Beyond Returns: A Candlestick-Based Approach to Covariance Estimation*

Preliminary and incomplete draft. Please do not cite without permission.

Yasin Simsek

Department of Economics, Duke University

This version: February 19, 2025

Abstract

This paper proposes a novel framework for estimating asset return covariances using high-frequency candlestick data. We introduce univariate and multivariate realized measures derived from high-frequency candlestick observations and establish their consistency and asymptotic normality within a standard in-fill asymptotic framework. The first-order asymptotic results highlight the connection between candlestick data and dependence structure of asset returns. Drawing from these findings, we develop an integrated covariance matrix estimator which is formulated as the linear combination of candlestick based realized estimators. We apply the new estimator in the construction of global minimum variance portfolios using high-frequency price data of S&P 500 stocks between January 1996 and December 2023. Our approach significantly reduces portfolio risk in the out-of-sample while maintaining lower turnover, smaller aggregate short positions, and less extreme exposures to individual assets compared to standard procedures. We also find that this outperformance carries over effectively to low-frequency datasets (e.g., daily CRSP). Finally, we find that the stochastic discount factors implied by our novel estimator yields significantly smaller cross-sectional pricing errors for various factor datasets. Our findings underscore the broader implications of our approach in asset pricing as well as portfolio management.

Keywords: Covariance estimation, high-frequency candlestick data, portfolio selection.
J.E.L. Codes: C13, C32, C58, G11.

*We thank Tim Bollerslev, Anna Bykhovskaya, Andrew Patton, Peter Hansen, George Tauchen and seminar participants at Duke University, Email: yasin.simsek@duke.edu.

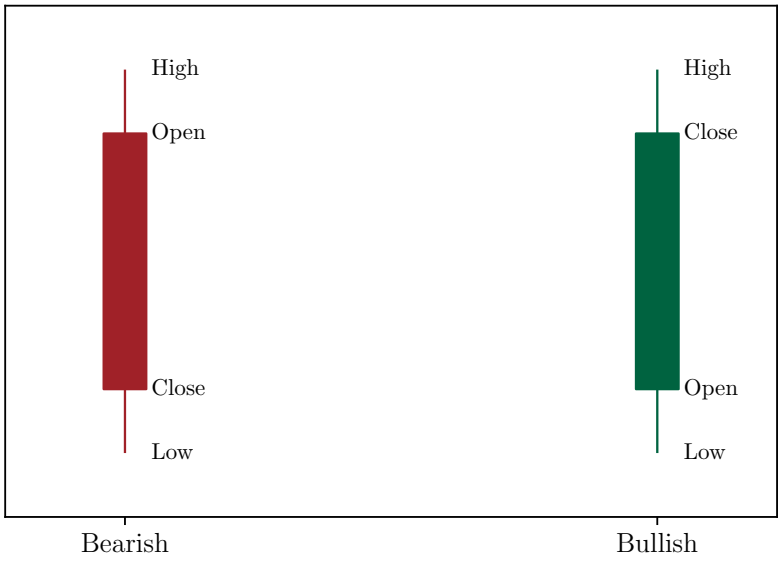
1 Introduction

The covariance matrix of asset returns is fundamental to many financial applications, such as portfolio optimization, risk management, and cross-sectional asset pricing. Despite its importance, estimating covariances is notoriously difficult due to factors including the large number of parameters involved, the inherent noise in asset returns, and the challenge of accurately capturing time-varying dynamics. Notably, following the seminal works of Andersen et al. (2003); Barndorff-Nielsen and Shephard (2002, 2004), a substantial body of econometric literature has emerged to study the modeling, estimation, and prediction of asset return covariances using high-frequency financial data (see Aït-Sahalia and Jacod (2014) and references therein). So far, research papers in this area have predominantly focused on returns within a trading period, commonly defined as the logarithmic difference between the open (first) and close (last) price observations. While widely adopted, this approach overlooks covariance information embedded in intra-period price dynamics, i.e., observed prices between the first and last one.

Building on this background, this paper introduces a novel framework for more accurately estimating covariances by explicitly leveraging intra-period price dynamics. The price movements within a period are often represented by candlestick charts in financial markets. A typical candlestick contains four different price observations over a period of interest: open, high, low, and close prices; see Figure 1 for a concrete illustration. These charts are widely available through online platforms, including Yahoo Finance, Robinhood, and TradingView, and serve as fundamental tools in technical analysis.

We develop univariate and multivariate realized measures derived from high-frequency candlestick observations. Subsequently, we establish the consistency and asymptotic normality of these measures within a standard in-fill asymptotic framework. Our first-order asymptotic results show the connection between candlestick-based realized es-

Figure 1: **Candlestick Examples**. The figure shows examples of bearish and bullish candlesticks. The open, high, low, and close prices are also highlighted.



timators and the dependence structure of asset returns. Driven by this theoretical finding, we introduce a novel integrated covariance matrix estimator, which is formulated as a linear combination of the initial estimators. This combination is designed to fully leverage the informational content inherent in the candlestick structure while maintaining computational efficiency, ultimately reducing estimation error in comparison to conventional benchmarks.

Our framework closely follows the methodology proposed in Li et al. (2024), where the authors introduce a linear combination of range and absolute return for spot volatility estimation, with the optimal weights determined by minimizing the asymptotic variance of the estimator. We generalize this approach to the covariance estimation problem; however, the resulting optimal weights depend on unobservable quantities, rendering the approach infeasible in practice. To address this challenge, we exploit a special case of data-generating process that eliminates dependence on unobservable components, thereby yielding a feasible and theoretically grounded weighting scheme.

Through a realistic simulation design, we demonstrate that our estimator closely approximates the theoretically optimal estimator regarding root mean squared error, leaving limited scope for further refinement.

The adverse effects of poorly estimated covariance matrix have been widely recognized by the literature, for example, in the context of portfolio selection. As Li (2015) and the references therein show, the estimation error in covariances often leads to highly leveraged positions on certain assets[1], consequently diverging from "true" optimal weights and offering low reward-risk ratio in the out-of-sample. Indeed, our approach may be seen as a way to address these issues by integrating candlestick data in the estimation process, ultimately enhancing the information content of the training data and yielding higher-quality covariance estimates. To elucidate this idea, we conduct an empirical analysis on estimating global minimum variance (GMV) portfolios. We study the high-frequency price data of S&P 500 stocks between 1996 and 2023. Using our novel approach, we construct minimum-variance portfolios with various sizes and compare their out-of-sample performance with standard benchmarks like realized covariance estimator and equally-weighted portfolio. Empirical results show that our approach achieves significantly reduced portfolio risk in the out-of-sample while maintaining lower turnover, smaller aggregate short positions, and less extreme exposures to individual assets.

To further substantiate these findings, we extend our analysis to the construction of GMV portfolios using low-frequency price data. Low-frequency data, such as daily stock prices, is widely used by investors due to its accessibility and availability relative to high-frequency datasets. For this analysis, we utilize daily stock price data sourced from the Center for Research in Security Prices (CRSP), a well-known academic data provider. Our results demonstrate that the proposed estimator consistently outperforms standard benchmarks in terms of out-of-sample performance, mirroring the

[1] Kelly et al. (2024) argue that efficient portfolios take large bets on low-variance principal component portfolios. This leads to poor out-of-sample performance since these portfolios have small risk premia.

advantages observed in the high-frequency context. These findings highlight the robustness and generalizability of our approach across varying data frequencies, reinforcing its practical applicability in diverse investment scenarios.

