weak).

We conduct 200 Monte Carlo replications for each Data Generating Process (DGP) to evaluate the finite-sample performance of our proposed Nuclear Norm (NN) method. We compare it against several widely-used methods for handling asynchronous data, which can be broadly classified into two categories:

(i) Imputation Methods: These methods aim to fill in missing observations. In addition to our NN method, we include:

- Previous-Tick interpolation (PI), as used in Zhang (2011).
- Linear Interpolation (LI).

**Figure 3 Box plots of the $L_1$ norm of the incremental relative error matrix.**




*Note.* Upper panel: varying number of assets; lower panel: different observation frequencies. Here, "LI" stands for linear interpolate; "PI" stands for previous-tick interpolate; "PA" for pre-averaging; "NN" stands for our nuclear norm. Small green squares denote the mean values.

(ii) Subsampling Methods: These methods create a synchronized dataset by selecting a sparse subset of the original data. We consider:

- Refresh Time (RT), proposed by Barndorff-Nielsen et al. (2008).
- Pre-Averaging (PA), as used in Mykland et al. (2019).

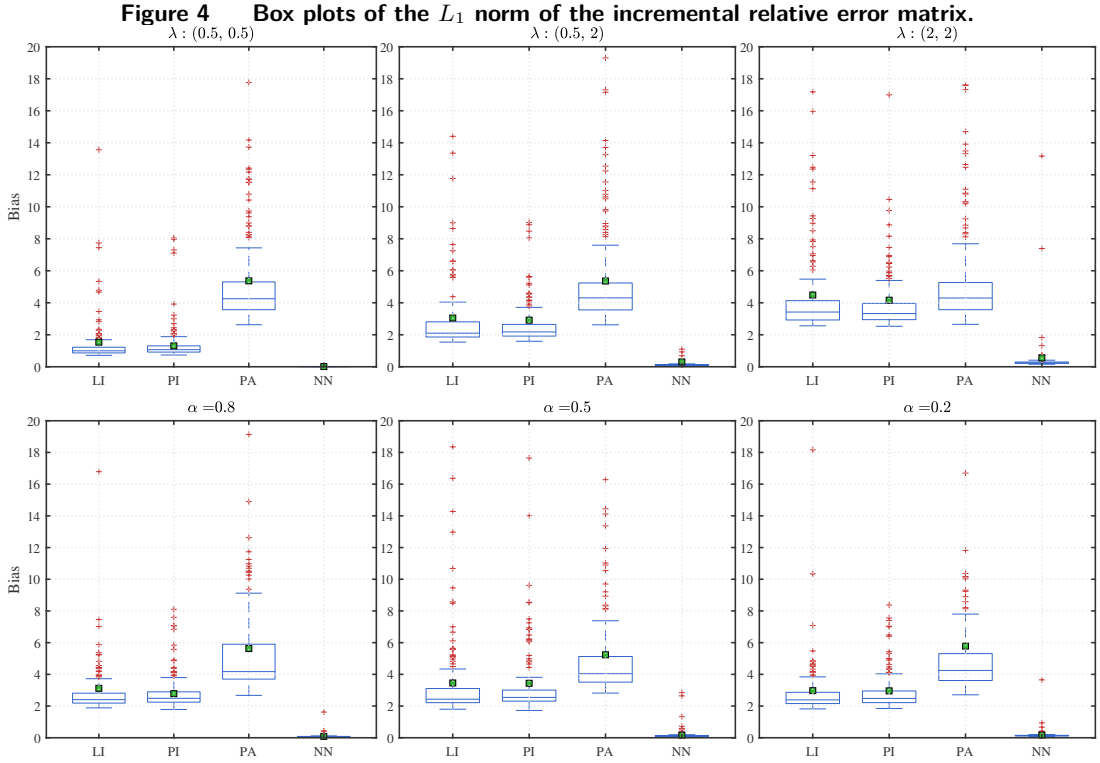
For all methods, after obtaining a synchronized return matrix, we estimate the covariance matrix using the Principal Component Analysis (PCA)-based approach developed by Ait-Sahalia and Xiu (2017).

We evaluate the estimation accuracy of increment matrix $\widehat{\Pi}$ by reporting scaled $L_1$ norm of relative error $(\widehat{\Pi}-\Pi)\oslash\Pi$, that is, $\frac{1}{nN}\|(\widehat{\Pi}-\Pi)\oslash\Pi\|_{L_1}$, where $(a_{ij})_{n\times m}\oslash(b_{ij})_{n\times m}=(a_{ij}/b_{ij})_{n\times m}$. We evaluate the estimation accuracy of covariance matrix $\widehat{\Sigma}$ by reporting various relative estimation error of norms, including the Frobenius $\|\widehat{\Sigma}-\Sigma\|_F/\|\Sigma\|_F$, the matrix $L_2$ $\|\widehat{\Sigma}-\Sigma\|/\|\Sigma\|$, and the maximum $\|\widehat{\Sigma}-\Sigma\|_{max}/\|\Sigma\|_{max}$ norms, where $\Sigma$ represents the true integrated covariance.

## 4.2. Simulation Results

Figure 3 presents the recovery bias of the returns matrix across various sample sizes and number of assets. Notably, our nuclear norm approach consistently outperforms alternative methods,[1] with

**Figure 4 Box plots of the $L_1$ norm of the incremental relative error matrix.**




*Note.* Upper panel: different asynchronous intensities; lower panel: different factor strengths. Here, "LI" stands for linear interpolate; "PI" stands for previous-tick interpolate; "PA" for pre-averaging; "NN" stands for our nuclear norm. Small green squares denote the mean values.

low estimation errors even with few assets ($N = 20$). Furthermore, our method demonstrates superior stability, as evidenced by its minimal performance variation, whereas other techniques exhibit substantially larger fluctuations.

Figure 4 illustrates the recovery bias of the returns matrix under varying levels of asynchrony and different factor strengths. Across all scenarios, our nuclear norm method leads the field, maintaining the lowest estimation errors. As expected, every method's performance deteriorates as asynchrony intensifies—unobservable prices become excessive at extreme levels of asynchrony. Nevertheless, our approach consistently yields small, stable errors. Moreover, variations in factor strength have only a minor impact on its performance.

We further evaluate the accuracy of the covariance matrix estimation across different asynchronous recovery methods, examining how it is influenced by four key variables: the number of assets and the sampling frequency (Table 1), the magnitude of asynchronous intensity and the strength of underlying factors (Table 2).

The left panel of Table 1 illustrates the impact of number of assets ($N$) on covariance estimation accuracy. Two key findings emerge. First, our NN method consistently outperforms all competing

**Table 1 Covariance estimation accuracy: Varying assets and frequencies.**

| | Dimension | | | | | | Frequency | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | RT | LI | PI | PA | NN | Minute | RT | LI | PI | PA | NN |
| | | | | | | $\|\hat{\Sigma}-\Sigma\|_F/\|\Sigma\|_F$ | | | | | |
| 20 | 0.131 (0.055) | 0.134 (0.035) | 0.132 (0.029) | 0.296 (0.159) | **0.073** (0.032) | 0.5 min | 0.121 (0.049) | 0.080 (0.029) | 0.080 (0.028) | 0.252 (0.123) | **0.060** (0.025) |
| 50 | 0.143 (0.060) | 0.137 (0.037) | 0.133 (0.033) | 0.301 (0.156) | **0.077** (0.031) | 1 min | 0.151 (0.067) | 0.139 (0.037) | 0.135 (0.035) | 0.301 (0.152) | **0.079** (0.032) |
| 120 | 0.149 (0.062) | 0.141 (0.036) | 0.137 (0.034) | 0.301 (0.153) | **0.076** (0.032) | 5 min | 0.307 (0.127) | 0.233 (0.058) | 0.227 (0.053) | 0.431 (0.197) | **0.145** (0.058) |
| | | | | | | $\|\hat{\Sigma}-\Sigma\|/\|\Sigma\|$ | | | | | |
| 20 | 0.126 (0.063) | 0.133 (0.041) | 0.115 (0.039) | 0.289 (0.178) | **0.074** (0.036) | 0.5 min | 0.119 (0.056) | 0.080 (0.033) | 0.078 (0.032) | 0.246 (0.137) | **0.058** (0.028) |
| 50 | 0.138 (0.069) | 0.137 (0.042) | 0.125 (0.041) | 0.296 (0.175) | **0.077** (0.035) | 1 min | 0.147 (0.076) | 0.139 (0.043) | 0.131 (0.042) | 0.292 (0.171) | **0.079** (0.037) |
| 120 | 0.144 (0.072) | 0.141 (0.041) | 0.133 (0.041) | 0.293 (0.172) | **0.076** (0.036) | 5 min | 0.294 (0.145) | 0.233 (0.066) | 0.216 (0.065) | 0.419 (0.220) | **0.145** (0.064) |
| | | | | | | $\|\hat{\Sigma}-\Sigma\|_{max}/\|\Sigma\|_{max}$ | | | | | |
| 20 | 0.155 (0.078) | 0.119 (0.038) | 0.135 (0.035) | 0.277 (0.194) | **0.070** (0.036) | 0.5 min | 0.149 (0.064) | 0.071 (0.030) | 0.084 (0.027) | 0.230 (0.132) | **0.053** (0.026) |
| 50 | 0.162 (0.080) | 0.121 (0.038) | 0.139 (0.035) | 0.270 (0.179) | **0.071** (0.034) | 1 min | 0.169 (0.083) | 0.126 (0.040) | 0.147 (0.037) | 0.274 (0.164) | **0.071** (0.036) |
| 120 | 0.164 (0.079) | 0.130 (0.040) | 0.152 (0.039) | 0.280 (0.170) | **0.071** (0.036) | 5 min | 0.304 (0.159) | 0.224 (0.069) | 0.274 (0.068) | 0.388 (0.208) | **0.138** (0.065) |

*Note.* This table summarizes results from 200 Monte Carlo simulations. The reported values are averages, with standard deviations in parentheses. The left and right panels show results under varying numbers of assets ($N$) and observation frequencies, respectively. Covariance matrices are estimated using a PCA-based approach (Ait-Sahalia and Xiu 2017). The methods compared are: Refresh Time (RT), Linear Interpolation (LI), Previous-Tick Interpolation (PI), Pre-Averaging (PA), and our proposed Nuclear Norm (NN) method. Bold entries highlight the best-performing method in each scenario.

methods across both error norms and all sizes ($N = 20$, 50, 120). Second, the NN method also exhibits the lowest standard deviation, highlighting its superior stability and reliability.

The right panel examines the effect of observation frequency. As expected, the estimation accuracy of all methods degrades as the frequency decreases. Despite this general trend, the NN method maintains its top-ranking performance, demonstrating its robustness even with less frequent data. In contrast, while the interpolation methods (LI and PI) perform reasonably well, particularly at higher frequencies, the PA method consistently yields the poorest results.

Next, we verify the effect of asynchronous intensity and factor intensity on covariance estimation. First, we examine the impact of asynchronous intensity. As expected, the performance of most methods deteriorates as the intensity of asynchrony increases (i.e., as more prices become unobservable). However, our proposed NN method demonstrates remarkable robustness, showing only a slight degradation in performance. In contrast, the interpolation-based methods (LI and PI) are highly sensitive to this parameter. While they perform reasonably well at a low asynchronous intensity

**Table 2 Covariance estimation accuracy: Varying asynchrony and factor strength.**

| | Asynchronous intensity | | | | | | Factor strength | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda_{\text{asy}}$ | RT | LI | PI | PA | NN | $\alpha$ | RT | LI | PI | PA | NN |
| | | | | | $\|\hat{\Sigma}-\Sigma\|_F/\|\Sigma\|_F$ | | | | | | |
| $(0.5, 0.5)$ | 0.134 (0.054) | **0.060** (0.023) | 0.062 (0.023) | 0.301 (0.152) | 0.079 (0.033) | 0.8 | 0.158 (0.066) | 0.139 (0.036) | 0.135 (0.034) | 0.314 (0.150) | **0.084** (0.031) |
| $(0.5, 2.0)$ | 0.192 (0.091) | 0.243 (0.034) | 0.252 (0.033) | 0.301 (0.152) | **0.078** (0.032) | 0.5 | 0.201 (0.061) | 0.148 (0.030) | 0.146 (0.028) | 0.382 (0.140) | **0.120** (0.025) |
| $(2.0, 2.0)$ | 0.201 (0.093) | 0.361 (0.036) | 0.340 (0.035) | 0.301 (0.152) | **0.077** (0.032) | 0.2 | 0.383 (0.045) | **0.258** (0.015) | 0.261 (0.014) | 0.615 (0.112) | 0.285 (0.018) |
| | | | | | $\|\hat{\Sigma}-\Sigma\|/\|\Sigma\|$ | | | | | | |
| $(0.5, 0.5)$ | 0.134 (0.061) | **0.057** (0.027) | 0.058 (0.027) | 0.292 (0.171) | 0.079 (0.037) | 0.8 | 0.151 (0.077) | 0.138 (0.042) | 0.130 (0.042) | 0.300 (0.171) | **0.081** (0.037) |
| $(0.5, 2.0)$ | 0.185 (0.103) | 0.234 (0.043) | 0.234 (0.044) | 0.292 (0.171) | **0.078** (0.037) | 0.5 | 0.172 (0.078) | 0.136 (0.040) | 0.127 (0.039) | 0.338 (0.172) | **0.097** (0.035) |
| $(2.0, 2.0)$ | 0.193 (0.105) | 0.361 (0.042) | 0.333 (0.043) | 0.292 (0.171) | **0.077** (0.036) | 0.2 | 0.274 (0.072) | **0.160** (0.024) | 0.153 (0.022) | 0.503 (0.169) | 0.181 (0.027) |
| | | | | | $\|\hat{\Sigma}-\Sigma\|_{max}/\|\Sigma\|_{max}$ | | | | | | |
| $(0.5, 0.5)$ | 0.150 (0.067) | **0.061** (0.029) | 0.080 (0.035) | 0.274 (0.164) | 0.072 (0.036) | 0.8 | 0.170 (0.082) | 0.120 (0.040) | 0.139 (0.032) | 0.279 (0.161) | **0.072** (0.036) |
| $(0.5, 2.0)$ | 0.198 (0.109) | 0.259 (0.072) | 0.273 (0.076) | 0.274 (0.164) | **0.071** (0.036) | 0.5 | 0.178 (0.076) | 0.119 (0.037) | 0.132 (0.030) | 0.306 (0.155) | **0.081** (0.032) |
| $(2.0, 2.0)$ | 0.212 (0.114) | 0.319 (0.055) | 0.342 (0.074) | 0.274 (0.164) | **0.069** (0.036) | 0.2 | 0.228 (0.063) | 0.156 (0.037) | 0.161 (0.035) | 0.390 (0.141) | **0.154** (0.034) |

*Note.* This table summarizes results from 200 Monte Carlo simulations. The reported values are averages, with standard deviations in parentheses. The left and right panels show results under varying asynchronous intensities ($\lambda_{\text{asy}}$) and factor strengths, respectively. Covariance matrices are estimated using a PCA-based approach (Ait-Sahalia and Xiu 2017). The methods compared are: Refresh Time (RT), Linear Interpolation (LI), Previous-Tick Interpolation (PI), Pre-Averaging (PA), and our proposed Nuclear Norm (NN) method. Bold entries highlight the best-performing method in each scenario.

(e.g., 0.5), their accuracy collapses when the intensity is high (e.g., 2). This is because these methods rely on nearby observed prices for imputation. As asynchrony intensifies, the average time between valid price observations widens, forcing interpolation over longer gaps and inevitably introducing significant bias.

Second, we analyze the effect of factor strength ($\alpha$). The PA method performs very poorly, especially when the underlying factors are weak. The NN method achieves the best performance across nearly all scenarios. The only exception occurs at $\alpha = 0.2$, where the PI and LI methods yield comparable results. This is likely because when the common factor is extremely weak, the cross-sectional dependence is minimal, which reduces the relative advantage of our global, factor-based approach. Nevertheless, even in this edge case, the NN method still provides the most accurate recovery of the underlying return matrix, as shown in Figure 4.

# 5. Empirical Analysis

To demonstrate the practical applicability and utility of our proposed methodology, we conduct several empirical applications. The remainder of this section is organized as follows. Section 5.1 describes the asynchronous characteristics of high-frequency data. Section 5.2 highlights the difference between the eigenvalues derived from traditional stale prices (derived from the PI method) and those recovered by our methodology. Section 5.3 compares the imputation errors across different approaches. Section 5.4 evaluates the impact of these methods on portfolio selection. Finally, Section 5.5 analyzes the beta discrepancies during periods of market turbulence.