In asset pricing models, similar to portfolio selection problem, accurate covariance estimation is crucial for identifying systematic risk factors and their associated risk premia. Specifically, estimation errors in betas, reflecting the covariance between asset returns and systematic risk factors, can hinder reliable inference on the pricing factors and the risk premia, resulting in misleading conclusions about the cross-section of asset prices, see for example Shanken (1992); Jagannathan and Wang (1996); Kan and Zhang (1999); Gospodinov et al. (2014); Kleibergen (2009). Our approach offers a new perspective on this problem by providing a more accurate estimation of covariances. To see this in action, we construct the corresponding stochastic discount factors (SDF) using our candlestick-based covariance estimator which is estimated from S&P 500 stocks between 1996 and 2023. We then compare the results with that of realized covariance. We study recently created factor data sets by Jensen et al. (2023) and Chen and Zimmermann (2021) as test assets. The former dataset includes 153 factors, while the latter dataset contains 212 factors. Our empirical analysis shows that our approach yields significantly smaller cross-sectional pricing errors for both set of factors, showing strength of SDF based on candlestick returns. Furthermore, we investigate the performance of constructed SDFs across different subgroups determined by factor themes. We find that our results appear to be broadly consistent across different factor themes, which suggests additional robustness to our approach.

This paper contributes to multiple strands of the literature. Our work inspired from the range-based volatility estimation literature. Starting from seminal papers Garman and Klass (1980) and Parkinson (1980), this literature highlights remarkable efficiency gains in estimating variances by extracting more information from price paths. Notably, Christensen and Podolskij (2007) introduces the realized-range estimator, constructed

by high-frequency price ranges. Later, it is extended to be robust to microstructure noise (Martens and Van Dijk (2007); Christensen et al. (2009)); jumps (Christensen and Podolskij (2012)) and drifts (Li et al. (2023)). While these papers focus on the standard in-fill asymptotics, several works propose *fixed-k* inference on candlestick based volatility estimators, see Li et al. (2024), Bollerslev et al. (2024a), Bollerslev et al. (2024b) and Bollerslev et al. (2024c). Particularly, our econometric framework is closely related to Li et al. (2024) which offers linear combinations of candlestick-based spot volatility estimates, meeting certain optimality conditions. Our paper complements this literature by extending these ideas to estimating covariances. This allows us to incorporate high-frequency candlestick data into various asset pricing applications, not only estimation and inference of volatility.

Prior to this study, Brandt and Diebold (2006) and Bannouh et al. (2009) presented a multivariate range-based covariance estimator by utilizing triangular no-arbitrage conditions. In contrast, our paper does not depend on this type of no-arbitrage condition, which may not be applicable to stock markets. Additionally, Rogers and Zhou (2008) estimated the correlation coefficient of a bivariate Brownian motion using candlestick observations. We depart from this approach by offering a realized-type estimator for the integrated covariance matrix of an Itô-semimartingale process, along with its in-fill asymptotic results. Furthermore, we conducted a comprehensive empirical analysis to demonstrate the advantages of our method.

From a technical perspective, our paper is closely related to the high-frequency econometrics literature. Estimating and making inferences on volatility with high-frequency data dates back to Andersen et al. (2003) and Barndorff-Nielsen and Shephard (2002). Since then, various aspects of high-frequency data have been explored by researchers; see, for example, Bollerslev (2022) and references therein. Similarly, Barndorff-Nielsen and Shephard (2004) extends high-frequency variance estimation to multivariate settings, resulting in ex-post covariation estimators. A rich collection of

these estimators has been offered recently. For instance, some papers from this line of work address the microstructure noise and synchronicity problems commonly observed in high-frequency prices. Notable references could be Aït-Sahalia et al. (2010); Christensen et al. (2010); Barndorff-Nielsen et al. (2011); Zhang (2011); Bibinger et al. (2014); Shephard and Xiu (2017), among many others. More recently, researchers have been developing sophisticated frameworks to address high-dimensionality in covariance structures of high-frequency returns. For example, Fan et al. (2016) proposes a block-diagonal structure using industry classifications to estimate vast covariance matrices; Dai et al. (2019) and Bollerslev et al. (2019) study on factor model-based estimators of realized covariances. Moreover, the benefits of using high-frequency methods in portfolio allocation and cross-sectional pricing are thoroughly studied in the literature, see for example Fleming et al. (2003); Bandi et al. (2008); Pooter et al. (2008); Liu (2009); Fan et al. (2012); Hautsch et al. (2015); Bollerslev et al. (2016, 2018a); Cai et al. (2020); Kim and Oh (2024). We complement this strand of literature by proposing the use of candlestick observations for estimating ex-post covariation of Ito-semimartingale processes. To the best of our knowledge, this paper is the first study exploring the asymptotic properties and empirical applications of multivariate realized candlestick measures in order to unveil the information embedded in candlesticks.

The rest of the paper is organized as follows: Section 2 discusses the asymptotic properties of candlestick-based realized measures and describes the optimal estimator for covariances, Section 3 presents the results of empirical applications and finally Section 4 concludes. Additional technical details, including proofs, are deferred to appendices.

2 Econometric Framework

2.1 The continuous-time price model

We denote the log-price vector by $\boldsymbol{X}_t = [X_{1,t}, ..., X_{N,t}]^\top$ for $0 \leq t \leq T$. Suppose that $\boldsymbol{X}_t$ is an Ito semimartingale defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$,

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{b}_s ds + \int_0^t \boldsymbol{\sigma}_s d\boldsymbol{W}_s \tag{1}$$

where $\boldsymbol{b}_s$ and $\boldsymbol{\sigma}_s$ stand for instantaneous drift and stochastic volatility processes, respectively, and $\boldsymbol{W}_s$ is an N-dimensional vector of independent Brownian motions. We exclude jumps for simplicity, though our framework can be extended to more general price processes that contain jumps[2].

The spot covariance matrix at a given time point s, described as $\boldsymbol{c}_s \equiv \boldsymbol{\sigma}_s \boldsymbol{\sigma}_s^\top$ and further define

$$\nu_{j,s} \equiv \sqrt{c_{jj,s}} \quad \text{and} \quad \rho_{jk,s} \equiv \frac{c_{jk,s}}{\nu_{j,s}\nu_{k,s}} \tag{2}$$

in which $\nu_{j,s}$ denotes the spot volatility of asset j, and $\rho_{jk,s}$ stands for the spot correlation coefficient between assets j and k.

The goal of this study is to estimate ex-post covariation, commonly called *integrated covariance* of price process $\boldsymbol{X}_t$. The integrated covariance, is formally defined as

$$ICV_T \equiv \int_0^T \boldsymbol{c}_s ds. \tag{3}$$

over the time period $[0, T]$. In a similar vein, we can describe *integrated variance* for individual asset j:

$$IV_{j,T} \equiv \int_0^T \nu_{j,s}^2 ds. \tag{4}$$

[2]Jumps have distinct dynamics compared to the continuous part of the price process. For example, jumps are less persistent than continuous returns, and they are usually associated with news arrivals such as earnings announcements and macroeconomic announcements (see Andersen et al. (2007) and Lee and Mykland (2008)). Therefore, one can separate the jump component using, for example, the truncation method of Mancini (2009) and then apply the same ideas to the continuous component.

corresponds to the j-th diagonal element of ICV_T. Throughout the paper, T is fixed, and we suppress the subscript T for simplicity for the rest of the paper.

The Itô-semimartingale representation in Equation 1 is motivated by no-arbitrage arguments. It serves as a workhorse model in continuous-time finance and economics thanks to its applicability and flexibility; for a discussion, see Jacod and Protter (2012). However, it is well known that the model in Equation 1 is misspecified in the presence of microstructure noise, and, therefore, we treat this model as a reasonable approximation to reality. In line with standard practice, we rely on "not-too-finely" sampled price data to circumvent these difficulties.

2.2 The observation scheme and notation

We implement a standard equidistant observation scheme: The price process $\boldsymbol{X}_t$ is sampled on a regular time grid, $\{i\Delta_n : i = 0, 1, ..., nT\}$ where $\Delta_n = 1/n$ be the sampling frequency and $T > 0$ be a fixed time span, usually taken as a month or a day. We denote a typical sampling interval $[(i-1)\Delta_n, i\Delta_n]$ by $\mathcal{T}_i$.

The high-frequency return over period $\mathcal{T}_i$ defined as:

$$\boldsymbol{r}_i \equiv \boldsymbol{X}_{i\Delta_n} - \boldsymbol{X}_{(i-1)\Delta_n} \tag{5}$$

where we use first and last price observations within a Δ_n-time interval. The existing high-frequency econometrics literature on covariance estimation has primarily focused on estimators based on high-frequency returns. Motivated by candlesticks, this paper augments the information set for covariance estimation by incorporating intra-interval price dynamics. Specifically, for a given Δ_n-time interval, we can obtain the highest and lowest prices, denoted by $\sup_{t\in\mathcal{T}_i} \boldsymbol{X}_t$ and $\inf_{t\in\mathcal{T}_i} \boldsymbol{X}_t$. Two new high-frequency

returns can summarize this additional information:

$$\begin{aligned} \boldsymbol{h}_i &\equiv \sup_{t\in\mathcal{T}_i} \boldsymbol{X}_t - \boldsymbol{X}_{(i-1)\Delta_n} \\ \boldsymbol{\ell}_i &\equiv \inf_{t\in\mathcal{T}_i} \boldsymbol{X}_t - \boldsymbol{X}_{(i-1)\Delta_n} \end{aligned} \tag{6}$$

where $\boldsymbol{h}_i$ and $\boldsymbol{\ell}_i$ refer to high and low returns over $\mathcal{T}_i$, respectively. Note that sup and inf operators are applied element-wise.

To assist subsequent discussion in the paper and remain consistent with the literature, we define two new variables, inspired by Bollerslev et al. (2024a),

$$\boldsymbol{a}_i \equiv \boldsymbol{h}_i + \boldsymbol{\ell}_i - \boldsymbol{r}_i \quad \text{and} \quad \boldsymbol{w}_i \equiv \boldsymbol{h}_i - \boldsymbol{\ell}_i \tag{7}$$

where $\boldsymbol{a}_i$ and $\boldsymbol{w}_i$ are called asymmetry and range, respectively. If we consider candlestick shapes, range $\boldsymbol{w}_i$ corresponds to the length of the candlestick, and asymmetry $\boldsymbol{a}_i$ captures the difference between the length of the upper and lower wicks of the candlestick. Wicks refer to the thin lines above and below the body of the candlestick; see Figure 1. Intuitively, the length of the candlestick, and so range, shows the maximum span of asset price over the trading period, and therefore $\boldsymbol{w}_i$ can be related to the variability of the asset returns. Furthermore, the wicks of the candlestick demonstrate how much the price reverses back from its high and low attained in that period, indicating that $\boldsymbol{a}_i$ might reflect the fluctuations in asset prices, similar to the range variable. As a result, a high-frequency candlestick provides three different returns (or information) potentially related to the price movements over a period of interest.

Lastly, to help the discussion in the theoretical results, we define counterparts of $\boldsymbol{a}_i$ and $\boldsymbol{w}_i$ for a generic standard Brownian motion on $[0,1]$ interval. Specifically, suppose that $\widetilde{W}_t$ is a standard Brownian motion over $[0,1]$ interval. Corresponding asymmetry

and range variables can be defined simply as:

$$\begin{aligned} \widetilde{a} &= \sup_{t\in[0,1]} \widetilde{W}_t - \widetilde{W}_0 + \inf_{t\in[0,1]} \widetilde{W}_t - \widetilde{W}_0 - (\widetilde{W}_1 - \widetilde{W}_0) \\ \widetilde{w} &= \sup_{t\in[0,1]} \widetilde{W}_t - \widetilde{W}_0 - \inf_{t\in[0,1]} \widetilde{W}_t - \widetilde{W}_0. \end{aligned} \tag{8}$$

2.3 Realized candlestick measures: Univariate case

To fix the ideas, we begin by defining estimators for the integrated variance of a single asset based on candlestick information. This can be thought of as a special case of our multivariate framework when $N = 1$.

The classical nonparametric estimator of integrated variance, *realized variance*, is developed by seminal papers Andersen et al. (2001, 2003) and Barndorff-Nielsen and Shephard (2002). We formulate *realized variance* of asset j at sampling frequency Δ_n as:

$$\widehat{C}_j = \sum_{i=1}^{T/\Delta_n} r_{j,i}^2 \tag{9}$$

where $r_{j,i}$ is the high-frequency return of asset j over the interval $\mathcal{T}_i$ defined in Equation 5.

Drawing on classical works of Parkinson (1980) and Garman and Klass (1980), Christensen and Podolskij (2007) propose *realized range* estimator for the integrated variance of asset j using high-frequency range $\boldsymbol{w}_i$ in Equation 7, which can be expressed as follows:

$$\widehat{W}_j = \frac{1}{\gamma_{w,2}} \sum_{i=1}^{T/\Delta_n} w_{j,i}^2 \tag{10}$$

where $\gamma_{w,p} = \mathbb{E}[\widetilde{w}^p]$ and $\gamma_{w,2} \approx 2.7719$. Inspired by realized variance and realized range estimator, we can define a new estimator *realized asymmetry* exploiting asymmetry variable, $\boldsymbol{a}_i$, formulated in Equation 7. This new integrated variance estimator simply equals to

$$\widehat{A}_j = \frac{1}{\gamma_{a,2}} \sum_{i=1}^{T/\Delta_n} a_{j,i}^2 \tag{11}$$

where $\gamma_{a,p} = \mathbb{E}[\widetilde{a}^p]$ and $\gamma_{a,2} \approx 0.2273$. Note that the numerical values of $\gamma_{w,2}$ and $\gamma_{a,2}$ are obtained by computing the moments using simulated paths of a generic Brownian Motion on $[0, 1]$. Also, these coefficients bring the estimators $\widehat{W}_j$ and $\widehat{A}_j$ to the same scale as $\widehat{C}_j$, allowing direct comparison between them.

The consistency of $\widehat{C}_j$ immediately follows since it is built on the theory of *quadratic variation (QV)*, see for example Andersen et al. (2003). Moreover, Christensen and Podolskij (2007) shows that $\widehat{W}_j$ is consistent for IV_j under some regularity conditions. Indeed, the consistency of $\widehat{A}_j$ can be established in a similar way. The following theorem summarizes the first-order asymptotic properties of these estimators.

Theorem 1. *Under the Assumption 1 in the Appendix B, for an asset j, we have $(\widehat{C}_j, \widehat{W}_j, \widehat{A}_j) \xrightarrow{p} IV_j$ where IV_j is defined in Equation 4.*

Theorem 1 demonstrates that high-frequency candlestick data yields three different consistent estimators for the integrated variance of asset returns. This result is crucial for the subsequent analysis, ensuring that the estimators converge to the actual integrated variance as the sampling frequency increases. Next, we turn to the asymptotic distribution of these estimators to evaluate their relative efficiency and understand the different information conveyed by these estimators.