## 5.1. Data Description

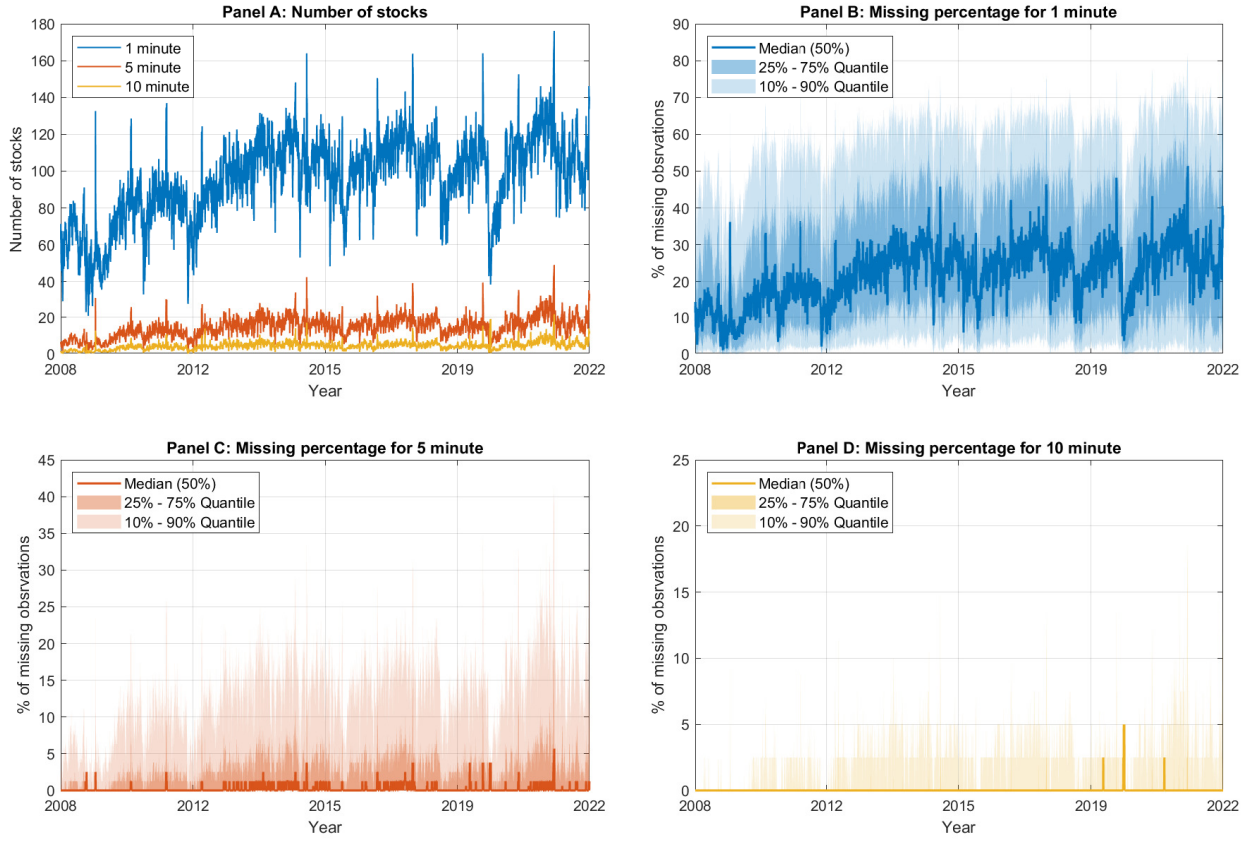
We combine data for S&P 500 constituents from the TAQ database with data for the SPDR S&P 500 ETF Trust (SPY) from Pi Trading, covering the period from January 2008 to December 2022. Following Bollerslev et al. (2024), we exclude holidays and days with shortened trading hours. Our analysis is based on high-frequency data at 1-minute, 5-minute, and 10-minute frequencies. For each frequency, an observation is deemed "missing" if no trade for the respective asset occurs within the corresponding time interval (e.g., within a given one-minute window for the 1-minute series). We construct a balanced panel, which ultimately comprises 360 stocks.

Asynchrony is a pervasive feature of high-frequency data. In our analysis, we define the 1-minute interval as the base observational unit, corresponding to a discrete time grid $\mathcal{T} = \{t_1, ..., t_n\}$.[2] The challenge of asynchronous trading can thus be framed as an imputation or data completion problem. Our primary goal is to recover complete time series of prices at 1-minute, 5-minute, and 10-minute frequencies, respectively, for every stock in the panel.

Let $\mathcal{T} = \{t_1, ..., t_n\}$ represent the complete grid of discrete observation times within a given period (e.g., a trading day with $n = 390$ one-minute intervals starting from an initial time $t_0$). For any stock $i$, let $\{\tau_{i1}, ..., \tau_{ii_n}\}$ denote the subset of times at which its price is actually observed. Asynchronous trading occurs when a stock's observation set does not cover the entire grid, i.e., $\mathcal{T} \setminus \{\tau_{i1}, ..., \tau_{ii_n}\} \neq \{\emptyset\}$. Consequently, the set of time points where the price of stock $i$ is considered missing is defined by the set difference $\mathcal{T} \setminus \{\tau_{i1}, ..., \tau_{ii_n}\}$. Figure 5 provides a summary of such missingness across our high-frequency panel data.

Panel A of Figure 5 plots the daily average number of missing stocks at the 1-minute frequency. Despite a well-documented increase in overall trading activity over the past 15 years, the number of missing stocks at this high frequency has remained persistently high, showing no significant downward trend. While this number temporarily declined during the COVID-19 pandemic, the trend was short-lived, reverting to its previous level of approximately 100 within a year. In contrast, at lower frequencies (5- and 10-minute intervals), the issue of missing data is substantially less severe, with the count of missing stocks typically remaining below 40.

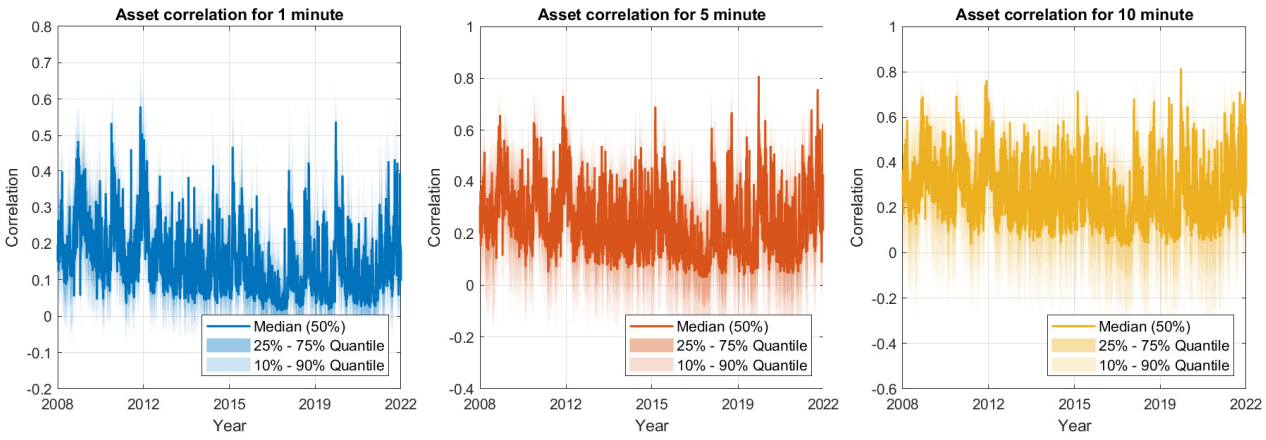
**Figure 5 Missing values over time.**




*Note.* This figure illustrates patterns of missing data in our sample. Panel A displays the daily average number of stocks with missing observations at the 1-minute frequency. Panels B, C, and D depict the daily evolution of the cross-sectional quantiles of missing observations for data sampled at 1-minute, 5-minute, and 10-minute frequencies, respectively.

Panel B of Figure 5 illustrates the intraday evolution of the cross-sectional quantiles of the missing probability—also known as staleness probability[3]—at the 1-minute frequency. We observe a median missing rate of 20–30%, a level consistent with the findings of Bandi et al. (2020) for NYSE-listed stocks. The prevalence of such missing data, or price stagnation (i.e., price staleness), is a critical issue in high-frequency analysis. The problem becomes even more acute at finer time scales; for instance, Bandi et al. (2024) reports a missing rate exceeding 50% for highly liquid stocks at the 10-second frequency. Using these stale prices as if they were true observations of the underlying price process can introduce significant biases into key financial analyses, such as volatility and correlation estimation. Returning to our specific findings, Panel B reveals that the extent of missingness can be extreme for some assets, with the 90th percentile reaching 60% even at the 1-minute frequency. As we decrease the sampling frequency, this issue is substantially mitigated. Panel C shows that the 90th percentile of the missing rate drops to approximately 15% for 5-minute data. For 10-minute

**Figure 6 Asset realized correlation over time.**




*Note.* This figure reports the results of the cross-sectional quantiles for realized correlation (daily).

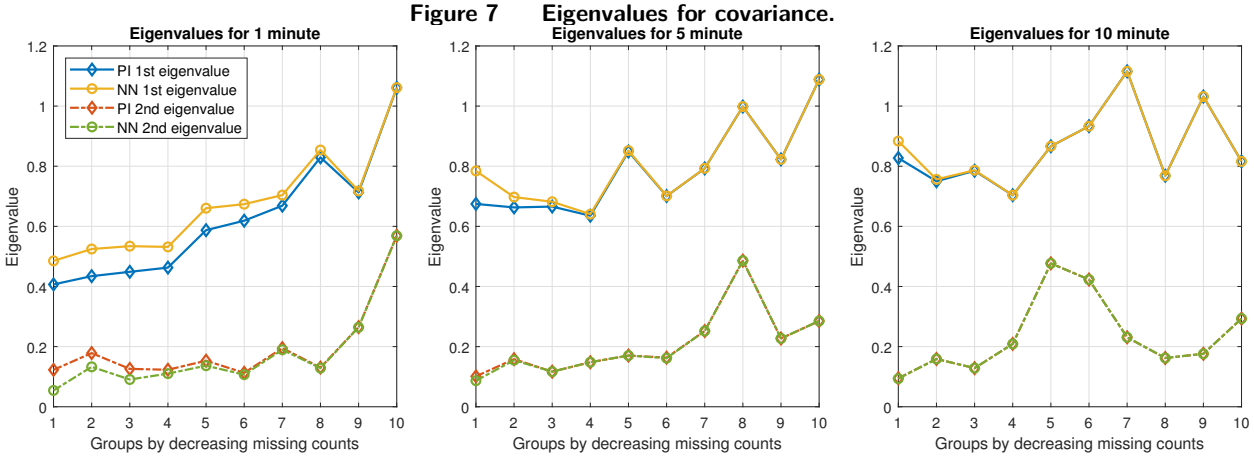
data (Panel D), the missing rate becomes negligible, with the 90th percentile falling below 5%. This trend is intuitive, as non-trading spells of 10 minutes are relatively uncommon for the stocks in our sample.

Figure 6 confirms the presence of significant cross-sectional correlations among stocks, although the magnitude of this correlation tends to decrease as the sampling frequency increases. This strong inter-dependence, often driven by common factors like industry effects, represents a valuable source of information for handling asynchrony. Ignoring it when imputing prices risks discarding crucial data.

However, conventional imputation methods that rely solely on an asset's own time series fail to exploit this cross-sectional information. For instance: The PI method, which carries forward the last observed price, utilizes only past univariate information and ignores contemporaneous trades in related stocks. The LI method artificially smooths prices between an asset's trades. Consequently, these univariate approaches cannot propagate the impact of market-wide information—reflected in the trading of correlated assets—to the price of an asynchronously traded stock. Therefore, any imputation scheme based purely on an individual time series is inherently limited, as it neglects vital information from the cross-section, such as latent factors. Capturing this complex dependence structure effectively requires sophisticated modeling tools (see, e.g., the discussion in Pelger 2020).

An alternative approach, distinct from imputation, is the refresh time scheme, which synchronizes data by subsampling, retaining only the time points where a specified set of assets has traded (Aït-Sahalia et al. 2010). This method, however, faces a significant trade-off in high-dimensional settings. A full-cross-section refresh time discards a vast amount of temporal data, whereas a pairwise refresh time is restricted to two assets and fails to scale. Ultimately, both univariate imputation and subsampling

**Figure 7 Eigenvalues for covariance.**




*Note.* This figure shows the first two eigenvalues of the panel covariance matrix for different groups, which are divided sequentially according to the missingness of the original data, with group 1 indicating the highest number of missing values and group 10 indicating the lowest number of missing values.

schemes struggle to effectively balance the preservation of temporal data with the integration of high-dimensional cross-sectional information.

## 5.2. Eigenvalues

The eigenvalues of the covariance matrix provide direct insights into the structure of systematic risk. To investigate this, we first categorize our 360 stocks into 10 equally-sized groups (36 stocks per group) based on their total number of missing observations, sorted from highest to lowest. Figure 7 then plots the magnitudes of the first two eigenvalues of the covariance matrix for each of these groups. These leading eigenvalues represent the amount of variance explained by the first two principal components, which are often interpreted as dominant systematic factors.

To evaluate our methodology, we compare it against the PI imputation, a standard benchmark in high-frequency finance. The PI method is the primary source of price staleness, as it simply carries forward the last observed price, creating an artificially static price series. Our analysis contrasts the eigenvalue structure of the covariance matrix derived from these stale, PI-imputed prices with that derived from the "effective prices" recovered by our proposed method.

Our findings reveal that price staleness significantly distorts the covariance matrix's eigenvalue structure, an effect that is most pronounced under two conditions: at higher frequencies (1-minute) and for stocks with more missing data.

(i) At the 1-minute frequency, the discrepancy is stark. In Group 1 (stocks with the most missing data), the leading eigenvalue from the PI method is approximately 0.4, whereas our method yields a value of around 0.5. This gap diminishes as data quality improves, becoming minimal in Group 10. A similar, substantial difference is also observed for the second eigenvalue.

(ii) At lower frequencies (5- and 10-minute), the distortion from the PI method is less severe, with significant differences in eigenvalues confined primarily to the first few groups (i.e., those with the most missing data).

Furthermore, focusing on the more reliable eigenvalues generated by our method for the 1-minute data, we observe a clear increasing trend in their magnitude from Group 1 to Group 10. This suggests a fundamental difference in the risk composition of these stocks. Stocks that trade frequently (e.g., in Group 10) exhibit stronger co-movement and are more exposed to systematic factors. Conversely, stocks prone to infrequent trading (e.g., in Group 1) possess a larger component of idiosyncratic risk, which is not captured by the leading principal components, even after our advanced imputation.

### 5.3. Imputation Error

In this subsection, we conduct an extensive study to compare the imputation accuracy of our proposed method against several benchmarks, including the LI and PI methods. We evaluate performance based on the imputation errors of log-prices across various sampling frequencies.[4]

To create a controlled setting for evaluating imputation errors, we follow a two-step procedure to artificially mask observed data points. For each day $t$ in our sample ($t = 1, ..., T$):

Step 1: Identify testable data: We start with the potential log-price matrix for that day, denoted as $\mathcal{P}_t$ (an $N \times n$ matrix, where $N$ the number of stocks and $n$ is the number of intraday observations). We first exclude any prices that were already missing in the original dataset. The remaining set of observed prices forms our "ground truth" data for the test.

Step 2: Construct mask matrix: We then create a binary mask matrix, $\mathcal{M}_t$, of the same dimensions. For each position corresponding to an observable price in $\mathcal{P}_t$, we randomly assign a value of 1 (indicating the price will be masked) with a pre-specified probability $p$, and 0 otherwise. We vary this masking probability $p$ across the set $\{0.1, 0.2, ..., 0.7\}$ to assess performance under different levels of data sparsity. [5]

The observed data available to the imputation methods is thus the element-wise product $\mathcal{P}_t^{obs} = \mathcal{P}_t \circ \mathcal{M}_t^c$, where $\mathcal{M}_t^c$ is the complement of $\mathcal{M}_t$ and $\circ$ denotes the Hadamard product. After applying an imputation method to obtain an estimate $\hat{\mathcal{P}}_t$, we measure its accuracy only on the artificially masked data points. The main results are summarized in Table 3. We report two error metrics, averaged over all trading days:

$$\begin{aligned} \text{Absolute error} &:= \frac{1}{T}\sum_{t=1}^{T} \frac{\|(\hat{\mathcal{P}}_t - \mathcal{P}_t) \circ \mathcal{M}_t\|_F}{\|\mathcal{M}_t\|_F}, \\ \text{Relative error} &:= \frac{1}{T}\sum_{t=1}^{T} \frac{\|(\hat{\mathcal{P}}_t - \mathcal{P}_t) \circ \mathcal{M}_t\|_F}{\|\mathcal{P}_t \circ \mathcal{M}_t\|_F}, \end{aligned} \tag{14}$$

where $\hat{\mathcal{P}}_t$ is the imputed price matrix, and the Frobenius norm $\|\cdot\|_F$ in the numerator is calculated only over the set of masked entries.