Theorem 2. *Suppose that the Assumption 1 and 2 in the Appendix B hold. Then, for an asset j, we have*

$$\Delta_n^{-1/2} \begin{pmatrix} \widehat{C}_j - IV_j \\ \widehat{W}_j - IV_j \\ \widehat{A}_j - IV_j \end{pmatrix} \xrightarrow{\mathcal{L}-s} \mathcal{MN}(0, \Upsilon \cdot IQ_j) \tag{12}$$

where

$$\Upsilon = \begin{pmatrix} 2 & 0.7616 & -0.4534 \\ 0.7616 & 0.4091 & 0.0627 \\ -0.4534 & 0.0627 & 2.2395 \end{pmatrix} \tag{13}$$

and IQ_j is the integrated quarticity of asset j:

$$IQ_j = \int_0^T \nu_{j,s}^4 ds \tag{14}$$

where $\nu_{j,s}$ is defined as spot standard deviation of asset j in Equation 2.

Proof. See Appendix B. □

The matrix Υ is derived from the moments of $\widetilde{a}$ and $\widetilde{w}$, see Equation 8. Also, $\xrightarrow{\mathcal{L}-s}$ denotes stable convergence in law, and $\mathcal{MN}$ indicates the mixed normal distribution. The standard practice in high-frequency econometrics is to employ stable convergence in law rather than ordinary convergence in distribution so that the limiting result remains valid even when conditioning on the entire volatility path $\{\nu_{j,s}\}$, see the discussions in Barndorff-Nielsen et al. (2008). The limiting distribution in such settings is typically a mixed normal because the variance depends on the integrated quarticity derived from the stochastic volatility.

Theorem 2 provides the joint convergence results of $\widehat{C}_j, \widehat{W}_j$, and $\widehat{A}_j$. It reveals that the most efficient integrated variance estimator is realized range followed by realized variance, and the least efficient one is realized asymmetry. Particularly, the realized range estimator is roughly 2.2 times more efficient than the realized volatility estimator[3]. This finding is not surprising since realized range estimator uses the entire price path while realized variance neglects most of it. Surprisingly, the estimation error variance gets bigger for $\widehat{A}_j$ compared to $\widehat{C}_j$. One can argue that it is simply an alternative estimator that happens to be negatively correlated with $\widehat{C}_j$, thereby allowing for the construction of more efficient estimators by utilizing the additional information embedded in candlesticks.

[3]We measure the efficiency in terms of the ratio of the standard deviation of estimation errors.

2.4 Realized candlestick measures: Multivariate case

Extending the idea of realized variance to a multivariate setting, Barndorff-Nielsen and Shephard (2004) proposes the *realized covariance* as a consistent estimator for integrated covariance matrix, ICV:

$$\widehat{\boldsymbol{C}} = \sum_{i=1}^{T/\Delta_n} \boldsymbol{r}_i \boldsymbol{r}_i^\top \tag{15}$$

where $\widehat{\boldsymbol{C}}$ is an $N \times N$ matrix and $\boldsymbol{r}_i$ is the vector of high-frequency returns defined in Equation 6. Similarly, one can define the *realized co-range* and *co-asymmetry* $N \times N$ matrices as multivariate peers of the univariate estimators, realized range, and realized asymmetry:

$$\widehat{\boldsymbol{W}} = \frac{1}{\gamma_{w,2}} \sum_{i=1}^{T/\Delta_n} \boldsymbol{w}_i \boldsymbol{w}_i^\top \tag{16}$$

$$\widehat{\boldsymbol{A}} = \frac{1}{\gamma_{a,2}} \sum_{i=1}^{T/\Delta_n} \boldsymbol{a}_i \boldsymbol{a}_i^\top \tag{17}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{a}_i$ are the vectors of high-frequency range and asymmetry variables defined in Equation 7. The coefficients $\gamma_{w,2}$ and $\gamma_{a,2}$ are defined as in the previous section.

As discussed earlier, univariate realized measures based on candlestick data are shown to be consistent estimators of integrated variance. The following theorem extends this result, establishing a similar relationship in the multivariate case.

Theorem 3. *Under the Assumption B1 in the Appendix B, we have* $(\widehat{\boldsymbol{C}}, \widehat{\boldsymbol{W}}, \widehat{\boldsymbol{A}}) \xrightarrow{p} (\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{A})$ *where the stochastic limits are* $N \times N$ *matrices with the* $(j,k)-th$ *elements are given by:*

$$\begin{bmatrix} \boldsymbol{C}_{jk} \\ \boldsymbol{W}_{jk} \\ \boldsymbol{A}_{jk} \end{bmatrix} \equiv \int_0^T v_{j,s} v_{k,s} \begin{bmatrix} \rho_{jk,s} \\ \xi(\rho_{jk,s}) \\ \psi(\rho_{jk,s}) \end{bmatrix} ds$$

The functions $\xi(\cdot)$ and $\psi(\cdot)$ are defined as:

$$\begin{aligned} \xi(\rho_{jk}) &\equiv \frac{\mathbb{E}[\widetilde{w}_j \widetilde{w}_k]}{\gamma_{w,2}} = \frac{(g(\rho_{jk}) + g(-\rho_{jk}))}{\gamma_{w,2}} \\ \psi(\rho_{jk}) &\equiv \frac{\mathbb{E}[\widetilde{a}_j \widetilde{a}_k]}{\gamma_{a,2}} = \frac{2(g(\rho_{jk}) - g(-\rho_{jk})) - \rho_{jk}}{\gamma_{a,2}} \end{aligned} \tag{18}$$

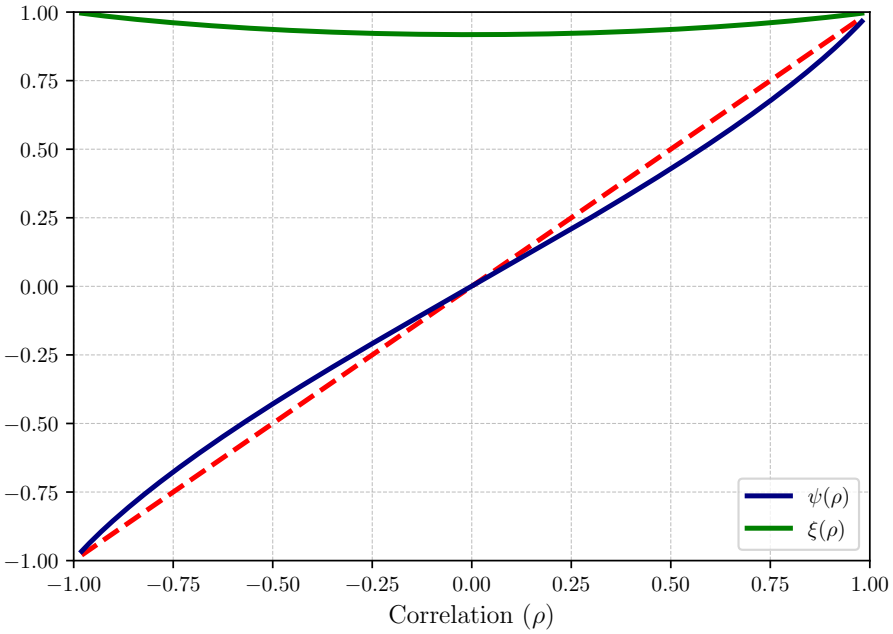
and $g(\cdot)$ function derived in Rogers and Shepp (2006):

$$g(\rho) = \cos\alpha \int_0^\infty \frac{\cosh \nu\alpha}{\sinh \nu\pi/2} \tanh \nu(\alpha/2 + \pi/4) \mathrm{d}\nu$$

with $\rho = \sin\alpha$.

Proof. See Appendix B. □

Figure 2: **The shape of $\psi(\cdot)$ and $\xi(\cdot)$ function, defined in Theorem 3**. The solid red line indicates $\psi(\cdot)$, the green line refers to $\xi(\cdot)$, and the blue dashed line points out a 45-degree line.



It is evident that $\boldsymbol{C}$ is the same as ICV. Hence, Theorem 3 implies that the realized covariance $\widehat{\boldsymbol{C}}$ is a consistent estimator for ICV, the integrated covariance matrix. However, this theorem establishes a nontrivial connection between the stochastic limits

of the realized co-measures $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{A}}$ and ICV. Specifically, these estimators converge to a multiplication of standard deviations with a nonlinear transformation of correlations with functions $\psi(\cdot)$ and $\xi(\cdot)$, potentially diverging from ICV. To explore this relationship in detail, we present the shape of these functions over different correlation values in Figure 2. First of all, this figure shows that these functions take 1 at $\rho = 1$ and therefore recover the consistency result of $N = 1$ case (or $j = k$ case), consistent with Theorem 1. A close examination of the figure suggests that the stochastic limit of $\widehat{\boldsymbol{A}}$ closely approximates the integrated covariance and behaves as a consistent estimator in the case of either no or perfect correlation. Conversely, the shape $\xi(\cdot)$ is symmetric around 0, and so, this implies $\widehat{\boldsymbol{W}}$ does not carry sign information of correlations. These findings motivate the next section, where we construct a new estimator that reasonably combines these realized co-measures.

Next, we investigate the asymptotic distribution of the candlestick-based realized co-measures to complement first-order asymptotic results. In order to simplify the discussion, we focus on the individual elements of the realized co-measures: $\widehat{\boldsymbol{C}}_{jk}$, $\widehat{\boldsymbol{W}}_{jk}$, and $\widehat{\boldsymbol{A}}_{jk}$.

Theorem 4. *Under the Assumptions B1 and B2 in the Appendix B, we have the following asymptotic distributions for $\widehat{\boldsymbol{C}}_{jk}$, $\widehat{\boldsymbol{A}}_{jk}$ and $\widehat{\boldsymbol{W}}_{jk}$ for $j, k = 1, ..., N$:*

$$\begin{aligned} \Delta_n^{-1/2}(\widehat{\boldsymbol{C}}_{jk} - \boldsymbol{C}_{jk}) &\xrightarrow{\mathcal{L}-s} \mathcal{MN}(0, \boldsymbol{\mathcal{C}}) \\ \Delta_n^{-1/2}(\widehat{\boldsymbol{A}}_{jk} - \boldsymbol{A}_{jk}) &\xrightarrow{\mathcal{L}-s} \mathcal{MN}(0, \boldsymbol{\mathcal{A}}) \\ \Delta_n^{-1/2}(\widehat{\boldsymbol{W}}_{jk} - \boldsymbol{W}_{jk}) &\xrightarrow{\mathcal{L}-s} \mathcal{MN}(0, \boldsymbol{\mathcal{W}}) \end{aligned} \tag{19}$$

with

$$\begin{aligned} \boldsymbol{\mathcal{C}} &= \int_0^T \nu_{j,s}^2 \nu_{k,s}^2 (1 + \rho_{jk,s}^2) ds \\ \boldsymbol{\mathcal{W}} &= \int_0^T \nu_{j,s}^2 \nu_{k,s}^2 \Xi(\rho_{jk,s}) ds \\ \boldsymbol{\mathcal{A}} &= \int_0^T \nu_{j,s}^2 \nu_{k,s}^2 \Psi(\rho_{jk,s}) ds \end{aligned} \tag{20}$$

and

$$\begin{aligned} \Psi(\rho_{jk}) &= \frac{\mathbb{E}[\widetilde{a}_j^2\widetilde{a}_k^2]-\mathbb{E}[\widetilde{a}_j\widetilde{a}_k]^2}{\gamma_{a,2}^2} \\ \Xi(\rho_{jk}) &= \frac{\mathbb{E}[\widetilde{w}_j^2\widetilde{w}_k^2]-\mathbb{E}[\widetilde{w}_j\widetilde{w}_k]^2}{\gamma_{w,2}^2}. \end{aligned} \tag{21}$$

Proof. See Appendix B. □

Theorem 4 shows that the asymptotic normality of the estimators is established with a reasonably complicated variance structure. The functions Ψ and Ξ are related to the fourth order moment of multivariate range and asymmetry variables and, to the best of our knowledge, their analytical form has not been derived yet[4]. Nonetheless, we compare the efficiency of these estimators in Section 2.6 in a realistic simulation design.

We examine special cases for correlation parameters to provide intuition about the asymptotic variances. First of all, when $\rho_{jk} = 1$ (or, equivalently, $j = k$ and $N = 1$), we find that $\Psi(\rho_{jk}) = 2.2395$ and $\Xi(\rho_{jk}) = 0.4091$[5]. These results recover asymptotic normality results in the univariate case; see Theorem 2. Moreover, when $\rho_{jk} = 0$, we have $\Psi(\rho_{jk}) = 1$ and this implies that asymptotic variance of $\widehat{\boldsymbol{A}}_{jk}$ is the same as that of $\widehat{\boldsymbol{C}}_{jk}$. Note that Theorem 3 shows the consistency of $\widehat{\boldsymbol{A}}_{jk}$ under zero correlation assumption. As a result, we can argue that $\widehat{\boldsymbol{C}}_{jk}$ and $\widehat{\boldsymbol{A}}_{jk}$ are both consistent estimators for ICV and have the same asymptotic variance under zero correlation assumption. In this case, their asymptotic variance can be written as

$$\int_0^T \nu_{j,s}^2\nu_{k,s}^2 ds. \tag{22}$$

This finding will be utilized in the next section, where we construct an optimal estimator for integrated covariance.

[4]For univariate case, see Appendix of Garman and Klass (1980).

[5]These numbers are computed using moments of simulated Brownian motion paths over $[0, 1]$ as in the previous section.

2.5 Combining realized estimators

Recent work by Li et al. (2024) proposes linear combinations of candlestick-based estimators for spot volatility together with a feasible fixed-k inference procedure. The resulting estimators are proven to be more accurate than standard benchmarks. Inspired by this work, we consider a linear class of estimators, $\widehat{\boldsymbol{\Sigma}}(\theta)$, for integrated covariance, ICV:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{jj}(\theta) &= \mu_1 \cdot \widehat{\boldsymbol{C}}_{jj} + \mu_2 \cdot \widehat{\boldsymbol{A}}_{jj} + (1-\mu_1-\mu_2) \cdot \widehat{\boldsymbol{W}}_{jj} \\ \widehat{\boldsymbol{\Sigma}}_{jk}(\theta) &= \lambda \cdot \widehat{\boldsymbol{C}}_{jk} + (1-\lambda) \cdot \widehat{\boldsymbol{A}}_{jk}; \quad j \neq k \end{aligned} \tag{23}$$

where $\theta = (\mu_1, \mu_2, \lambda)$ be the coefficients or weights which will be determined "optimally". This estimator combines realized candlestick measures in a linear way. It is quite flexible because we allow different weights for on-diagonal and off-diagonal entries of the integrated covariance matrix. Note that this estimator excludes realized co-range $\widehat{W}$ in covariance terms. As shown in Section 2, $\widehat{W}$ does not convey the sign information, and consequently, it may plague the optimal estimator by introducing bias. Nonetheless, $\widehat{W}$ can have valuable information regarding the magnitude of covariances. This potential, however, might be incorporated into an optimal estimator nonlinearly, which requires a more complicated and detailed treatment and, therefore, is left as future research.