**Table 3 Imputation error for different methods.**

| Mask | 1 minute | | | 5 minute | | | 10 minute | | |
|---|---|---|---|---|---|---|---|---|---|
| probability | PI | LI | NN | PI | LI | NN | PI | LI | NN |
| | | | | Absolute error | | | | | |
| 0.1 | 0.1088 | 0.0757 | 0.0681 | 0.2070 | 0.1426 | 0.1225 | 0.2853 | 0.1947 | 0.1646 |
| 0.2 | 0.1151 | 0.0785 | 0.0711 | 0.2195 | 0.1488 | 0.1289 | 0.3027 | 0.2033 | 0.1737 |
| 0.3 | 0.1220 | 0.0819 | 0.0747 | 0.2338 | 0.1558 | 0.1363 | 0.3213 | 0.2128 | 0.1839 |
| 0.4 | 0.1311 | 0.0864 | 0.0794 | 0.2514 | 0.1648 | 0.1459 | 0.3455 | 0.2254 | 0.1973 |
| 0.5 | 0.1427 | 0.0925 | 0.0857 | 0.2739 | 0.1764 | 0.1582 | 0.3768 | 0.2425 | 0.2153 |
| 0.6 | 0.1586 | 0.1006 | 0.0942 | 0.3036 | 0.1922 | 0.1749 | 0.4178 | 0.2650 | 0.2390 |
| 0.7 | 0.1814 | 0.1128 | 0.1067 | 0.3485 | 0.2170 | 0.2008 | 0.4781 | 0.3004 | 0.2759 |
| | | | | Relative error | | | | | |
| 0.1 | 0.0270 | 0.0188 | 0.0169 | 0.0509 | 0.0351 | 0.0300 | 0.0699 | 0.0478 | 0.0402 |
| 0.2 | 0.0286 | 0.0195 | 0.0176 | 0.0540 | 0.0366 | 0.0316 | 0.0742 | 0.0499 | 0.0425 |
| 0.3 | 0.0303 | 0.0204 | 0.0185 | 0.0575 | 0.0383 | 0.0334 | 0.0788 | 0.0522 | 0.0449 |
| 0.4 | 0.0326 | 0.0215 | 0.0197 | 0.0618 | 0.0405 | 0.0357 | 0.0847 | 0.0553 | 0.0482 |
| 0.5 | 0.0354 | 0.0230 | 0.0212 | 0.0673 | 0.0434 | 0.0388 | 0.0924 | 0.0595 | 0.0526 |
| 0.6 | 0.0394 | 0.0250 | 0.0233 | 0.0746 | 0.0473 | 0.0429 | 0.1024 | 0.0650 | 0.0584 |
| 0.7 | 0.0450 | 0.0280 | 0.0264 | 0.0857 | 0.0534 | 0.0492 | 0.1172 | 0.0737 | 0.0674 |

*Note.* See (14) for error calculations in this table. Key to abbreviations: "LI" stands for linear interpolate; "PI" stands for previous-tick interpolate; "NN" stands for our nuclear norm.

Table 3 presents several key findings. First and foremost, our proposed method consistently and significantly outperforms the conventional PI method, yielding imputation errors that are approximately half the size of those from the PI approach. This superiority is robust across all tested sampling frequencies and masking probabilities. We therefore conclude that relying on stale prices, which inherently ignores both cross-sectional dependence and dynamic time-series patterns, leads to substantial and avoidable inaccuracies in price imputation. In contrast, our global optimization framework, by recovering an underlying low-rank structure, provides a much more accurate approximation of the true, unobserved high-frequency prices.

Second, our method also demonstrates a clear advantage over the LI method. A plausible explanation for this result is that the gains from capturing cross-sectional dependence, a key feature of our model, outweigh the benefits of simple temporal interpolation. Furthermore, the performance gap between our method and the benchmarks widens as the masking probability increases. This highlights the particular strength of our approach in scenarios with high data sparsity, where traditional methods like PI become increasingly unreliable.

Finally, we observe that the imputation performance of all methods tends to degrade at lower frequencies (e.g., 10-minute). This presents a general challenge for high-frequency data imputation, as longer time intervals between observations can obscure the underlying price dynamics. Despite this challenge, the superiority of our method remains striking. For instance, even under a severe 70% masking probability, our method achieves a lower error rate than the PI method does under a

minimal 10% masking probability. This compelling result underscores the remarkable efficiency and robustness of our proposed imputation framework.

### 5.4. Applications to Portfolio Selection

In this subsection, we evaluate the economic significance of the covariance matrix estimates by examining their performance in a practical portfolio allocation context. We consider the following constrained minimum-variance portfolio problem:

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}, \quad s.t. \quad \boldsymbol{w}'\mathbf{1} = 1, \|\boldsymbol{w}\|_1 \leq c, \tag{15}$$

where $\|\boldsymbol{w}\|_1 \leq c$ impose an risk-exposure constraint, with $\|\cdot\|_1$ denoting the $L_1$ norm. A value of $c = 1$ corresponds to a long-only portfolio (no short sales), whereas $c > 1$ allows for short selling.

Our empirical analysis uses intraday data only, excluding overnight returns to avoid complications from dividend issuances and stock splits. To ensure robustness to price jumps, we further apply a five-standard-deviation truncation rule. The truncation threshold is calibrated using the bipower variation estimator of Barndorff-Nielsen and Shephard (2004), adjusted for diurnal volatility patterns following the methodology of Li et al. (2017) and Bollerslev et al. (2024).

Following the literature Fan et al. (2012), Ait-Sahalia and Xiu (2017), and Cui et al. (2024), we adopt a monthly rebalancing strategy. At the end of each month, we construct the optimal portfolio weights by solving problem (15) using an estimate of the integrated covariance matrix. This covariance matrix is estimated using the high-frequency data from the past month, first imputed by one of the competing methods and then processed using the estimator of Ait-Sahalia and Xiu (2017). We then evaluate the out-of-sample performance of these portfolios over the next month under various exposure constraints $c$.

The results are presented in Table 4, where we report three key performance metrics: out-of-sample annualized average return (AR), annualized standard deviation (SD), and the Sharpe ratio (SR).

Table 4 reports the out-of-sample portfolio performance from 2016 to 2022 across various gross exposure constraints ($c$) and for five equally-sized stock groups. These groups are formed by sorting stocks based on their degree of data missingness, from highest (Group 1) to lowest (Group 5), allowing us to assess performance under different data quality scenarios.

A primary finding is that portfolios constructed using our imputation method achieve the highest Sharpe ratios in nearly all scenarios. This result underscores the economic value of accurately estimating the covariance matrix by accounting for asynchrony. While the Sharpe ratios tend to decline as leverage increases (i.e., as $c$ grows), they stabilize for $c > 4$. It is important to note that while more advanced covariance estimators, such as that of Cui et al. (2024), could potentially further enhance performance, our focus here is on isolating the impact of the imputation method itself, for which we use a standard estimation approach.

**Table 4 The out-of-sample performance of monthly-rebalanced optimal portfolios.**

| | | Constraint: 2 | | | Constraint: 3 | | | Constraint: 4 | | | Constraint: 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G. | | PI | LI | NN | PI | LI | NN | PI | LI | NN | PI | LI | NN |
| 1 | AR | 11.796 | 12.925 | 13.777 | 12.782 | 12.852 | 13.836 | 12.782 | 12.852 | 13.938 | 12.782 | 12.852 | 13.938 |
| | SD | 23.179 | 24.265 | 24.439 | 23.345 | 24.576 | 25.191 | 23.345 | 24.576 | 25.281 | 23.345 | 24.576 | 25.281 |
| | SR | 0.508 | 0.532 | **0.564** | 0.547 | 0.522 | **0.549** | 0.547 | 0.522 | **0.551** | 0.547 | 0.522 | **0.551** |
| 2 | AR | 6.743 | 6.999 | 6.993 | 6.155 | 6.321 | 6.861 | 6.115 | 6.064 | 6.537 | 6.115 | 6.064 | 6.537 |
| | SD | 20.954 | 21.072 | 21.355 | 20.819 | 20.923 | 21.220 | 20.811 | 20.927 | 21.262 | 20.811 | 20.927 | 21.262 |
| | SR | 0.321 | **0.332** | 0.327 | 0.295 | 0.302 | **0.323** | 0.293 | 0.289 | **0.307** | 0.293 | 0.289 | **0.307** |
| 3 | AR | 11.373 | 13.375 | 13.858 | 11.617 | 13.809 | 14.575 | 11.598 | 13.974 | 14.660 | 11.598 | 13.974 | 14.660 |
| | SD | 24.311 | 24.467 | 24.513 | 24.624 | 24.962 | 24.978 | 24.681 | 25.055 | 25.094 | 24.681 | 25.055 | 25.094 |
| | SR | 0.467 | 0.546 | **0.565** | 0.471 | 0.553 | **0.584** | 0.469 | 0.557 | **0.584** | 0.469 | 0.557 | **0.584** |
| 4 | AR | 5.655 | 5.431 | 5.844 | 6.097 | 6.011 | 6.690 | 6.351 | 6.204 | 6.936 | 6.350 | 6.204 | 6.935 |
| | SD | 23.790 | 23.950 | 23.844 | 23.604 | 23.703 | 23.567 | 23.566 | 23.669 | 23.538 | 23.566 | 23.669 | 23.538 |
| | SR | 0.237 | 0.226 | **0.245** | 0.258 | 0.253 | **0.283** | 0.269 | 0.262 | **0.294** | 0.269 | 0.262 | **0.294** |
| 5 | AR | 4.002 | 4.183 | 4.377 | 4.028 | 4.226 | 4.312 | 4.023 | 4.212 | 4.316 | 4.023 | 4.212 | 4.316 |
| | SD | 21.204 | 21.268 | 21.334 | 21.469 | 21.539 | 21.580 | 21.500 | 21.574 | 21.627 | 21.500 | 21.574 | 21.627 |
| | SR | 0.188 | 0.196 | **0.205** | 0.187 | 0.196 | **0.199** | 0.187 | 0.195 | **0.199** | 0.187 | 0.195 | **0.199** |

*Note.* This table reports the out-of-sample performance of portfolios constructed using different covariance matrix estimators, evaluated monthly from 2016 to 2022. We report the annualized average return (AR), annualized standard deviation (SD), and the Sharpe ratio (SR). "G." refers to the stock groups defined before, and $c$ denotes the gross exposure constraint from the optimization problem (15). Values in bold indicate the highest Sharpe ratio for each given case.

A striking pattern emerges from the inter-group comparison: portfolios of stocks with more missing data (e.g., Groups 1-2) consistently generate higher average returns and Sharpe ratios than those with less missing data (e.g., Groups 4-5). The Sharpe ratio of Group 1, for instance, is more than double that of Group 5. This observation aligns with the well-documented liquidity premium in asset pricing. Infrequent trading, which directly causes data missingness in our high-frequency setting, is a hallmark of illiquidity. Seminal works by Amihud and Mendelson (1986) and Pástor and Stambaugh (2003) have established that illiquid stocks tend to offer higher expected returns to compensate investors for higher trading costs and exposure to systematic liquidity risk.

However, the link is not one-to-one. Price staleness and traditional liquidity measures are highly correlated but distinct concepts (Bandi et al. 2020, 2024). Bandi et al. (2024) posit that staleness can be decomposed into a systematic component, driven by factors like capital "shadow costs", and an idiosyncratic component, related to asset-specific spreads. The performance differentials we observe across groups may therefore reflect a complex interplay of these factors, leading to conclusions that may differ from those based on low-frequency data alone.

Finally, it is worth noting that relying on price staleness for out-of-sample portfolio allocations may reduce out-of-sample performance, which is consistent with the findings of Kong et al. (2024). Moreover, here we also find that when missing values are severe (staleness probability is high), the

gap between portfolio allocation using estimates of effective prices (NN method) and using stale prices (PI method) may be larger.

### 5.5. Spot Beta

In this subsection, we investigate the impact of price staleness on the estimation of spot betas for individual stocks relative to an exchange-traded fund (ETF). We estimate the spot beta using the fixed-$k$ local regression method proposed by Bollerslev et al. (2024), to which we refer the reader for detailed computational procedures. To the best of our knowledge, the effect of price staleness (as induced by the PI method) on high-frequency beta estimation has not been systematically studied, and this analysis aims to fill that gap. This is a critical gap, as recent literature demonstrates the profound impact of high-frequency beta estimation on asset pricing tests. For example, Hollstein et al. (2020) find that the empirical failures of the Conditional CAPM can be largely resolved by moving from daily to high-frequency betas. While their study successfully underscores the value of high-frequency data, it relies on lag-based adjustments to mitigate non-synchronicity. Our analysis, therefore, also serves to show how our synchronization framework can provide cleaner and more reliable beta estimates, which are essential inputs for the type of asset pricing tests conducted by Hollstein et al. (2020), thereby preventing potentially misleading inferences caused by data artifacts like price staleness.

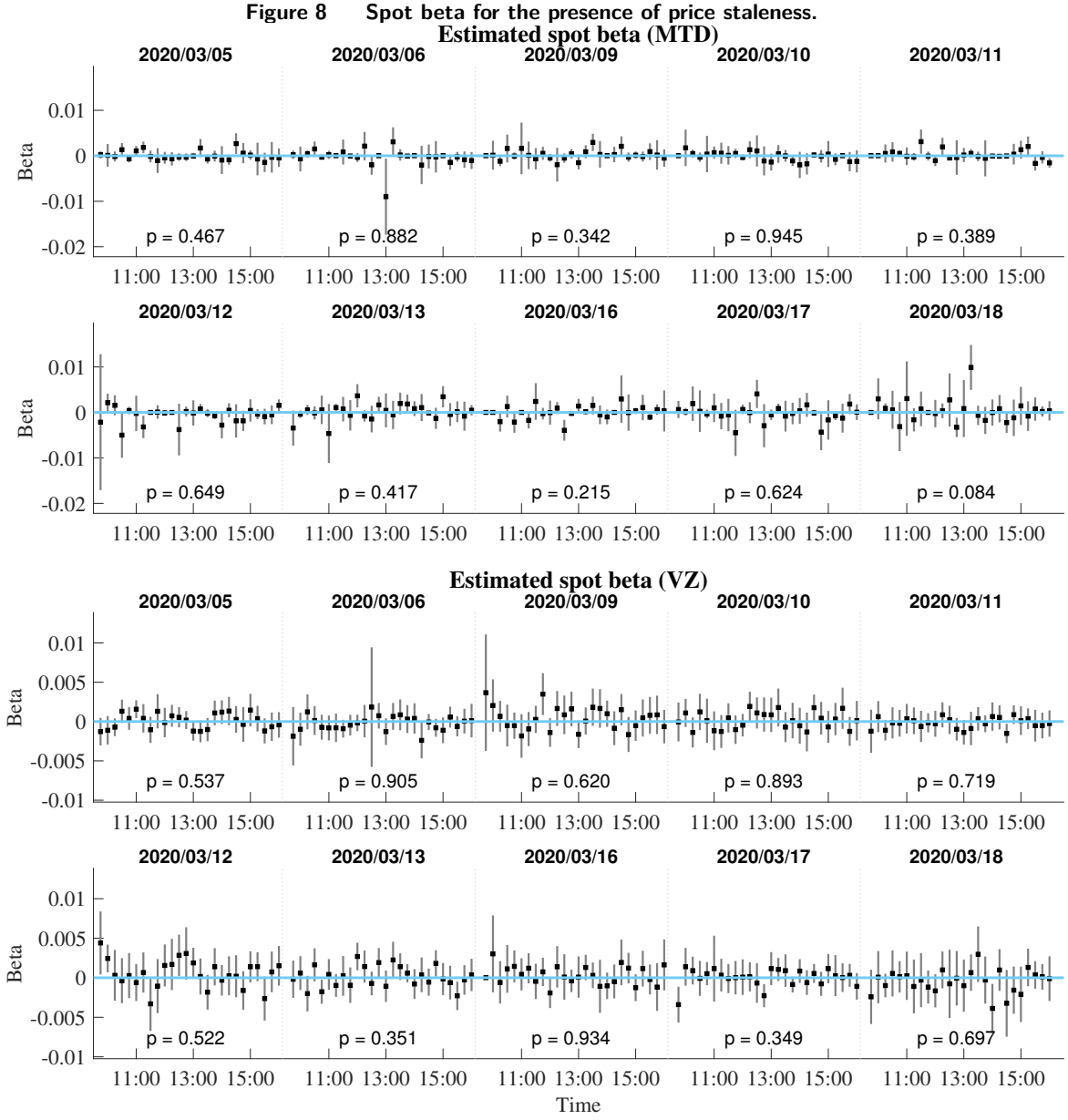
Our market proxy is the SPDR S&P 500 ETF Trust (SPY). While the S&P 500 Index (SPX) is the theoretical benchmark, it is not directly tradable. The SPY, designed to track the SPX, is one of the most liquid and widely traded ETFs globally, making it an ideal instrument for this analysis.

To highlight the effects of asynchrony, we focus our analysis on a period of extreme market turbulence: the ten trading days from March 5 to March 18, 2020. This period witnessed the onset of the COVID-19 financial crisis, characterized by an oil price war (March 9, "Black Monday I" crash), the declaration of a global pandemic (March 11), and emergency actions by the Federal Reserve (March 15), leading to extreme price volatility. In our regression framework, the SPY return series serves as the regressor, and the individual stock return series is the dependent variable.[6]

We specifically compare two types of stocks: those with a high number of missing observations during this period and those with very few. For instance, Mettler-Toledo International (MTD, Industrials) and Verizon Communications (VZ, Communication Services) represent the extremes of high and low data missingness, respectively, within our sample for this period. Our subsequent analysis will focus on these two representative stocks.

In our empirical tests, we primarily focus on testing the null hypothesis of zero beta ($H_0 : \beta = 0$). This focus is motivated by the strong idiosyncratic nature of individual stock returns, which often makes it statistically challenging to detect a consistently significant, non-zero beta. Following

**Figure 8 Spot beta for the presence of price staleness.**




*Note.* This figure plots the estimated spot betas of Mettler-Toledo International (MTD) and Verizon Communications (VZ) against the SPY. The betas are estimated using 1-minute price data over 15-minute rolling windows, along with their corresponding 90% confidence intervals. The analysis covers the two-week period of high market volatility from March 5, 2020, to March 18, 2020. The $p$-value reported in each panel corresponds to a test of the functional null hypothesis that the entire spot beta process for a given day is equal to zero ($H_0 : \beta_t = 0$ for all $t$).

Bollerslev et al. (2024), our main analysis uses a 15-minute estimation window ($k = 15$). Robustness checks using $k = 5$ and $k = 10$ are provided in the Supplementary Appendix.