We choose the weights by minimizing the expected quadratic loss, represented as follows:

$$\theta^* = (\mu_1^*, \mu_2^*, \lambda^*) \equiv \underset{\mu_1,\mu_2,\lambda}{\operatorname{argmin}} \, \mathbb{E}[\|ICV - \widehat{\boldsymbol{\Sigma}}(\theta)\|^2]. \tag{24}$$

This optimization problem can be divided into two parts: we initially determine the optimal coefficients for the on-diagonal elements, followed by the off-diagonal elements. Based on the results from Theorem 2, the first-order conditions yield the following optimal coefficients:

$$\mu_1^* = -0.3935; \quad \mu_2^* = 1.3917^{6}. \tag{25}$$

[6]These optimal coefficients are reminiscent of the well-known estimator proposed by Garman and Klass

For covariance terms, we derive the optimal solution, λ^*, in the from of the following equation:

$$\lambda^* = 0.5 - \frac{\sum_{j\neq k}\mathbb{E}\left[(\widehat{\boldsymbol{C}}_{jk} - ICV_{jk})^2\right] - \sum_{j\neq k}\mathbb{E}\left[(\widehat{\boldsymbol{A}}_{jk} - ICV_{jk})^2\right]}{2\sum_{j\neq k}\mathbb{E}\left[(\widehat{\boldsymbol{C}}_{jk} - \widehat{\boldsymbol{A}}_{jk})^2\right]}. \tag{26}$$

Intuitively, the optimal weight decreases as the estimation error of the realized covariance increases in relation to the realized asymmetry. When their (squared) estimation errors are equal, the optimal weight is determined to be 0.5. However, the optimal weight depends on unobservable quantities, making its direct application impractical. To address this challenge, we utilize the asymptotic theory established in the previous section, along with the special case of no correlation. Specifically, we first assume that the correlation between all pairs of assets is zero. In this scenario, the expectations in the numerator become the variances of $\widehat{\boldsymbol{C}}_{jk}$ and $\widehat{\boldsymbol{A}}_{jk}$. As a result, we can substitute these expectations with their asymptotic counterparts. This simplification leads to a numerator of 0, since both variances are equivalent under the zero correlation assumption. Consequently, the optimal weight in this particular case is $\lambda^* = 0.5$. We then extend this finding to the general case, which provides us with a feasible estimator characterized by a straightforward and easy-to-implement structure.

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{jj}(\theta) &= -0.3935 \cdot \widehat{\boldsymbol{C}}_{jj} + 1.3917 \cdot \widehat{\boldsymbol{W}}_{jj} + 0.0018 \cdot \widehat{\boldsymbol{A}}_{jj}, \\ \widehat{\boldsymbol{\Sigma}}_{jk}(\theta) &= 0.5 \cdot \widehat{\boldsymbol{C}}_{jk} + 0.5 \cdot \widehat{\boldsymbol{A}}_{jk}; \quad j \neq k. \end{aligned} \tag{27}$$

We would like to point out that the weight on off-diagonal elements is not necessarily the theoretically optimal solution. However, it is near optimal and feasible, achieving a good balance between simplicity and effectiveness. Importantly, our Monte Carlo simulations (see Section 2.6) demonstrate that the 50/50 weighted estimator performs remarkably well in finite samples, closely approximating the performance of the optimal

(1980). Diverging from this seminal paper, we employ realized estimators for the second moments of a semimartingale process.

estimator and leaving minimal scope for further improvement[7].

2.6 Assessing the gap to optimality

As we emphasize in the previous section, our novel estimator $\widehat{\mathbf{\Sigma}}$ corresponds to the "true" optimal combination only under certain circumstances. Therefore, the goal of this section is to determine how closely our estimator approximates the exact optimal version in a reasonably general environment. To this end, we conduct a simulation study following Barndorff-Nielsen et al. (2011). Specifically, we consider a bivariate stochastic volatility model for $i = 1, 2$:

$$dX_t^{(i)} = b^{(i)}dt + \rho^{(i)}\sigma_t^{(i)}dB_t^{(i)} + \sqrt{1-\rho^{(i)2}}\sigma_t^{(i)}dW_t$$

where B^i and W are independent Brownian motions, representing idiosyncratic and common components, respectively. This setting implies that $X^{(i)}$ is a diffusive process with constant drift and random spot volatility. Indeed, we model the spot volatility as follows:

$$\begin{aligned}\sigma_t^{(i)} &= \exp\left(\beta_0^{(i)} + \beta_1^{(i)}\zeta_t^{(i)}\right)\\ d\zeta_t^{(i)} &= \alpha_t^{(i)}\zeta_t^{(i)}dt + dB_t^{(i)}\end{aligned}$$

where the second line obeys the Ornstein-Uhlenbeck process. This data-generating process (DGP) employs a leverage effect, the correlation between $dX_t^{(i)}$ and $\sigma_t^{(i)}$. Moreover, we can deduce that the correlation between assets is $\sqrt{1-\rho^{(1)2}}\sqrt{1-\rho^{(2)2}}$.

[7] In the context of estimating the correlation coefficient of bivariate Brownian Motions, Rogers and Zhou (2008) shows that the "optimal" estimator is simple average of correlation coefficients of asymmetry and return variables, similar to our result. In that paper, the optimality refers to minimizing the variance of the estimation error when the actual correlation is 0 and subject to unbiasedness condition when the actual correlation is either 0 or ± 1. This optimality is limited to specific conditions and does not generalize to more complex dynamics.

The parameters of the DGP are the same as in Barndorff-Nielsen et al. (2011):

$$(b^{(i)}, \beta_0^{(i)}, \beta_1^{(i)}, \alpha^{(i)}) = (0.03, -\frac{5}{16}, \frac{1}{8}, -\frac{1}{40})$$

which ensures that $\mathbb{E}(\int_0^1 \sigma_s^{(i)2} ds) = 1$. The initial values of $\zeta_t^{(i)}$ are drawn from $\mathcal{N}(0, -1/2\alpha^{(i)})$ which is the stationary distribution of the parameter itself. We complete the above setting by assuming $\rho = \rho^{(1)} = \rho^{(2)}$ and assigning multiple values for correlation parameter:

$$\sqrt{1-\rho^{(1)2}}\sqrt{1-\rho^{(2)2}} = 1 - \rho^2 \in \{0, 0.1, 0.2, ..., 0.9\}$$

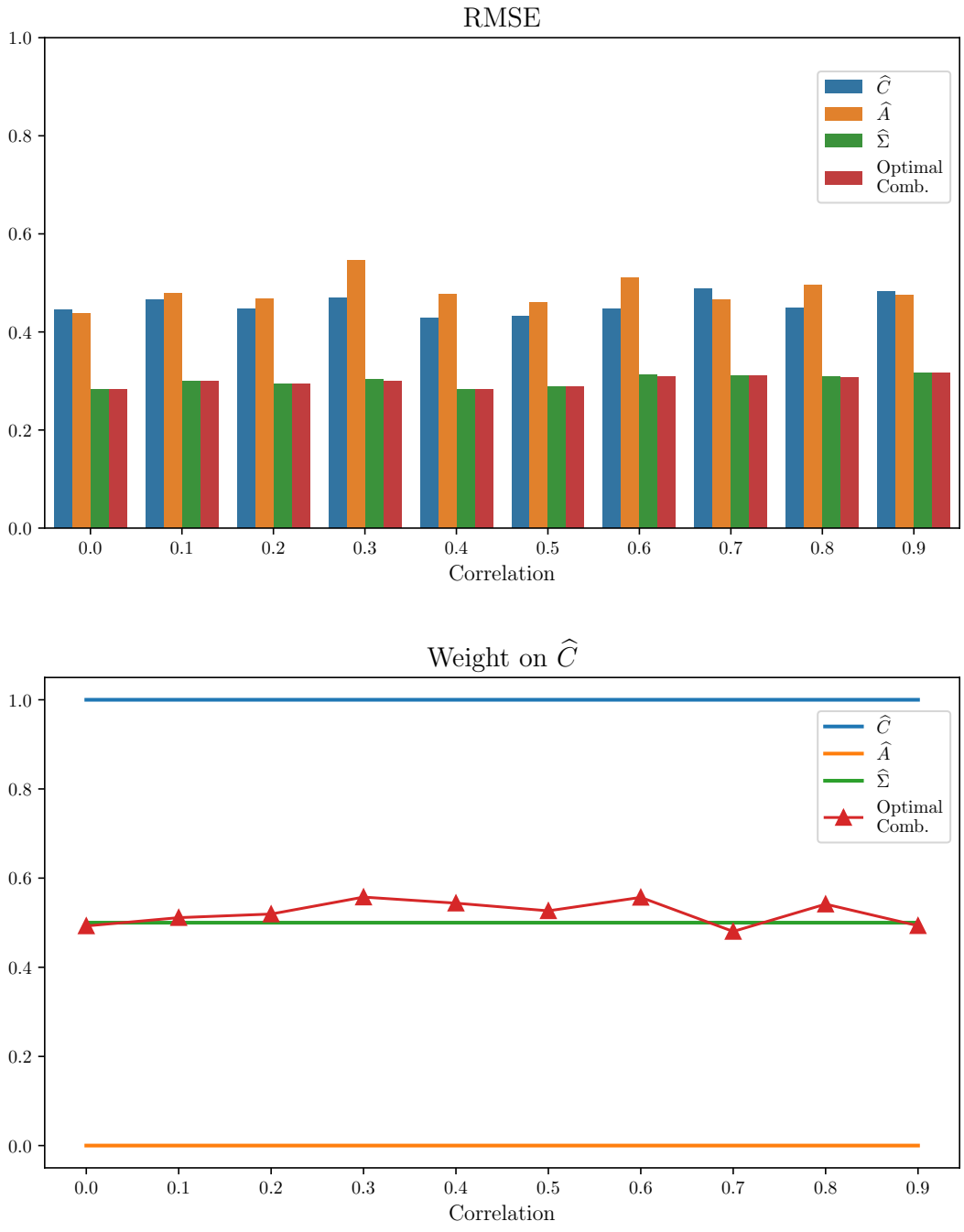
For each value of $\rho^{(i)}$, we simulate 10000 paths over $[0, 1]$. The interval $[0, 1]$ indicates a day or equivalently 6.5 hours. We implement the standard Euler scheme at the seconds level, resulting in 23400 sub-intervals.

We create 5−min high-frequency candlestick observations from seconds-level simulated prices. Then, we compare the new (feasible) estimator $\widehat{\boldsymbol{\Sigma}}$ with realized covariance $\widehat{\boldsymbol{C}}$, realized co-asymmetry $\widehat{\boldsymbol{A}}$ and finally the "true" optimal estimator denote by $\widehat{\boldsymbol{\Sigma}}_{opt}$. The performance of these methods is evaluated based on the root-mean-squared-error (RMSE) metric for the off-diagonal elements covariance matrix.

Figure 3 reports the average outcomes from the Monte Carlo simulations. The top panel illustrates the root mean squared error (RMSE) of the proposed and competing estimators across varying levels of correlation, while the bottom panel displays the optimal weight, λ^*, across different correlation parameters represented by triangular markers, alongside corresponding λ values for other benchmark estimators.

The top panel of Figure 3 demonstrates that $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{A}}$ perform similarly across all correlation levels, with $\widehat{\boldsymbol{C}}$ slightly outperforming $\widehat{\boldsymbol{A}}$ particularly when correlation is around 0.4. The RMSE values of $\widehat{\boldsymbol{\Sigma}}$ are consistently lower than those of $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{A}}$, and thus our estimator is more precise. Importantly, the RMSE values of $\widehat{\Sigma}$ are closely

Figure 3: **The simulation of our estimator vs. the optimal estimator.** The top panel shows the root mean squared error (RMSE) of various estimators. The bottom panel demonstrates the evolution of the optimal coefficient λ^* with red line and triangular markers.



aligned with those of the optimal estimator (optimal combination), demonstrating the robustness of our approach. This similarity suggests that there is minimal scope for improvement toward the actual optimal target once the feasible estimator has been established. Consistent with this finding, the bottom panel of Figure 3 shows that λ^* remains approximately 0.5 across all correlation parameters and converges to 0.5 in the extreme cases of perfect and zero correlation, aligning with theoretical predictions. Consequently, our novel estimator presents a simple yet effective alternative to the optimal estimator, consistently outperforming individual estimators and closely approximating the optimal estimator in all scenarios.

3 Empirical Applications

In this section, we evaluate the performance of the proposed estimator ($\widehat{\mathbf{\Sigma}}$) in different empirical questions. Specifically, we consider the problem of constructing global minimum variance (GMV) portfolios based on the large cross-section of individual stocks. We study this problem in both high-frequency and daily datasets.

3.1 Data

We obtain high-frequency price data for all stocks in the S&P 500 index from the NYSE Trades and Quotes (TAQ) database via Wharton Research Data Services (WRDS). The sample period spans January 1996 to December 2023, covering 336 months and over 6,500 trading days. We exclude incomplete trading days, such as Christmas and Thanksgiving. The high-frequency prices are cleaned according to the procedures outlined in Barndorff-Nielsen et al. (2009), Patton and Sheppard (2015) and Aleti (2022). The data are sampled at a 5-minute frequency from 9:30 AM to 4:00 PM EST, yielding 79 intraday observations per stock per trading day. For each 5-minute interval, we record the last transacted price, along with the highest and lowest transacted prices, designated as close, high, and low prices, respectively. The previous period's close

price serves as the open price observation. We examine all recorded prices in the TAQ Trades database within the specified intervals for high and low prices. Additional details regarding the data construction process can be found in Appendix A.

3.2 Global minimum variance portfolios

We evaluate the proposed estimator and various benchmarks regarding their out-of-sample portfolio performance. To this end, we construct *Global Minimum Variance (GMV)* portfolios using covariance estimates of competing methods. Since the construction of the GMV portfolio does not require estimating expected returns, it offers a clean environment to evaluate the performance of covariance matrix estimators. Additionally, a growing body of literature has highlighted the appealing out-of-sample characteristics of GMV portfolios, demonstrating their effectiveness in minimizing risk and enhancing the Sharpe ratio, for example, Jagannathan and Ma (2003).

Let $\widehat{\boldsymbol{\Sigma}}_t$ be a (integrated) covariance estimate for month t. We use $\widehat{\boldsymbol{\Sigma}}_t$ as the forecast for the next month $t+1$, assuming it follows a random walk. We denote the weight of the GMV portfolio at month $t+1$ by ω_{t+1}, and it is constructed by solving the following optimization problem:

$$\min_{\boldsymbol{\omega}} \boldsymbol{\omega}^\top \widehat{\boldsymbol{\Sigma}}_t \boldsymbol{\omega} \quad \text{subject to} \quad \boldsymbol{\omega}^\top \mathbf{1} = 1. \tag{28}$$

The optimal solution is given by:

$$\widehat{\boldsymbol{\omega}}_{t+1} = \frac{\widehat{\boldsymbol{\Sigma}}_t^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\boldsymbol{\Sigma}}_t^{-1}\mathbf{1}}. \tag{29}$$

We start our empirical analysis by estimating the GMV portfolio implied by our candlestick-based estimator. Then, as a benchmark case, we consider realized covariance $\widehat{\boldsymbol{C}}$, which relies solely on information in standard returns. Lastly, we add an equally weighted portfolio to our analysis. This portfolio does not require an estima-

tion procedure and is called naive diversification. Prior research highlights that equally weighted portfolios often outperform certain estimated investment strategies; see, for example, DeMiguel et al. (2009).