The upper panel of Figure 8 reveals a striking pattern for the MTD stock: the spot beta estimates are frequently exactly zero, with confidence intervals degenerating to zero width. This is not a reflection of economic reality but rather a methodological artifact induced by the PI imputation. As

established in the classic literature on non-synchronous trading (Scholes and Williams 1977), when a stock's price does not update, the PI method records a zero return. This mechanically sets the stock's local covariance with the market (i.e. SPY) to zero, forcing the corresponding beta estimate to be zero as well. This issue is, unsurprisingly, most severe for the less liquid stock, MTD.

The consequences extend beyond distorted point estimates; this artifact critically undermines the validity of statistical inference. For instance, on March 9, 2020 ("Black Monday I"), the PI-based data yields an exceptionally high $p$-value of 0.342 for the functional test of the null hypothesis that the beta process is identically zero. This misleading result leads to a failure to reject the null of zero systematic risk during a major market crash.

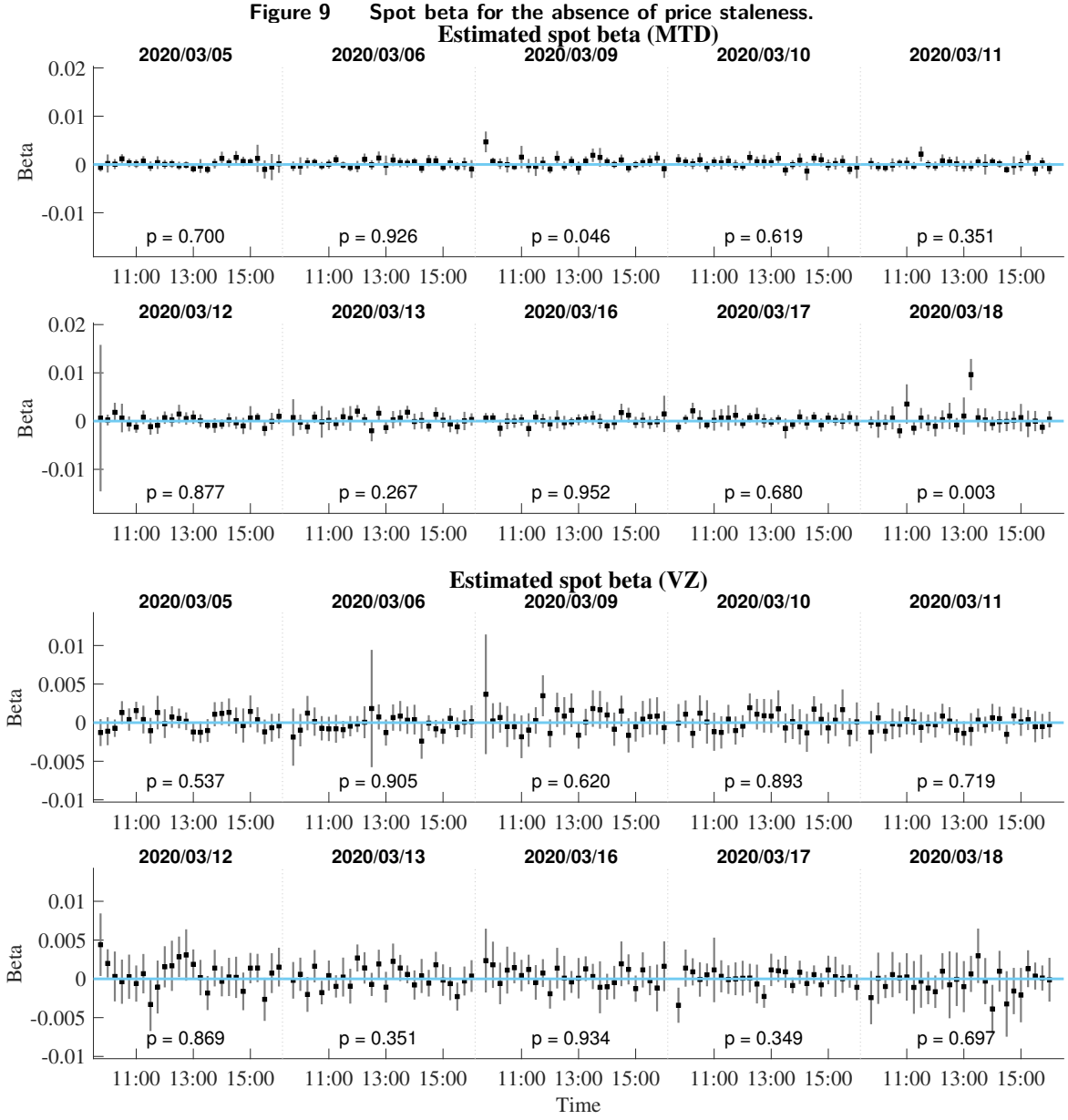
In stark contrast, when using the effective prices recovered by our method (see Figure 9), the $p$-value for the same test on the same day is merely 0.046, leading to a clear rejection of the null hypothesis. This discrepancy highlights a fundamental flaw: the zero-inflated return series generated by the PI method violates the "local Gaussianity" assumption that underpins the inferential framework of Bollerslev et al. (2024). Consequently, any statistical tests based on such contaminated data become unreliable.

In stark contrast to the erratic zero-beta estimates produced by the PI method, the beta paths derived from our approach (Figure 9) are notably smoother and more continuous. This demonstrates that once the microstructure noise from non-synchronous trading is properly filtered, a stock's systematic risk exposure is revealed to be a dynamically evolving intraday process, not a series of binary jumps. This finding aligns with the modern consensus in high-frequency finance that asset price processes are well approximated by continuous semi-martingales (Aït-Sahalia and Jacod 2014).

The resulting statistical inference is also substantially more credible. For MTD on March 18, 2020, for example, the uniform test yields a $p$-value of 0.003, allowing for a decisive rejection of the zero-beta null hypothesis and accurately capturing the stock's significant market risk on that day. Even on days with less statistical power, the confidence intervals remain informative, showing the beta fluctuating around a non-zero mean. This underscores how a robust imputation method, when combined with the optimal inference framework of Bollerslev et al. (2024), enables reliable inference even from short estimation windows.

The period of March 2020 included several market-wide "circuit breaker" trading halts (on March 9, 12, 16, and 18), and Figure 9 provides a clear window into how risk characteristics evolved under such extreme stress. For instance, VZ, a defensive telecommunications stock, generally exhibits a beta near zero, consistent with its sector profile. However, during the most turbulent sessions, the volatility of its beta and the width of its confidence bands visibly expand, reflecting a heightened spillover of systematic risk even to traditionally "safe" assets. This provides compelling visual evidence for the intraday variation of systematic risk, a phenomenon documented by Andersen et al. (2021).

**Figure 9 Spot beta for the absence of price staleness.**




*Note.* This figure plots the estimated spot betas of Mettler-Toledo International (MTD) and Verizon Communications (VZ) against the SPY. The betas are estimated using 1-minute price data over 15-minute rolling windows, along with their corresponding 90% confidence intervals. The analysis covers the two-week period of high market volatility from March 5, 2020, to March 18, 2020. The $p$-value reported in each panel corresponds to a test of the functional null hypothesis that the entire spot beta process for a given day is equal to zero ($H_0 : \beta_t = 0$ for all $t$).

Furthermore, MTD's beta path on March 18 reveals a distinct intraday pattern, trending upwards in the afternoon, potentially reflecting the market's continuous repricing of the stock's risk as new information was digested. Such fine-grained dynamics are completely obscured in daily or lower-frequency data.

In summary, the proper handling of non-synchronous trading is a critical prerequisite for drawing valid economic conclusions from high-frequency spot regressions. The conventional previous-tick method not only produces severely biased zero-beta estimates but also invalidates statistical inference. By employing a robust imputation technique, we can uncover the true, dynamic, and economically meaningful evolution of a stock's beta, particularly during periods of market turmoil. This capacity for precise, real-time risk measurement has significant practical implications for risk management, algorithmic trading, and event study analysis, allowing market participants to assess and respond to changes in asset risk with a high degree of precision when it matters most.

## 6. Conclusion

Asynchronous trading is a fundamental challenge in high-frequency finance that biases risk estimates and impairs asset allocation. We address this by recasting data synchronization as a constrained matrix completion problem. Our framework recovers the complete matrix of synchronous price increments by minimizing its nuclear norm—capturing the underlying low-rank factor structure—subject to linear constraints derived from observed, asynchronous prices.

Theoretically, we prove the existence and uniqueness of our estimator, establish its convergence rate, and show that it efficiently pools information across both liquid and illiquid assets, overcoming a key limitation of traditional methods. Empirically, using extensive simulations and a large panel of S&P 500 stocks, we demonstrate that our approach substantially outperforms established benchmarks. It corrects systematic biases in risk estimates and, most critically, generates portfolios with economically and statistically significant higher out-of-sample Sharpe ratios. Our research provides a powerful and practical tool for uncovering the true dynamics of asset prices, opening promising avenues for future work in areas such as microstructure modeling and mixed-frequency analysis. By providing a theoretically sound and empirically validated solution to a long-standing problem, this paper enables more precise risk measurement and offers a clearer lens into the dynamic nature of modern financial markets.

## Acknowledgments

Authors are listed in alphabetical order.

## Endnotes

[1]The RT method is excluded from this comparison. This is because RT is a subsampling technique, not an imputation method; it only retains a sparse subset of true observed returns and does not generate any estimated values to be compared against the ground truth.

[2]We select the 1-minute frequency, rather than a higher frequency (e.g., 1-second), to mitigate the effects of microstructure noise. Nevertheless, our methodology is also applicable to such ultra-high-frequency data, a direction we reserve for future research.

[3]Staleness probability refers to the likelihood that a price update does not occur within a given interval. See, e.g., Bandi et al. (2017), Bandi et al. (2020), and Kong et al. (2024).

[4]This choice is motivated by the semi-martingale nature of high-frequency asset price dynamics. Unlike low-frequency data, the properties of high-frequency price increments depend on the length of the time interval, making a direct analysis of log-prices more appropriate across different frequencies (e.g., 1-minute, 5-minute, and 10-minute).

[5]Four different missing observations in low-frequency data were investigated by Duan et al. (2024): (i) missing-at-random, (ii) simultaneous adoption, (iii) staggered adoption, (iv) switchback. Random missing in high-frequency data, or random missing but with missing probabilities that are not constant but semi-martingale processes may be more acceptable (e.g., Bandi et al. 2017 and Kong et al. 2024).

[6]The period's turmoil ignited on March 9 when an oil price war erupted, sparking recession fears and the "Black Monday I" crash. The crisis deepened on March 11 as the WHO declared a global pandemic and the U.S. announced a European travel ban, devastating key industries. In a dramatic response, the Federal Reserve executed an emergency rate cut to zero on March 15, but this was perceived as a panic move, failing to reassure investors and triggering another severe market plunge.

## References

Aït-Sahalia Y, Fan J, Xiu D (2010) High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association* 105(492):1504–1517.

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

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

Amihud Y, Mendelson H (1986) Asset pricing and the bid-ask spread. *Journal of Financial Economics* 17(2):223–249.

Andersen TG, Thyrsgaard M, Todorov V (2021) Recalcitrant betas: Intraday variation in the cross-sectional dispersion of systematic risk. *Quantitative Economics* 12(2):647–682.

Bandi FM, Kolokolov A, Pirino D, Renò R (2020) Zeros. *Management Science* 66(8):3466–3479.

Bandi FM, Pirino D, Reno R (2017) Excess idle time. *Econometrica* 85(6):1793–1846.

Bandi FM, Pirino D, Renò R (2024) Systematic staleness. *Journal of Econometrics* 238(1):105522.

Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2008) Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* 76(6):1481–1536.

Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2011) Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* 162(2):149–169.

Barndorff-Nielsen OE, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* 2(1):1–37.

Bollerslev T, Li J, Ren Y (2024) Optimal inference for spot regressions. *American Economic Review* 114(3):678–708.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J, et al. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3(1):1–122.

Candes E, Recht B (2012) Exact matrix completion via convex optimization. *Communications of the ACM* 55(6):111–119.

Chen D, Mykland PA, Zhang L (2020) The five trolls under the bridge: Principal component analysis with asynchronous and noisy high frequency data. *Journal of the American Statistical Association* 115(532):1960–1977.

Chen Y, Fan J, Ma C, Yan Y (2019) Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences* 116(46):22931–22937.

Cui L, Hong Y, Li Y, Wang J (2024) A regularized high-dimensional positive definite covariance estimator with high-frequency data. *Management Science* 70(10):7242–7264.

Duan J, Pelger M, Xiong R (2024) Factor analysis for causal inference on large non-stationary panels with endogenous treatment. *2024b):"Target PCA: Transfer learning large dimensional panel data," Journal of Econometrics* 244(2):105521.

Fan J, Furger A, Xiu D (2016) Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *Journal of Business & Economic Statistics* 34(4):489–503.

Fan J, Li Y, Yu K (2012) Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association* 107(497):412–428.

Gandy S, Recht B, Yamada I (2011) Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems* 27(2):025010.

Hayashi T, Yoshida N (2005) On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* 11(2):359–379.

Hollstein F, Prokopczuk M, Wese Simen C (2020) The conditional capital asset pricing model revisited: Evidence from high-frequency betas. *Management Science* 66(6):2474–2494.

Kong X (2017) On the number of common factors with high-frequency data. *Biometrika* 104(2):397–410.

Kong X (2018) On the systematic and idiosyncratic volatility with large panel high-frequency data. *The Annals of Statistics* 46(3):1077–1108.

Kong X, Lin JG, Liu C, Liu GY (2023) Discrepancy between global and local principal component analysis on large-panel high-frequency data. *Journal of the American Statistical Association* 118(542):1333–1344.

Kong X, Liu C (2018) Testing against constant factor loading matrix with large panel high-frequency data. *Journal of Econometrics* 204(2):301–319.

Kong X, Wu B, Ye W (2024) Staleness factor model and volatility estimation. *arXiv preprint arXiv:2410.07607* .

Li J, Todorov V, Tauchen G (2017) Jump regressions. *Econometrica* 85(1):173–195.

Liu C, Tang CY (2014) A quasi-maximum likelihood approach for integrated covariance matrix estimation with high frequency data. *Journal of Econometrics* 180(2):217–232.

Mykland PA, Zhang L, Chen D (2019) The algebra of two scales estimation, and the s-tsrv: High frequency estimation that is robust to sampling times. *Journal of Econometrics* 208(1):101–119.

Onatski A, Wang C (2024) Spurious factors in data with local-to-unit roots. *Econometric Theory* 1–38.

Pástor L, Stambaugh RF (2003) Liquidity risk and expected stock returns. *Journal of Political Economy* 111(3):642–685.

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

Pelger M (2020) Understanding systematic risk: A high-frequency approach. *The Journal of Finance* 75(4):2179–2220.

Recht B, Fazel M, Parrilo PA (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* 52(3):471–501.

Scheinberg K, Ma S, Goldfarb D (2010) Sparse inverse covariance selection via alternating linearization methods. *Advances in Neural Information Processing Systems* 23.

Scholes M, Williams J (1977) Estimating betas from nonsynchronous data. *Journal of Financial Economics* 5(3):309–327.

Shephard N, Xiu D (2017) Econometric analysis of multivariate realised qml: Estimation of the covariation of equity prices under asynchronous trading. *Journal of Econometrics* 201(1):19–42.

Shin M, Kim D, Fan J (2023) Adaptive robust large volatility matrix estimation based on high-frequency financial data. *Journal of Econometrics* 237(1):105514.

Szarek SJ (1983) The finite dimensional basis problem with an appendix on nets of grassmann manifolds. *Acta Mathematica* 151(1):153–179.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press).

Zhang B, Pan G, Gao J (2018) Clt for largest eigenvalues and unit root testing for high-dimensional nonstationary time series. *The Annals of Statistics* 46(5):2186–2215.

Zhang L (2011) Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics* 160(1):33–47.

Zhang N, Zhang L (2016) No gap second-order optimality conditions for a matrix cone programming induced by the nuclear norm. *Asia-Pacific Journal of Operational Research* 33(02):1650010.

# Technical Proofs and Additional Results

This appendix contains technical proofs and additional results for the paper.

- Appendix EC.1 provides the proofs of theorems in the main text.
- Appendix EC.2 provides the additional simulation results.
- Appendix EC.3 provides additional empirical results.

## EC.1. Proofs

*Proof of Theorem 1* The proof proceeds by establishing the restricted isometry property (RIP) for the linear operator $\mathcal{A}$. We will prove the two bounds, (7) and (8), separately.

**Proof of the Bound in** (7)

Recall the model setting that

$$dX_t = \mu_t dt + \sigma_t dW_t + \sigma_t^* dW_t^* =: dX^\mu + dX_t^c + dX_t^*, \tag{EC.1}$$

which implies a corresponding decomposition for the increment matrix, where $\sigma_t = \sigma^0 \Sigma_t$. Let $V = \left(\int_{t_{j-1}}^{t_j} \Sigma_t dW_t\right)_{r\times n}$, $\Pi^\mu = \left(X^\mu_{it_j} - X^\mu_{it_{j-1}}\right)_{N\times r}$, $\Pi = \left(X^c_{it_j} - X^c_{it_{j-1}}\right)_{N\times n}$ and $\Pi^{*,\mu} = \left(X^*_{it_j} - X^*_{it_{j-1}}\right)_{N\times n} + \Pi^\mu := \Pi^* + \Pi^\mu$. Now, we first prove that the solution for $\Pi$ in (3) is equal to $\Pi_0$ with probability approaching one (conditional on $\{R_j, D_{ik}; i \le N, j \le n, k \le n\}$). Rearranging the second equation in (3), we have $\mathcal{A}(\Pi) = b - \mathcal{A}(\Pi^{*,\mu}) := b^*$ and the solution is to find a matrix of rank no larger than $r$ so that this constraint is satisfied and $\|\Pi\|_*$ is minimized. To this end, we first prove that $\Pi$ satisfies the nearly isometry condition for small enough $\delta$,

$$\begin{aligned} &P_{R,D}\left((1-\delta)\|\Pi\|_F \le \|\mathcal{A}(\Pi)\| \le (1+\delta)\|\Pi\|_F \text{ for all } \sigma^0 \in \Sigma^0\right) \\ \ge\ & 1 - C\exp\left\{-c^2r^2 / \left(\sum_{j=1}^n R_j^2\right)^{1/4}\right\} - C\exp\{-c^2r^2\delta^2 N^\alpha / \overline{L}_1\}, \end{aligned}$$

which amounts to the following condition for another small $\delta$ ($\delta$ is a generic small constant that may vary at different appearance),

$$\begin{aligned} &P_{R,D}\left(\left|\frac{\|\mathcal{A}(\Pi)\|^2 - \|\Pi\|_F^2}{\|\Pi\|_F^2}\right| \le \delta \text{ for all } \sigma^0 \in \Sigma^0\right) \\ \ge\ & 1 - C\exp\left\{-c^2r^2 / \left[\left(\sum_{j=1}^n R_j^2\right)^{1/4}\right]\right\} - C\exp\{-c^2r^2\delta^2 N^\alpha / \overline{L}_1\}. \end{aligned} \tag{EC.2}$$

Since $\|\Pi\|_F^2 = N^\alpha \|V\|_F^2$, (EC.2) reduces to the following relative concentration condition for $V$.

$$\begin{aligned} &P_{R,D}\left(\left|\frac{\|\mathcal{A}(\sigma^0 V)\|^2 - \|\sigma^0 V\|_F^2}{\|V\|_F^2}\right| > N^\alpha\delta \text{ for some } \sigma^0 \in \Sigma^0\right) \\ \le\ & C\exp\left\{-c^2r^2\delta^2 N^\alpha / \overline{L}_1\right\} + C\exp\left\{-c^2r^2 / \left(\sum_{j=1}^n R_j^2\right)^{1/4}\right\}. \end{aligned} \tag{EC.3}$$

**Step 1.1: Bounding the Denominator $\|V\|_F^2$.**

In the sequel, we aim to prove (EC.3). We first consider the limit of $\|V\|_F^2$ in the denominator which is irrelevant to $\sigma^0$. Notice that $V$ consists of $r$ rows of martingale differences with shrinking intervals, by standard stochastic calculus, for $k=1,...,r$,

$$\begin{aligned}
&P_{R,D}\left(\|V_{k\cdot}\|^2-\int_0^T\Sigma_{tk\cdot}\Sigma'_{tk\cdot}dt>x\sqrt{\sum_{j=1}^n R_j^2}\right)\\
\le\ & E\exp\left(-\theta x\sqrt{\sum_{j=1}^n R_j^2}+\theta\sum_{j=1}^n\left[(V_{kj}-V_{k(j-1)})^2-\int_{t_{j-1}}^{t_j}\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt\right]\right)\\
\le\ & \exp\left(-\theta x\sqrt{\sum_{j=1}^n R_j^2}\right)E\left\{\prod_{j=1}^{n-1}e^{\theta\left[\left(V_{kj}-V_{k(j-1)}\right)^2-\int_{t_{j-1}}^{t_j}\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt\right]}\right.\\
&\left.\times\ E_{\mathcal{F}_{t_{n-1}}}e^{\theta\left[\left(V_{kn}-V_{k(n-1)}\right)^2-\int_{t_{n-1}}^{t_n}\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt\right]}\right\}.
\end{aligned}$$

Now we compute the conditional expectation inside the unconditional expectation. By the change of time, there exists a standard Brownian motion $B$ so that

$$(V_{kn}-V_{k(n-1)})^2-\int_{t_{n-1}}^{t_n}\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt=B^2_{<V_k>_{t_{n-1}}^{t_n}}-<V_k>_{t_{n-1}}^{t_n},$$

where $<V_k>_a^b$ is the quadratic variation of $V_k$ in the interval $(a,b]$. Notice that $e^{\theta[(V_{kj}-V_{k(j-1)})^2-\int_{t_{j-1}}^{t_j}\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt]}$ is the end point of a submartingale for $\theta>0$, by Assumption 2, $<V_k>_{t_{j-1}}^{t_j}\le CR_j$, and hence

$$\begin{aligned}
&E_{\mathcal{F}_{t_{n-1}}}e^{\theta\left[(V_{kn}-V_{k(n-1)})^2-\int_{t_{n-1}}^{t_n}\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt\right]}\\
\le\ & Ee^{\theta\left[B^2_{CR_n}-CR_n\right]}=Ee^{\theta CR_n(Z^2-1)}\le e^{2\theta^2C^2R_n^2},
\end{aligned}\tag{EC.4}$$

by Lemma 1 of Fan et al. (2012) for $|\theta CR_n|\le 1/4$, where $Z$ is a standard normal random variable. Iteratively using (EC.4),

$$\begin{aligned}
&P_{R,D}\left(\sum_{j=1}^n(V_{kj}-V_{k(j-1)})^2-\int_0^T\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt>x\sqrt{\sum_{j=1}^n R_j^2}\right)\\
\le\ & \exp\left(-\theta x\sqrt{\sum_{j=1}^n R_j^2}+2\theta^2C^2\sum_{j=1}^n R_j^2\right),
\end{aligned}$$

which is minimized when $\theta=x/(4C^2\sqrt{\sum_{j=1}^n R_j^2})$ and the minimum is $e^{(-x^2/(8C^2))}$. By the range of $\theta$, the range of $x$ is $0<x\le C\sqrt{\sum_{j=1}^n R_j^2}/R_j$ for all $j=1,...,n$. Therefore,

$$P_{R,D}\left(\|V\|_F^2-\sum_{k=1}^r\int_0^T\Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt>x\sqrt{\sum_{j=1}^n R_j^2}\right)$$

$$\leq \sum_{k=1}^{r} P\left(\sum_{j=1}^{n}(V_{kj}-V_{k(j-1)})^2-\int_0^T \Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt > x\sqrt{\sum_{j=1}^{n}R_j^2}/r\right)$$
$$\leq r\max_{k\leq r} P\left(\sum_{j=1}^{n}(V_{kj}-V_{k(j-1)})^2-\int_0^T \Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt > x\sqrt{\sum_{j=1}^{n}R_j^2}/r\right).$$

Similarly, we can obtain the other side

$$P_{R,D}\left(\|V\|_F^2-\sum_{k=1}^{r}\int_0^T \Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt < -x\sqrt{\sum_{j=1}^{n}R_j^2}\right)\leq e^{-x^2/(8C^2)+\log r}.$$

Summarizing the above two equations, we have

$$P_{R,D}\left(|\|V\|_F^2-\sum_{k=1}^{r}\int_0^T \Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt| > x\sqrt{\sum_{j=1}^{n}R_j^2}\right)\leq 2e^{-x^2/(8C^2)+\log r}. \tag{EC.5}$$

We take $x=cr/\{2(\sum_{j=1}^n R_j^2)^{1/4}\}$ for $c$ small enough which satisfies the range condition for $x$ because of the assumption that $\frac{(\sum_{j=1}^n R_j^2)^{3/4}}{\max_j R_j}\geq \frac{cr}{2C}$. By the condition that $\int_0^T \Sigma_{t,k\cdot}\Sigma'_{t,k\cdot}dt\geq c$ and choose $c$ small,

$$P_{R,D}\left(\|V\|_F^2\leq\frac{cr}{2}\right)\leq 2\exp\left\{-c^2r^2/\left\{32C^2\left(\sum_{j=1}^{n}R_j^2\right)^{1/2}\right\}\right\}. \tag{EC.6}$$

**Step 1.2: Bounding the Numerator $\|\mathcal{A}(\sigma^0V)\|_F^2-\|\sigma^0V\|_F^2$.**

Next, we consider the numerator $\|\mathcal{A}(\sigma^0V)\|_F^2-\|\sigma^0V\|_F^2$. To express the difference by a sum of crossing products, we recall

$$d_{ik}:=\sharp\{j;\tau_{i(l-1)}\leq t_j\leq t_{k-1}\leq\tau_{il}\text{ for some }1\leq l\leq n_i\}$$

to be the number of potential increments before $\Delta_k^n V$ in the observed interval $(\tau_{i(l-1)},\tau_{il}]$ which contains $t_{k-1}$. We make a convention that $d_{ik}:=0$ if the set is empty and the resulting sum $\sum_{l=1}^0\Delta_{k-l}^nV=0$. We rewrite

$$\|\mathcal{A}(\sigma^0V)\|_F^2-\|\sigma^0V\|_F^2=\sum_{k=1}^{n}\sum_{i=1}^{N}(\sigma^0(i,\cdot)\Delta_k^nV)(\sigma^0(i,\cdot)\sum_{l=1}^{d_{ik}}\Delta_{k-l}^nV).$$

Recall that $\overline{L}_1=\sum_{k=1}^n R_k\sum_{i=1}^N\|\sigma^0(i,\cdot)\|^2(\sum_{l=1}^{D_{ik}}R_{k-l})\log(\sum_{l=1}^{D_{ik}}R_{k-l})$.

$$\begin{aligned}
&P_{R,D}\left(\|\mathcal{A}(\sigma^0V)\|_F^2-\|\sigma^0V\|_F^2>N^\alpha x\right)\\
\leq&\exp\{-\theta N^\alpha x\}\,E\prod_{k=1}^{n-1}\exp\left\{\theta\sum_{i=1}^{N}(\sigma^0(i,\cdot)\Delta_k^nV)(\sigma^0(i,\cdot)\sum_{l=1}^{d_{ik}}\Delta_{k-l}^nV)\right\}\\
&\times E_{\mathcal{F}_{n-1}}\exp\left\{\theta\sum_{i=1}^{N}(\sigma^0(i,\cdot)\Delta_n^nV)(\sigma^0(i,\cdot)\sum_{l=1}^{d_{in}}\Delta_{n-l}^nV)\right\}\\
\leq&\exp\left\{-\theta N^\alpha x+Cr\theta^2\sum_{k=1}^{n}R_k\|\sum_{i=1}^{N}\sigma^0(i,\cdot)\sum_{l=1}^{D_{ik}}\Delta_{k-l}^nV\sigma^0(i,\cdot)\|_F^2\right\}
\end{aligned}$$

$$\leq \exp\left\{-\theta N^{\alpha}x + Cr\theta^2 \sum_{k=1}^{n} R_k \sum_{m=1}^{r}\sum_{i=1}^{N}(\sigma^0(i,m))^2 \sum_{i=1}^{N}(\sigma^0(i,\cdot)\sum_{l=1}^{D_{ik}}\Delta_{k-l}^n V)^2\right\}$$
$$\leq \exp\left\{-\theta N^{\alpha}x + Cr^2\theta^2 N^{\alpha}\overline{L}_1\right\}, \tag{EC.7}$$

for any $\theta > 0$, where in the last step we have made use of the fact that $|\sum_{l=1}^{d_{ik}}\Delta_{k-l}^n V| \leq C\left(\sum_{l=1}^{D_{ik}} R_{k-l}\right)^{1/2}\log\left(\sum_{l=1}^{D_{ik}} R_{k-l}\right)^{1/2}$ by Assumption 2 and the law of iterated logarithm for diffusion paths, and the fact that given $\mathcal{F}_{k-1}$,

$$\sum_{i=1}^{N}(\sigma^0(i,\cdot)\Delta_k^n V)(\sigma^0(i,\cdot)\sum_{l=1}^{d_{ik}}\Delta_{k-l}^n V)$$

is an end point of a continuous martingale and hence can be represented by $B_{\mathcal{T}_{k,nN}}$ for some stopping time $\mathcal{T}_{k,nN}$ satisfying

$$\mathcal{T}_{k,nN} \leq CR_k N^{\alpha}\sum_{i=1}^{N}\|\sigma^0(i,\cdot)\|^2\left(\sum_{l=1}^{D_{ik}} R_{k-l}\right)\log\left(\sum_{l=1}^{D_{in}} R_{k-l}\right).$$

A simple use of the optional stopping theorem for submartingales leads to the last inequality.

**Step 1.3: Combining the Bounds.**

Take $\theta = x/(2Cr^2\overline{L}_1)$, the upper bound of (EC.7) is minimized with the minimum being $\exp\{-x^2N^{\alpha}/(4Cr^2\overline{L}_1)\}$ for any $x > 0$. Combining the bounds for the numerator and the denominator (EC.6) via a union bound, we have:

$$P_{R,D}\left(\frac{\|\mathcal{A}(\Pi)\|_F^2 - \|\Pi\|_F^2}{\|\Pi\|_F^2} > \delta\right)$$
$$\leq P_{R,D}\left(\|V\|_F^2 \leq cr/2\right) + P\left(|\|\mathcal{A}(\Pi)\|_F^2 - \|\Pi\|_F^2| > N^{\alpha}\delta cr/2\right)$$
$$\leq 2\exp\left\{-c^2r^2/\left(\sum_{j=1}^{n} R_j^2\right)^{1/2}\right\} + C\exp\left\{-c^2\delta^2 N^{\alpha}/\overline{L}_1\right\}, \tag{EC.8}$$

where $c$ is taken small and the first probability bound is irrelevant to $\sigma^0$. (EC.8) proves the theorem for fixed $\sigma^0$.

**Proof of the Bound in** (8)

The proof follows a similar strategy, but the analysis is more complex due to the presence of the idiosyncratic component $\Pi^*$. We use the decomposition $\Delta = \Pi + \Pi^*$.

**Step 2.1: Bounding the Denominator $\|\Delta\|_F^2$.**

To prove (8), we present a decomposition, $\|\Delta\|_F^2 = N^{\alpha}\|V\|_F^2 + \|\Pi^{*,\mu}\|_F^2 + 2tr(\Pi'\Pi^{*,\mu})$. The result for $\|V\|_F^2$ is already given as above. To derive the concentration result for $\|\Pi^*\|_F^2$, we decompose

$$\Pi^* = \Pi_1^* + \Pi_2^* := \left(\int_{t_{j-1}}^{t_j}\sigma_{t_{j-1}}^* dW_t^*\right)_{N\times n} + \left(\int_{t_{j-1}}^{t_j}(\sigma_t^* - \sigma_{t_{j-1}}^*)dW_t^*\right)_{N\times n}.$$

Notice that $\|\Pi_2^*\|_F^2 = \sum_{i=1}^N \sum_{j=1}^n \left(\int_{t_{j-1}}^{t_j} (\sigma_{it}^* - \sigma_{it_{j-1}}^*) dW_{it}^*\right)^2$ and $\int_{t_{j-1}}^{t_j} (\sigma_{it}^* - \sigma_{it_{j-1}}^*)^2 dt \le C R_j^{2-\epsilon}$. For $|\theta C R_j^{2-\epsilon}| < 1/4$,

$$
\begin{aligned}
&P_{R,D}\left( |\|\Pi_2^*\|_F^2 - \sum_{i=1}^N \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \left(\sigma_{it}^* - \sigma_{it_{j-1}}^*\right)^2 dt| > N y \sqrt{\sum_{j=1}^n R_j^{4-2\epsilon}} \right) \\
&\le N \max_i P\left( |\sum_{j=1}^n \left\{ \left( \int_{t_{j-1}}^{t_j} \left(\sigma_{it}^* - \sigma_{it_{j-1}}^*\right) dW_t^* \right)^2 - \int_{t_{j-1}}^{t_j} \left(\sigma_{it}^* - \sigma_{it_{j-1}}^*\right)^2 dt \right\}| \right. \\
&\quad \left. > y \sqrt{\sum_{j=1}^n R_j^{4-2\epsilon}} \right) \\
&\le 2N \exp\left\{ -\theta y \sqrt{\sum_{j=1}^n R_j^{4-2\epsilon}} + 2\theta^2 C^2 \sum_{j=1}^n R_j^{4-2\epsilon} \right\}, \qquad \text{(EC.9)}
\end{aligned}
$$

which is minimized at $\theta = y/(4C^2\sqrt{\sum_{j=1}^n R_j^{4-2\epsilon}})$ with $2e^{\{-y^2/(8C^2)+\log(N)\}}$ being the minimum for $0 < y < C\sqrt{\sum_{j=1}^n R_j^{4-2\epsilon}}/R_j^{2-\epsilon}$ for all $j = 1, ..., n$.

For $\Pi_1^*$, we have

$$
\|\Pi_1^*\|_F^2 - \sum_{i=1}^N \sum_{j=1}^n \sigma_{it_{j-1}}^{*2} (t_j - t_{j-1}) = \sum_{j=1}^n (\sigma_{t_{j-1}}^*)^2 (t_j - t_{j-1}) \sum_{i=1}^N [(Z_i)^2 - 1],
$$

where $Z_i$'s are independent standard normal random variables. By Lemma 2.27 of Wainwright (2019), for independent standard Gaussian random variables $Z_1, ..., Z_N$ and $\frac{C\theta^2 R_j^2}{2} < 1/4$,

$$
\begin{aligned}
&E_{\mathcal{F}_{t_{j-1}}} \exp\left\{ \theta (\sigma_{t_{j-1}}^*)^2 (t_j - t_{j-1}) \sum_{i=1}^N [(Z_i)^2 - 1] \right\} \\
&\le E_{\mathcal{F}_{t_{j-1}}} \exp\left\{ \frac{\theta^2 (\sigma_{t_{j-1}}^*)^4 \pi^2 (t_j - t_{j-1})^2}{2} \|\rho^*(Z_1, ..., Z_N)'\|_2^2 \right\} \\
&\le E_{\mathcal{F}_{t_{j-1}}} \exp\left\{ \frac{C\theta^2 R_j^2}{2} \sum_{i=1}^N Z_i^2 \right\} \le \exp\left\{ \frac{CNR_j^2\theta^2}{2} \left(1 + C\theta^2 R_j^2\right) \right\} \\
&\le \exp\left\{ \frac{3CNR_j^2\theta^2}{4} \right\}.
\end{aligned}
$$

Again, following the steps in proving the result for $\|V\|_F^2$, we have

$$
\begin{aligned}
&P_{R,D}\left( |\|\Pi_1^*\|_F^2 - \sum_{i=1}^N \sum_{j=1}^n \sigma_{it_{j-1}}^{*2} (t_j - t_{j-1})| > z \sqrt{N \sum_{j=1}^n R_j^2} \right) \\
&\le 2 \exp\left\{ -\theta z \sqrt{N \sum_{j=1}^n R_j^2} + \frac{3C\theta^2}{4} N \sum_{j=1}^n R_j^2 \right\}, \qquad \text{(EC.10)}
\end{aligned}
$$

which is minimized at $\theta = 2z/(3C\sqrt{N\sum_{j=1}^n R_j^2})$ with the minimum $e^{-z^2/3C}$ for $0 < z \le 3\sqrt{C}\sqrt{N\sum_{j=1}^n R_j^2}/(2\sqrt{2}R_j)$ for all $j$.

Now we summarize the results to give the concentration result for $\Delta$. Let $IV = N^\alpha \int_0^T (\sum_{k=1}^r \Sigma_{t,k\cdot}\Sigma'_{t,k\cdot})dt + \sum_{i=1}^N \int_0^T (\sigma_{it}^*)^2 dt =: N^\alpha IV_1 + IV_2$. For any $0 < u < 1$,

$$\begin{aligned}
&P_{R,D}\left(|\|\Delta\|_F^2 - IV| > Nu \text{ for some } \sigma^0 \in \mathcal{B}(N,r)\right)\\
\le\ &P_{R,D}\left(N^\alpha|\|V\|_F^2 - IV_1| > Nu/3\right)\\
&+P_{R,D}\left(\|\Pi^\mu\|_F^2 > Nu/3\right) + P_{R,D}\left(|\|\Pi^*\| - IV_2| > Nu/3\right).
\end{aligned}$$

For the first term, (EC.5) shows that

$$P_{R,D}\left(N^\alpha|\|V\|_F^2 - IV_1| > Nu/3\right) \le 2\exp\left\{-N^{2-2\alpha}u^2/\left(72C^2\sum_{j=1}^n R_j^2\right)\right\}$$

By the boundedness of the drift coefficient, for $n$ large enough,

$$P_{R,D}\left(\|\Pi^\mu\|_F^2 > Nu/3\right) = 0.$$

Taking $\theta = 1/(4CR_j^{2-\epsilon})$ and $y = u/(3\sqrt{\sum_{j=1}^n R_j^{4-2\epsilon}})$ in (EC.9), and $\theta = \frac{\epsilon^*}{\sqrt{2C}R_j}$ and $z = \sqrt{N}u/(3\sqrt{\sum_{j=1}^n R_j^2})$ for $\epsilon^*$ small enough in (EC.10), we have for small but fixed $u$,

$$\begin{aligned}
&P_{R,D}\left(|\|\Pi^*\| - IV_2| > Nu/3\right)\\
\le\ &2N\exp\left\{-\frac{u}{12C\max_j R_j^{2-\epsilon}} + \frac{\sum_{j=1}^n R_j^{4-2\epsilon}}{8(\max_j R_j^{2-\epsilon})^2}\right\}\\
&+2\exp\left\{-\frac{Nu\epsilon^*}{3\sqrt{2C}\max_j R_j} + \frac{3N(\epsilon^*)^2\sum_{j=1}^n R_j^2}{8(\max_j R_j)^2}\right\}\\
\le\ &2N\exp\left\{-\frac{u}{12C\max_j R_j^{2-\epsilon}}\right\} + 2\exp\left\{-\frac{Nu\epsilon^*}{3\sqrt{2C}\max_j R_j}\right\}.
\end{aligned}$$

Putting pieces together, for $N$ and $n$ large enough,

$$\begin{aligned}
&P_{R,D}\left(|\|\Delta\|_F^2 - IV| > Nu \text{ for some } \sigma^0 \in \mathcal{B}(N,r)\right)\\
\le\ &2\exp\left\{-N^{2-2\alpha}u^2/\left(72C^2\sum_{j=1}^n R_j^2\right)\right\} + 2N\exp\left\{-\frac{u}{12C\max_j R_j^{2-\epsilon}}\right\}\\
&+2\exp\left\{-\frac{Nu\epsilon^*}{3\sqrt{2C}\max_j R_j}\right\}.
\end{aligned}$$

Due to Assumption 2, we further deduce that

$$\begin{aligned}
&P_{R,D}\left(\|\Delta\|_F^2 \le Nc/2 \text{ for some } \sigma^0 \in \mathcal{B}(N,r)\right)\\
\le\ &2\exp\left\{-N^{2-2\alpha}c^2/\left(128C^2\sum_{j=1}^n R_j^2\right)\right\} + 2N\exp\left\{-\frac{c}{24C\max_j R_j^{2-\epsilon}}\right\}\\
&+2\exp\left\{-\frac{Nc\epsilon^*}{6\sqrt{2C}\max_j R_j}\right\}.
\end{aligned}$$

**Step 2.2: Bounding the Numerator $\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2$.**

So far we have given a bound for the denominator of $\{\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2\}/\|\Delta\|_F^2$. Next, we investigate into the numerator, i.e., the difference $\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2$. We begin with the difference between $\|\mathcal{A}(\Pi^*)\|_F^2$ and $\|\Pi^*\|_F^2$. This difference can also be expressed as

$$\|\mathcal{A}(\Pi^*)\|_F^2 - \|\Pi^*\|_F^2 = \sum_{k=1}^{n}\sum_{i=1}^{N}\Pi^*(i,k)\sum_{l=1}^{d_{ik}}\Pi^*(i,k-l).$$

Recall the notation $\overline{L}_2$. Similar to the derivation in (EC.7), we have

$$\begin{aligned}
&P_{R,D}\left(|\|\mathcal{A}(\Pi^*)\|_F^2 - \|\Pi^*\|_F^2| > y\sqrt{\overline{L}_2}\right)\\
&\le \exp\left\{-\theta y\sqrt{\overline{L}_2}\right\} E\prod_{k=1}^{n-1}\exp\left\{\theta\sum_{i=1}^{N}\Pi^*(i,k)\sum_{l=1}^{d_{ik}}\Pi^*(i,k-l)\right\}\\
&\quad\times E_{\mathcal{F}_{n-1}}\exp\left\{\theta\sum_{i=1}^{N}\Pi^*(i,n)\sum_{l=1}^{d_{in}}\Pi^*(i,n-l)\right\}\\
&\le \exp\left\{-\theta y\sqrt{\overline{L}_2} + C\theta^2\overline{L}_2\right\}, \qquad \text{(EC.11)}
\end{aligned}$$

for any $\theta > 0$ and where in the last step, again, we have made use of the optional stopping theorem for submartingales and the Lévy representation theorem for continuous martingales. Take $\theta = y/(2C\sqrt{\overline{L}_2})$, the upper bound of (EC.11) is minimized at $e^{-y^2/(4C)}$ for any $y > 0$.

Lastly, we consider the difference between $\|\mathcal{A}(\Pi^\mu)\|_F^2$ and $\|\Pi^\mu\|_F^2$. By the local boundedness of $\mu_t$ in Assumption 2, we soon have

$$|\|\mathcal{A}(\Pi^\mu)\|_F^2 - \|\Pi^\mu\|_F^2| \le C\sum_{i=1}^{N}\sum_{k=1}^{n}R_k\sum_{l=1}^{D_{ik}}R_{k-l} = o(N),$$

by Assumption 1. This shows that

$$P_{R,D}\left(|\|\mathcal{A}(\Pi^\mu)\|_F^2 - \|\Pi^\mu\|_F^2| > N\epsilon/3\right) = 0,$$

for large enough $N$ and $n$. Taking

$$x = \frac{\epsilon N^{1-\alpha}}{3},\ y = \frac{\epsilon N}{3\sqrt{\overline{L}_2}},$$

we have

$$\begin{aligned}
&P_{R,D}\left(|\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2| > N\epsilon\right)\\
&\le 2\exp\left\{-\frac{N^{2-\alpha}\epsilon^2}{36Cr^2\overline{L}_1}\right\} + 2\exp\left\{-\frac{N^2\epsilon^2}{36C\overline{L}_2}\right\}.
\end{aligned}$$

**Step 2.3: Combining the Bounds.**

Finally, combining the probability bounds for the numerator and denominator of $\frac{\|\mathcal{A}(\Delta)\|_F^2-\|\Delta\|_F^2}{\|\Delta\|_F^2}$ leads to the desired result in (8):

$$\begin{aligned}
&P_{R,D}\left(|\frac{\|\mathcal{A}(\Delta)\|_F^2-\|\Delta\|_F^2}{\|\Delta\|_F^2}|>\delta\right)\\
\le\ &P_{R,D}\left(|\|\mathcal{A}(\Delta)\|_F^2-\|\Delta\|_F^2|>\frac{cN\delta}{2}\right)+P_{R,D}\left(\|\Delta\|_F^2\le\frac{Nc}{2}\right)\\
\le\ &2\exp\left\{-\frac{N^{2-\alpha}c^2\delta^2}{144Cr^2\overline{L}_1}\right\}+2\exp\left\{-\frac{N^2c^2\delta^2}{144C\overline{L}_2}\right\}\\
&+2\exp\left\{-c^2N^{2-2\alpha}/\left(128C^2\sum_{j=1}^n R_j^2\right)\right\}+2N\exp\left\{-\frac{c}{24C\max_j R_j^{2-\epsilon}}\right\}\\
&+2\exp\left\{-\frac{Nc\epsilon^*}{6\sqrt{2C}\max_j R_j}\right\}. \qquad \text{(EC.12)}
\end{aligned}$$

□

We next prove Theorem 2 that says the upper bounds in (EC.8) and (EC.12) hold uniformly in $\sigma^0\in\mathcal{B}(N,r)$.

*Proof of Theorem 2* First, we recall that $U$ is an arbitrary subspace of $N\times n$ matrices with dimension $r$. Without loss of generality, we assume that $\|\Pi\|_F^2\le 1$ and let $M$ be the maximum of $\mathcal{A}(\Pi)$ for $\Pi\in U$. Theorem 1 demonstrates that (7) holds for each $Q$, and thus $(1-\delta/2)\|Q\|_F\le\|\mathcal{A}(Q)\|\le(1+\delta/2)\|Q\|_F$ for large enough $N$ and $n$ (replace $\delta$ in (7) by $\delta/2$). Then by the triangular inequality,

$$\|\mathcal{A}(\Pi)\|\le\|\mathcal{A}(Q)\|+\|\mathcal{A}(\Pi-Q)\|\le 1+\delta/2+M\delta/4\le 1+\delta,$$

where we noticed that $M\le 1+\delta$ due to $M\le 1+\delta/2+M\delta/4$, and

$$\|\mathcal{A}(\Pi)\|\ge\|\mathcal{A}(Q)\|-\|\mathcal{A}(\Pi-Q)\|\ge 1-\delta/2-(1+\delta)\delta/4\ge 1-\delta.$$

This proves that for any $\delta\in(0,1)$, for $N$ and $n$ large enough,

$$\begin{aligned}
&P_{R,D}\left(|\frac{\|\mathcal{A}(\Pi)\|_F^2-\|\Pi\|_F^2}{\|\Pi\|_F^2}|>\delta/2\text{ for some }\sigma^0\in U\right)\\
\le\ &2\exp\left\{-c^2r^2/\left(\sum_{j=1}^n R_j^2\right)^{1/2}\right\}+C\left(\frac{24}{\delta}\right)^r\exp\{-c^2\delta^2N^\alpha/\overline{L}_1\}. \qquad \text{(EC.13)}
\end{aligned}$$

Notice that the term $(\frac{24}{\delta})^r$ is only multiplied to the second term in the upper bound of Theorem 1 because the first term in the bound is irrelevant to $\sigma^0$. By Lemma 4.4 and equation (4.17) of Recht et al. (2010),

$$\sup_{\sigma^0\in\{U;\rho(U,U_i)\le\epsilon/2\}}|\frac{\|\mathcal{A}(\Pi)\|_F-\|\Pi\|_F}{\|\Pi\|_F}|\le\delta':=\delta/2+(\|\mathcal{A}\|+1)\epsilon. \qquad \text{(EC.14)}$$

To make $\delta' < \delta$, we should have $\|\mathcal{A}\| = \lambda_{\max}^{1/2}(AA') \le \delta/(2\epsilon) - 1$ where $AA'$ is a diagonal matrix. In the subsequent, we choose $\epsilon = (\delta/4)(\sqrt{Nn/p}+1)^{-1}$. By the definition of $\mathcal{A}$ and the Assumption on $\|\mathcal{A}\|$, we have

$$P_{R,D}\left(\|\mathcal{A}\| > \frac{\delta}{2\epsilon} - 1\right) \le P_{R,D}\left(\lambda_{\max}^{1/2}(AA') > 2\sqrt{Nn/n^* + 1}\right) \le C\exp\{-\gamma Nn\}. \tag{EC.15}$$

Combining (EC.14) and (EC.15), we have

$$P_{R,D}\left(\sup_{\sigma^0 \in \{U;\rho(U,U_i)\le \epsilon/2\}} |\frac{\|\mathcal{A}(\Pi)\|_F - \|\Pi\|_F}{\|\Pi\|_F}\| > \delta\right) \le C\exp\{-\gamma Nn\}.$$

By (EC.13) and the covering number of $\mathcal{B}(N,r)$,

$$\begin{aligned}
&P_{R,D}\left(|\frac{\|\mathcal{A}(\Pi)\|_F^2 - \|\Pi\|_F^2}{\|\Pi\|_F^2}| > \delta/2, \text{ for some } \sigma^0 \in U\right)\\
&\le 2\left(\frac{2C_0}{\epsilon}\right)^{r(N-r)}\left(\frac{24}{\delta}\right)^r \exp\left\{-c^2\delta^2 N^\alpha/\overline{L}_1\right\} + 2\exp\left\{-c^2r^2/\left(\sum_{j=1}^n R_j^2\right)^{1/2}\right\}\\
&\le C\exp\{-c^2\delta^2 N^\alpha/\overline{L}_1\} + 2\exp\left\{-c^2r^2/\sum_{j=1}^n R_j^2\right\},
\end{aligned}$$

for $c$ small enough and $C$ large enough which depend on $\delta$ and may change across lines, where in the last step, we have made use of the condition that

$$r(N-r)\log\left(\sqrt{\frac{Nn}{n^*}} + 1\right) = o\left\{\frac{N^\alpha}{\overline{L}_1}\right\}.$$

This proves (9).

Next, we prove the uniqueness. Suppose that there is another matrix $\Pi_1$ of rank at most $r$ so that $\mathcal{A}(\Pi_1) = b$ or $\mathcal{A}(\Pi_1) = b - \mathcal{A}(\Pi^*)$ and $\Pi_0 \ne \Pi_1$. Then $\Pi_1 - \Pi_0$ is a nonzero matrix of rank at most $2r$ and $\mathcal{A}(\Pi_1 - \Pi_0) = 0$. However, $0 = \|\mathcal{A}(\Pi_1 - \Pi_0)\| \ge (1-\delta)\|\Pi_1 - \Pi_0\|_F > 0$, where $\delta$ implicitly depends on $2r$ here. This contradiction shows the uniqueness of $\Pi$.

Similar to the proof of (10), due to

$$r(N-r)\log\sqrt{Nn/n^*} + 1 = o(N^{2-\alpha}/\overline{L}_1)$$

we have

$$\begin{aligned}
&P_{R,D}\left(|\frac{\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2}{\|\Delta\|_F^2}| > \delta \text{ for some } \sigma^0 \in \mathcal{B}(N,r)\right)\\
&\le P_{R,D}\left(|\|\mathcal{A}(\Delta)\|_F^2 - \|\Delta\|_F^2| > \frac{cN\delta}{2} \text{ for some } \sigma^0 \in \mathcal{B}(N,r)\right) + P_{R,D}\left(\|\Delta\|_F^2 \le \frac{Nc}{2}\right)\\
&\le 2\exp\left\{-\frac{N^{2-\alpha}c^2\delta^2}{144Cr^2\overline{L}_1}\right\} + 2\exp\left\{-\frac{N^2c^2\delta^2}{144C\overline{L}_2}\right\}\\
&\quad + 2\exp\left\{-c^2N^{2-2\alpha}/\left(128C^2\sum_{j=1}^n R_j^2\right)\right\} + 2N\exp\left\{-\frac{c}{24C\max_j R_j^{2-\epsilon}}\right\}\\
&\quad + 2\exp\left\{-\frac{Nc\epsilon^*}{6\sqrt{2C}\max_j R_j}\right\}.
\end{aligned}$$

Let $\Delta_1$ be another solution of (3), similarly, we have

$$0 = \|\mathcal{A}(\Delta_1) - \mathcal{A}(\Delta_0)\|^2 \geq (1-\delta)\|\Delta_1 - \Delta_0\|_F^2 > 0.$$

This proves the uniqueness of $\Delta$ solving (3).

□

*Proof of Theorem 3* Let $(\widehat{\Pi}_*, \widehat{\Delta}_*)$ be a linear combination of $(\widehat{\Pi}, \widehat{\Delta})$ and $(\widehat{\Pi}_0, \widehat{\Delta}_0)$, i.e., $(\widehat{\Pi}_*, \widehat{\Delta}_*) = \frac{\delta}{\beta}(\widehat{\Pi}, \widehat{\Delta}) + (1 - \frac{\delta}{\beta})(\widehat{\Pi}_0, \widehat{\Delta}_0)$ for $\beta > \delta$. Then we have $\|(\widehat{\Pi}_*, \widehat{\Delta}_*) - (\widehat{\Pi}_0, \widehat{\Delta}_0)\|_F = \delta$ and $\|(\widehat{\Pi}, \widehat{\Delta}) - (\widehat{\Pi}_0, \widehat{\Delta}_0)\|_F = \beta$. By the convexity of the function $G_n(\Pi, \Delta)$, we have

$$\frac{\delta}{\beta} G_n(\widehat{\Pi}, \widehat{\Delta}) + (1 - \frac{\delta}{\beta}) G_n(\widehat{\Pi}_0, \widehat{\Delta}_0) \geq G_n(\widehat{\Pi}_*, \widehat{\Delta}_*).$$

This together with (12) yields

$$\begin{aligned}
\frac{\delta}{\beta}(G_n(\widehat{\Pi}, \widehat{\Delta}) - G_n(\widehat{\Pi}_0, \widehat{\Delta}_0)) &\geq G_n(\widehat{\Pi}_*, \widehat{\Delta}_*) - G_n(\widehat{\Pi}_0, \widehat{\Delta}_0) \\
&\geq G_{n0}(\widehat{\Pi}_*, \widehat{\Delta}_*) - G_{n0}(\widehat{\Pi}_0, \widehat{\Delta}_0) - 2Na_{Nn} \\
&\geq c(1+\lambda)(\|\widehat{\Pi}_* - \widehat{\Pi}_0\|_F^2 + \|\widehat{\Delta}_* - \widehat{\Delta}_0\|) - 2Na_{Nn} \\
&= c(1+\lambda)\delta^2 - 2Na_{Nn}, \qquad \text{(EC.16)}
\end{aligned}$$

for some $c > 0$ with probability approaching one, where in the last step we have made use of Theorem 4.2 in Zhang and Zhang (2016). The inequality (EC.16) demonstrates that

$$0 \geq \frac{\delta}{\beta}\left[G_n(\widehat{\Pi}, \widehat{\Delta}) - G_n(\widehat{\Pi}_0, \widehat{\Delta}_0)\right] \geq c(1+\lambda)\delta^2 - 2Na_{Nn}.$$

This shows that $G_n(\Pi, \Delta)$ can not be minimized outside a $\delta$ neighborhood of $(\widehat{\Pi}_0, \widehat{\Delta}_0)$ when $\delta > \sqrt{\frac{cNa_{Nn}}{1+\lambda}}$, and hence $\|\widehat{\Pi} - \widehat{\Pi}_0\|_F = O_p\left(\sqrt{\frac{Na_{Nn}}{1+\lambda}}\right)$ and $\|\widehat{\Delta} - \widehat{\Delta}_0\|_F = O_p\left(\sqrt{\frac{Na_{Nn}}{1+\lambda}}\right)$.

□

*Proof of Theorem 4* The proof relies on a standard error decomposition. By adding and subtracting the oracle estimators $\widehat{\Delta}_0$ and $\widehat{\Pi}_0$, we can write:

$$\widehat{\Delta} - \Delta_0 = \widehat{\Delta} - \widehat{\Delta}_0 + \widehat{\Delta}_0 - \Delta_0, \quad \widehat{\Pi} - \Pi_0 = \widehat{\Pi} - \widehat{\Pi}_0 + \widehat{\Pi}_0 - \Pi_0.$$

Applying the triangle inequality to the norms of these expressions, the results of Theorem 4 follow directly from the bounds provided in Theorem 3 and the definitions of the oracle estimators $\widehat{\Delta}_0$ and $\widehat{\Pi}_0$.

□

*Proof of Theorem 5* The proof strategy is to decompose the total error into two components: the error of the de-biased estimator relative to the initial estimator, and the error of the initial estimator itself, which is bounded by Theorem 3. We begin with the following decompositions:

$$\widetilde{\Delta} - \Delta_0 = \widehat{\Delta} - \widehat{\Delta}_0 + \left[(\widetilde{\Delta} - \widehat{\Delta}) + (\widehat{\Delta}_0 - \Delta_0)\right], \tag{EC.17}$$

$$\widetilde{\Pi} - \Pi_0 = \widehat{\Pi} - \widehat{\Pi}_0 + \left[(\widetilde{\Pi} - \widehat{\Pi}) + (\widehat{\Pi}_0 - \Pi_0)\right]. \tag{EC.18}$$

We will analyze the second term on the right-hand side for both cases. Let $U_n$ and $V_n$ be the matrices of left and right singular vectors of $\widehat{\Delta}$, respectively. By the $SIN(\theta)$ theorem and Theorem 3,

$$\|U_n - U_*\|/\sqrt{N} + \|V_n - V_*\|/\sqrt{N} = O_p\left(\sqrt{\frac{a_{Nn}}{1+\lambda}}\right). \tag{EC.19}$$

Let $D_b$ be the difference between the matrices of the singular values of $\widetilde{\Delta}$ and $\widehat{\Delta}$, then we have by (EC.19)

$$\|\widetilde{\Delta} - \widehat{\Delta} + \widehat{\Delta}_0 - \Delta_0\| = \|U_n D_b V_n' - U_* D_b V_*'\| = O_P\left(\|D_b\|\sqrt{\frac{N a_{Nn}}{1+\lambda}}\right).$$

This together with (EC.17) and Theorem 3 proves the results for $\|\widetilde{\Delta} - \Delta_0\|$. For the closeness in the Frobenius norm, the steps are the same but with the operator norm replaced by the Frobenius norm.

The proofs for $\widetilde{\Pi}$ are similar to that for $\widetilde{\Delta}$ except for noticing that the difference between the matrices of the singular values of $\widetilde{\Pi}$ and $\widehat{\Pi}$ has $J$ nonzero singular values all equal to $\lambda(1 + 1/\mu)$, while the difference between the matrices of the singular values of $\widehat{\Pi}_0$ and $\Pi_0$ has the nonzero singular values being $(\underbrace{-\lambda(1+1/\mu), ..., -\lambda(1+1/\mu)}_{J}, -\lambda^*_{J+1}, ..., -\lambda^*_r)$ when $J < r$, being $(\underbrace{-\lambda(1+1/\mu), ..., -\lambda(1+1/\mu)}_{r}, \lambda^*_{r+1} - \lambda(1+1/\mu), ..., \lambda^*_J - \lambda(1+1/\mu))$ when $J > r$, and being $(\underbrace{-\lambda(1+1/\mu), ..., -\lambda(1+1/\mu)}_{r})$ when $J = r$. Then by the Wyle theorem and the $SIN(\theta)$ theorem,

$$\begin{aligned}
&\|\widetilde{\Pi} - \widehat{\Pi} + \widehat{\Pi}_0 - \Pi_0\|/\sqrt{N} \\
&= O_p\left\{\left[\lambda\left(1 + \frac{1}{\mu}\right) + 1\right]\sqrt{\frac{a_{Nn}}{1+\lambda}} + \frac{\lambda^*_{J+1} I(J<r) + \lambda^*_{r+1} I(J>r) + 0 I(J=r)}{\sqrt{N}}\right\},
\end{aligned}$$

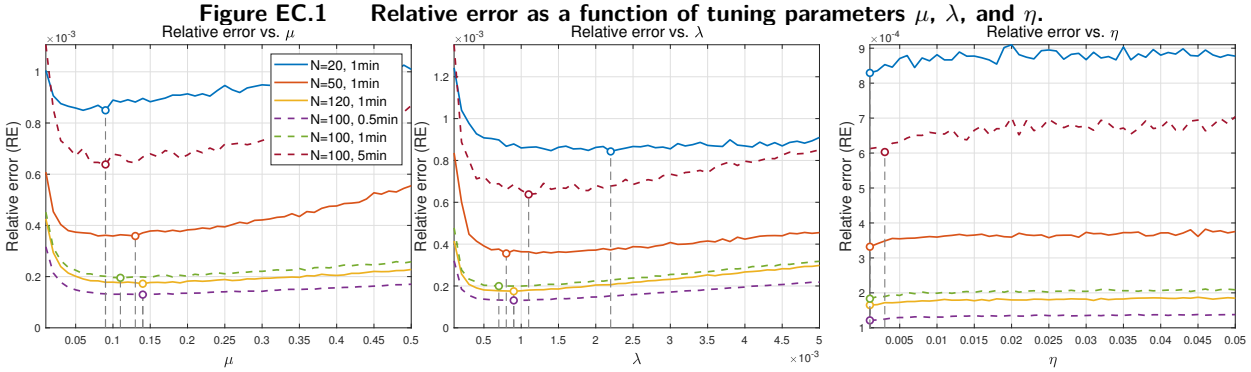
and replacing the operator norm by the Frobenius norm,

$$\begin{aligned}
&\|\widetilde{\Pi} - \widehat{\Pi} + \widehat{\Pi}_0 - \Pi_0\|_F/\sqrt{N} \\
&= O_p\left\{\left[\lambda\left(1 + \frac{1}{\mu}\right) + 1\right]\sqrt{\frac{a_{Nn}}{1+\lambda}} + \frac{\sum_{l=J+1}^{r} \lambda^*_l I(J<r) + \sum_{l=r+1}^{J} \lambda^*_{r+1} I(J>r) + 0 I(J=r)}{\sqrt{N}}\right\}.
\end{aligned}$$

This together with (EC.18) and Theorem 3 proves the results for $\widetilde{\Pi} - \Pi_0$.

□

**Figure EC.1 Relative error as a function of tuning parameters $\mu$, $\lambda$, and $\eta$.**




*Note.* The relative error is calculated according to (6). The vertical dotted line in each panel indicates the parameter value that minimizes the error.

## EC.2. Additional Simulation

In this subsection, we show the results of another tuning parameter selection in the main text, based on the relative error, as in Figure EC.1.

The analysis based on relative errors yields conclusions that are largely consistent with those derived from absolute errors. This consistency across different error metrics further underscores the robustness of our findings and the stability of the method with respect to its tuning parameters.

## EC.3. Additional Empirical Analysis

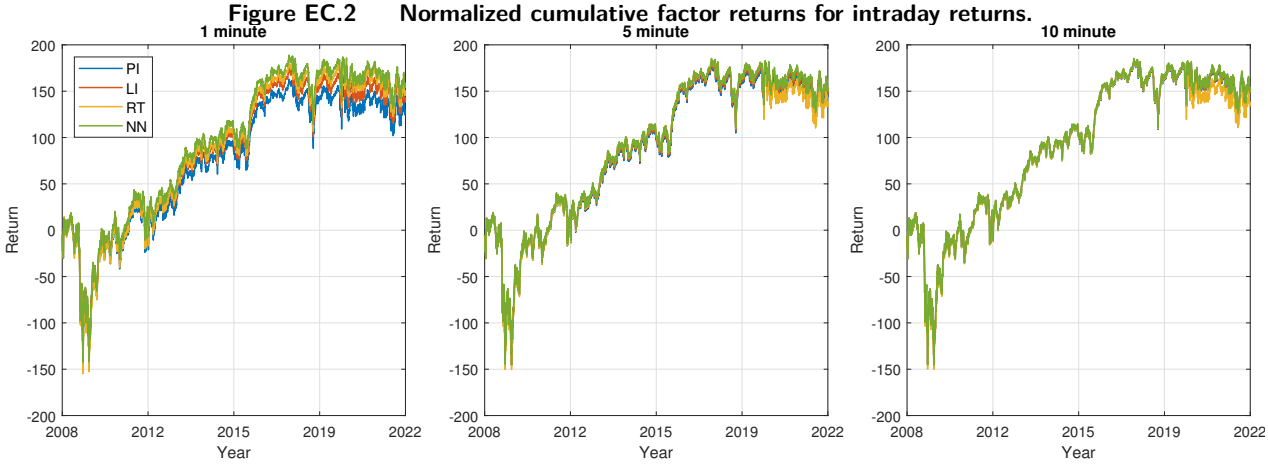
### EC.3.1. Factors

In this subsection, we further evaluate the economic implications of different imputation methods by constructing and analyzing a high-frequency return factor. Specifically, we extract the first principal component (the "market factor") from the imputed intraday return matrices and compute its cumulative performance over time, excluding overnight returns. Figure EC.2 plots the cumulative return of this factor, derived from data imputed by four different methods (NN, PI, LI, RT), across three sampling frequencies. As expected, the first latent factor captures a significant portion of the total variation under all imputation methods.

The results for the 1-minute frequency reveal substantial performance discrepancies among the methods. The factor constructed from our NN-imputed data yields the highest cumulative return over the 15-year sample period. In contrast, the factor derived from the PI method performs the worst. The performance gap is economically significant: the cumulative return of the top-performing factor (NN) exceeds that of the worst-performing factor (PI) by more than 30%.

At lower frequencies (5-minute and 10-minute), the performance gap between the NN, PI, and LI methods narrows considerably. This is consistent with our earlier findings that the distorting effects of price staleness are less severe in lower-frequency data. However, the factor derived from the RT method continues to exhibit markedly inferior performance. This is attributable to the inherent

**Figure EC.2 Normalized cumulative factor returns for intraday returns.**




drawback of the RT scheme, which discards a vast amount of valid price information by subsampling to a sparse, synchronized time grid, thereby compromising the quality of the extracted factor.

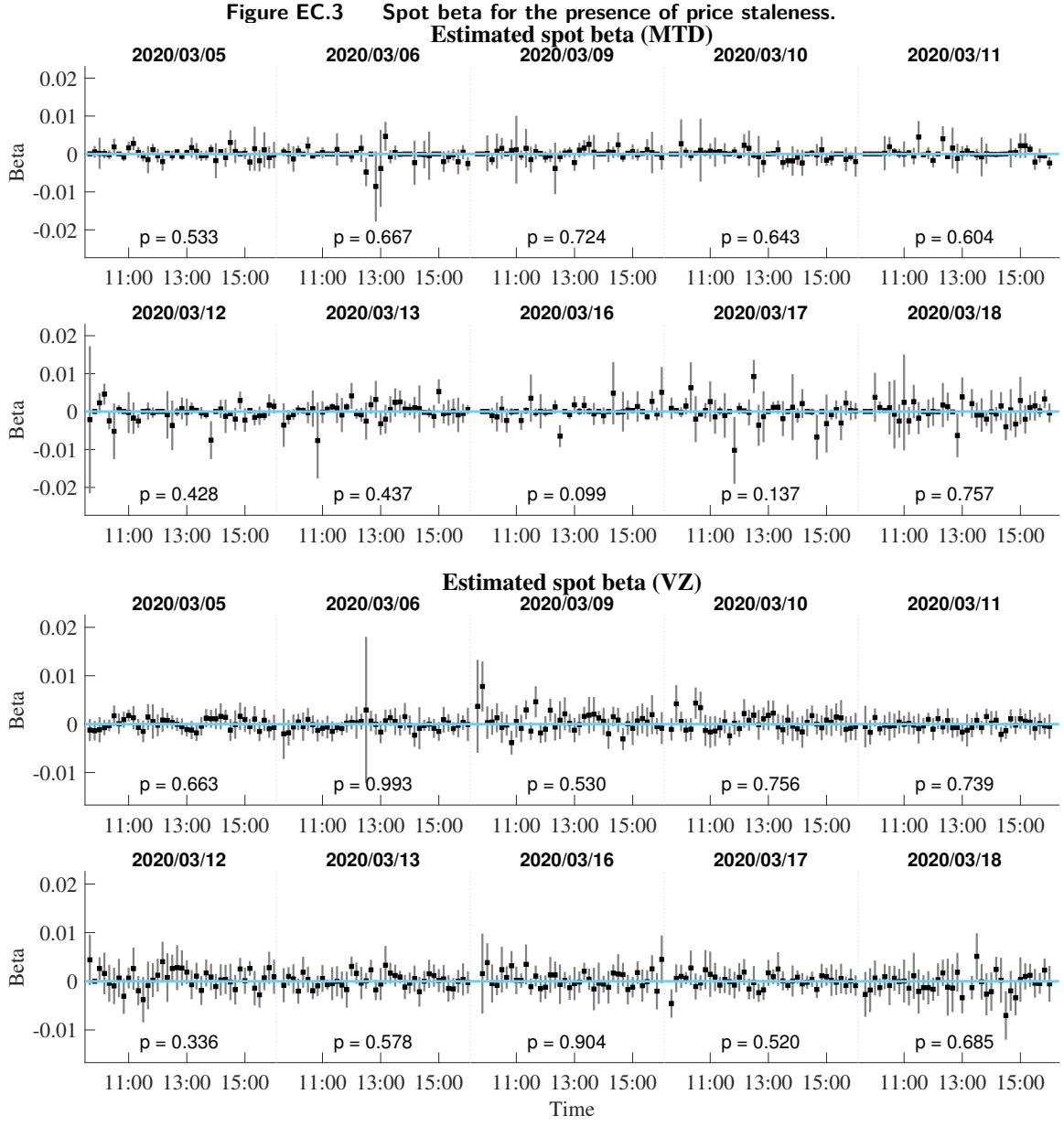
### EC.3.2. Additional Beta Results

Building on our initial analysis using a 15-minute estimation window, we now conduct a series of robustness checks by re-estimating the spot betas for Mettler-Toledo International (MTD) and Verizon Communications (VZ) using 5-minute (Figures EC.3 and EC.4) and 10-minute (Figures EC.5 and EC.6) windows. This multi-window approach allows us to directly examine the practical implications of the bias-variance trade-off discussed in Bollerslev et al. (2024) and to verify the stability of our core conclusions. The findings strongly reaffirm our initial results and provide deeper insights into the properties of the estimation methods.

Our initial analysis revealed that the "previous-tick" method, which generates stale prices, produces artificial zero-beta estimates. The robustness checks demonstrate that this is not an idiosyncratic issue tied to a 15-minute window but a fundamental flaw of the method. Comparing the "price staleness" plots across windows (Figures EC.3, EC.4, and the original 15-minute plot) reveals that the problem of zero betas for the less-liquid MTD becomes more severe as the estimation window shrinks. In the 5-minute window (Figure EC.3), there are visibly more instances where the beta estimate collapses to zero than in the 10- or 15-minute windows. This is perfectly logical: the probability of a stock not trading is higher over a 5-minute interval than a 15-minute one. This confirms that the zero-beta phenomenon is a direct, mechanical consequence of price staleness, as first explored in early literature by Scholes and Williams (1977). Across all window sizes, the statistical inference from the price-staleness method remains unreliable.

As expected from theory, the spot beta estimates become smoother as the window size $k$ increases. The 5-minute beta paths (Figure EC.4) are visibly "noisier" and have wider confidence intervals, reflecting higher variance from using fewer observations. The 10-minute paths (Figure EC.5) are

**Figure EC.3 Spot beta for the presence of price staleness.**

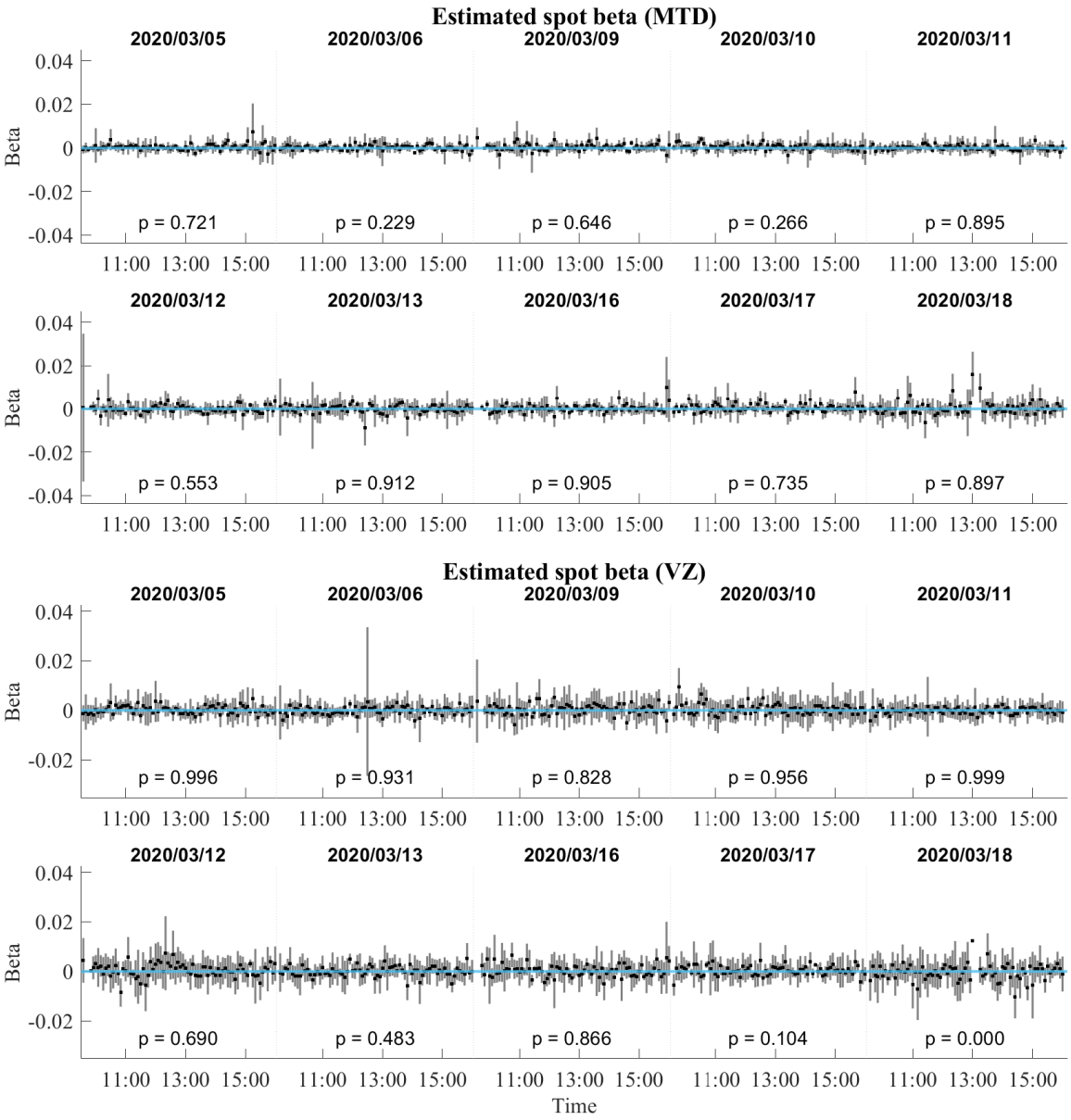



*Note.* This figure plots the estimated spot betas of Mettler-Toledo International (MTD) and Verizon Communications (VZ) against the SPY. The betas are estimated using 1-minute price data over 5-minute rolling windows, along with their corresponding 90% confidence intervals. The analysis covers the two-week period of high market volatility from March 5, 2020, to March 18, 2020. The $p$-value reported in each panel corresponds to a test of the functional null hypothesis that the entire spot beta process for a given day is equal to zero ($H_0 : \beta_t = 0$ for all $t$).

smoother, and the original 15-minute paths are the smoothest of all, reflecting lower estimation variance. This is the classic bias-variance trade-off in action. Despite the differences in variance, the underlying economic narrative remains remarkably consistent across all windows for the our method.

The consistency of our findings across different temporal resolutions greatly strengthens our conclusions about intraday risk dynamics, a topic of growing interest (e.g., Andersen et al. 2021). The

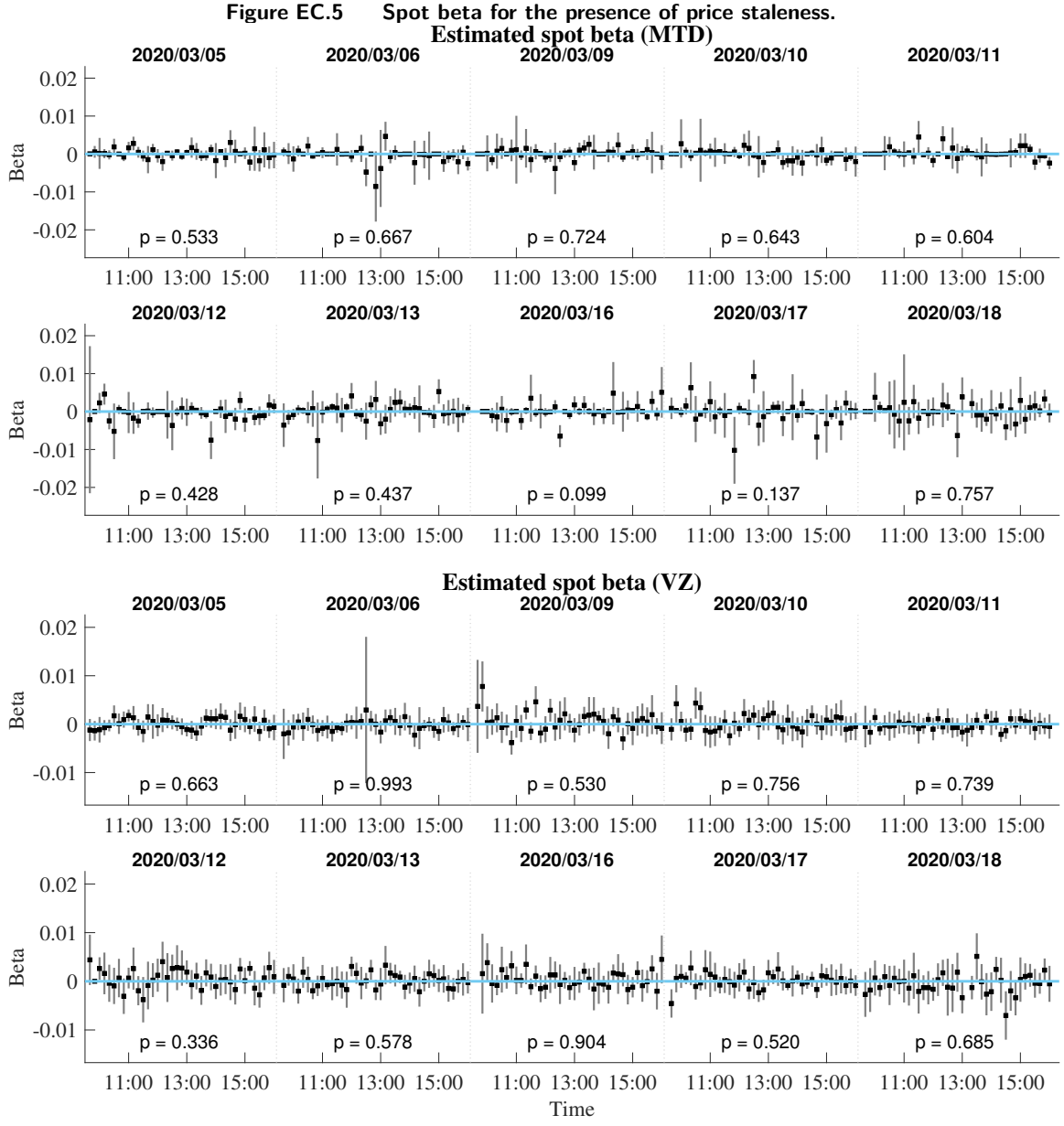
**Figure EC.4 Spot beta for the absence of price staleness.**




*Note.* This figure plots the estimated spot betas of Mettler-Toledo International (MTD) and Verizon Communications (VZ) against the SPY. The betas are estimated using 1-minute price data over 5-minute rolling windows, along with their corresponding 90% confidence intervals. The analysis covers the two-week period of high market volatility from March 5, 2020, to March 18, 2020. The $p$-value reported in each panel corresponds to a test of the functional null hypothesis that the entire spot beta process for a given day is equal to zero ($H_0 : \beta_t = 0$ for all $t$).

general intraday shapes of the beta paths, especially on the most volatile days, are preserved across the 5, 10, and 15-minute horizons. This suggests that the primary intraday risk fluctuations for these stocks occur at a frequency that is well-captured even by a 15-minute window. A shorter window like 5 minutes provides a more granular view but confirms the same broader patterns, increasing our confidence that these are genuine features of the market and not artifacts of a specific window choice.

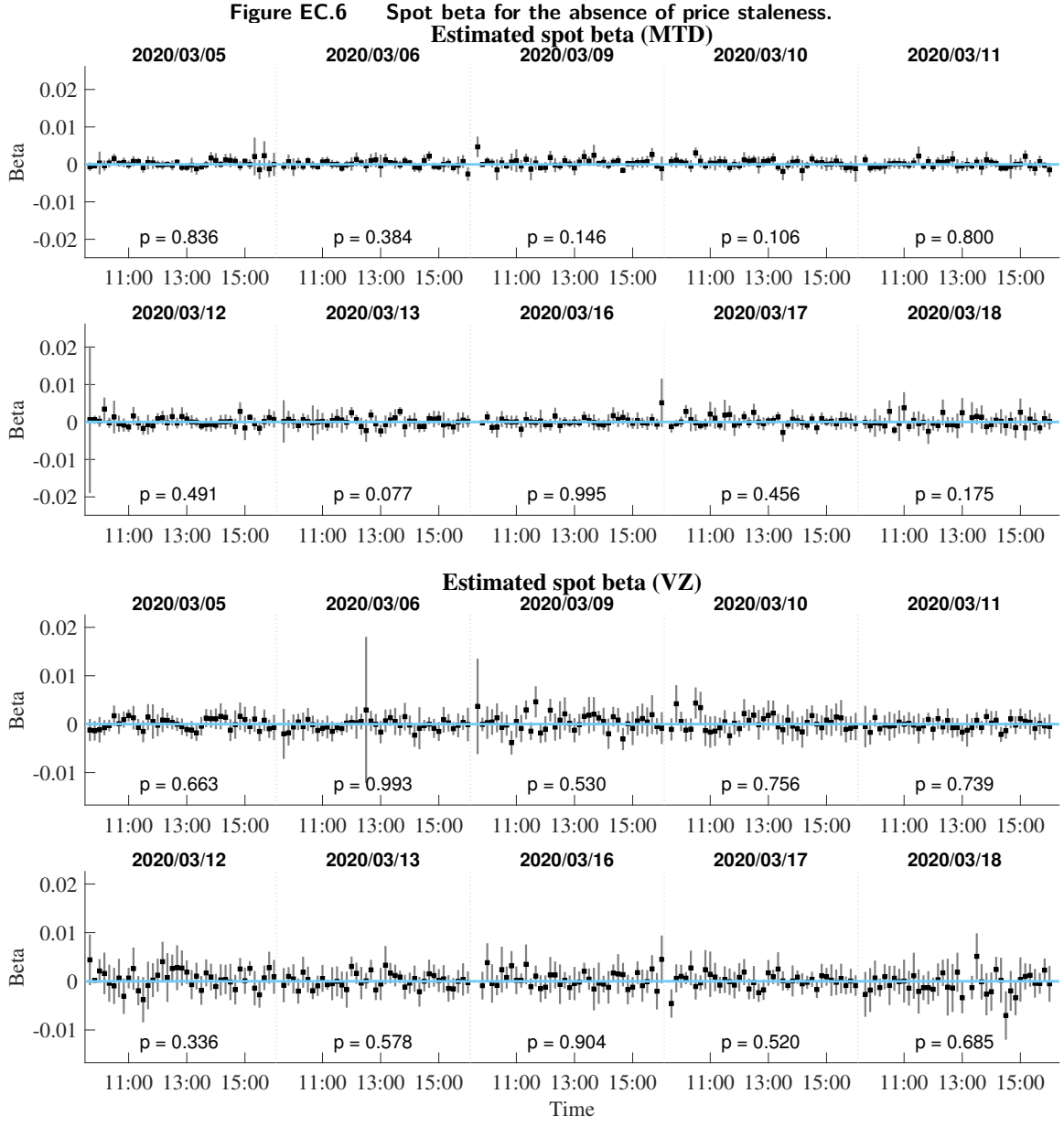
**Figure EC.5 Spot beta for the presence of price staleness.**




*Note.* This figure plots the estimated spot betas of Mettler-Toledo International (MTD) and Verizon Communications (VZ) against the SPY. The betas are estimated using 1-minute price data over 10-minute rolling windows, along with their corresponding 90% confidence intervals. The analysis covers the two-week period of high market volatility from March 5, 2020, to March 18, 2020. The $p$-value reported in each panel corresponds to a test of the functional null hypothesis that the entire spot beta process for a given day is equal to zero ($H_0 : \beta_t = 0$ for all $t$).

This analysis provides practical guidance. For assets where theory or observation suggests very high-frequency changes in risk exposure, a smaller $k$ (like 5 minutes) might be preferable, despite higher variance. For assets with smoother risk profiles or for analyses focused on broader intraday trends, a larger $k$ (like 15 minutes) can provide a clearer signal by reducing noise. The fact that

**Figure EC.6 Spot beta for the absence of price staleness.**




*Note.* This figure plots the estimated spot betas of Mettler-Toledo International (MTD) and Verizon Communications (VZ) against the SPY. The betas are estimated using 1-minute price data over 10-minute rolling windows, along with their corresponding 90% confidence intervals. The analysis covers the two-week period of high market volatility from March 5, 2020, to March 18, 2020. The $p$-value reported in each panel corresponds to a test of the functional null hypothesis that the entire spot beta process for a given day is equal to zero ($H_0 : \beta_t = 0$ for all $t$).

our conclusions hold across this range indicates the robustness of both the underlying economic phenomena and our methodology itself.

The comprehensive robustness analysis confirms and strengthens our initial findings. We have shown that the flaws of the "price staleness" method are systematic, while the robust estimation procedure provides stable and economically meaningful results that are not sensitive to the specific

choice of the estimation window. This demonstrates the power and reliability of the framework proposed by Bollerslev et al. (2024), providing a credible tool for uncovering the rich, dynamic nature of systematic risk at high frequencies, even under the most extreme market conditions.