We analyze minimum-variance portfolios consisting of $N \in \{100, 250, 500\}$ stocks. At the start of each month, we pick the N largest stock listed in S&P 500 index and then estimate covariances using methods outlined above with the most recent month data. This allows us to account for local volatility and correlations. Based on integrated covariance matrix estimates, portfolios are formed and held until the beginning of the following month.

First, the competing methods are evaluated based on their out-of-sample portfolio performance. We use three different evaluation metrics: average return (AR), standard deviation (SD), and Sharpe ratio (SR) of estimated portfolios. Specifically, SD indicates the out-of-sample risk of minimum-variance portfolios, which is a preferred metric since minimum-variance portfolios are particularly designed for that purpose. Average return and Sharpe ratio are popular measures among market professionals and academics to evaluate a particular investment strategy.

One potential advantage of incorporating candlestick variables into the estimation procedure is that they yield more stable covariance predictions. This stability can reduce turnover in the portfolio construction process. To evaluate this channel, following Bollerslev et al. (2018b), we report portfolio turnovers, denoted by TO, in addition to the performance metrics outlined above. The portfolio turnover from month t to $t+1$ is calculated as:

$$TO_t = \sum_{n=1}^{N} \left| \omega_{t+1}^{(n)} - \omega_t^{(n)} \frac{1 + r_t^{(n)}}{1 + \omega_t' r_t} \right|. \tag{30}$$

Reducing estimation error in covariances not only yields less turnover but also generates less extreme positions on individual assets. Thus, we also report the portfolio

concentration, denoted by CO:

$$CO_t = \left(\sum_{n=1} \omega_t^{(n)2} \right)^{1/2} \tag{31}$$

and total portfolio short positions, denoted by SP:

$$SP_t = \sum_{n=1}^{N} \omega_t^{(n)} \mathbf{1}_{\{\omega_t^{(n)}<0\}}. \tag{32}$$

Table 1: **Performance evaluation of Global Minimum Variance portfolios based on various covariance matrix estimators**: The columns indicate the minimum variance portfolios using equal weights, $\widehat{C}$ and $\widehat{\Sigma}$ with different number of stocks $N \in \{100, 250, 500\}$. Panel A reports performance metrics pertaining to out-of-sample returns and Panel B analyzes portfolio weights. The bold numbers indicates 5% level significance.

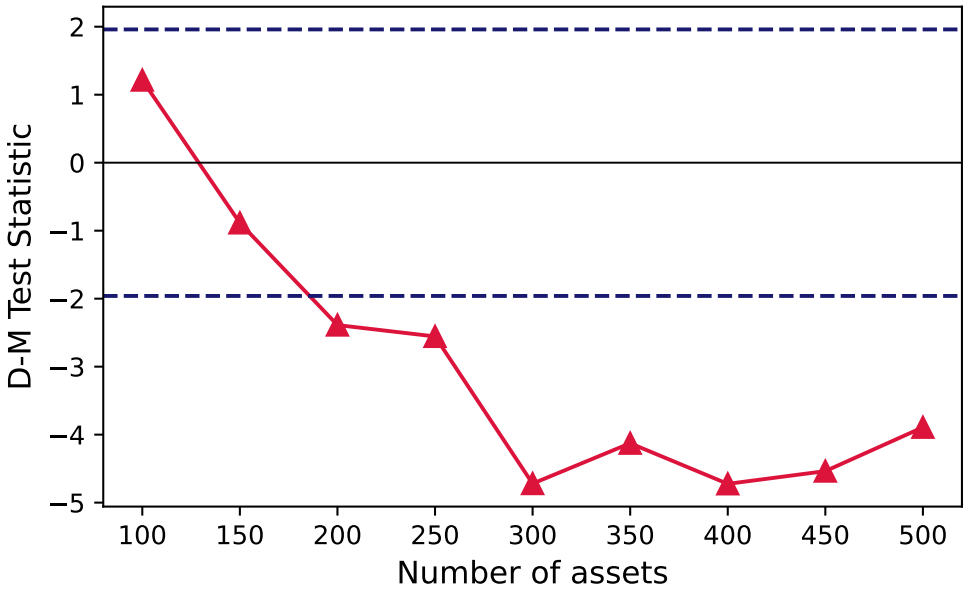
	$N = 100$			$N = 250$			$N = 500$		
	EW	$\widehat{C}$	$\widehat{\Sigma}$	EW	$\widehat{C}$	$\widehat{\Sigma}$	EW	$\widehat{C}$	$\widehat{\Sigma}$
Panel A: Returns:									
AR	17.45	13.47	13.51	17.12	13.75	12.73	9.19	4.46	4.40
SD	18.55	12.06	12.25	17.91	9.89	**9.62**	12.08	5.95	**5.59**
SR	0.94	1.12	1.10	0.96	1.40	1.32	0.76	0.75	0.79
Panel B: Weights:									
TO	0.11	2.81	2.65	0.11	4.17	3.41	0.07	5.14	3.98
CO		0.42	0.38		0.44	0.37		0.47	0.40
SP		-0.88	-0.78		-1.34	-1.08		-1.61	-1.25

Table 1 presents the out-of-sample comparison results for various integrated covariance matrix estimators. Panel A details the portfolio performance measures, while Panel B assesses the estimated weights. In Panel A, the minimum variance portfolio constructed using our novel estimator demonstrates the lowest ex-post variation for $N = 250$ and $N = 500$. Further analysis using Diebold and Mariano (1995) confirms that this difference is statistically significant at the 1% level when tested against $\widehat{C}$. Interestingly, for the case of $N = 100$, our new estimator does not outperform the realized covariance, although this difference is not statistically significant. Additionally, our empirical findings indicate that the optimal estimator and the realized covariance

matrix yield comparable out-of-sample Sharpe ratios, with a slight advantage for the former in the $N = 500$ scenario. Another important observation is that covariance estimators based on realized measures always outperform naively diversified portfolios in terms of portfolio risk and Sharpe ratio.

Turning to Panel B in Table 1, we observe further evidence supporting the above-mentioned results. Portfolio weights estimated using our new method consistently exhibit lower turnovers, more stable weights, and smaller short positions across all portfolio sizes. These findings align with our argument that incorporating candlestick information into covariance estimation enhances the quality of the resulting estimates.

Figure 4: **Global minimum variance portfolio comparison across different numbers of assets.** This figure shows Diebold and Mariano (1995) test statistics, which compares the proposed estimator with realized covariance for different numbers of assets. Blue dashed lines indicate 5% significance levels.



To elaborate on the comparison of different portfolios in terms of ex-post variation, we show the evolution of Diebold and Mariano (1995) test statistics across different number portfolio sizes, ranging from 100 to 500, in Figure 4. We compute the out-

of-sample variation of minimum variance portfolios based on our novel estimator and realized covariance estimator, then test the difference between out-of-sample portfolio risk. Figure 4 shows the corresponding test statistics. Looking at this figure, we can argue that while our estimator performs similarly to the realized covariance estimator for small portfolios, it significantly outperforms the benchmark for medium and large portfolios. Particularly the significant gains step in starting from portfolio size $N = 200$.

3.3 Applying to lower frequency returns

In order to complement the analysis conducted with high-frequency data in the previous section, we apply the proposed candlestick-based estimator in a lower-frequency setting. The literature reports superior performance in high-frequency methods compared to lower-frequency peers in strategic asset allocation; see, for example, Hautsch et al. (2015). However, lower-frequency data, such as daily observations, is quite common among market participants since it is more accessible and consistent with the investment horizon of a typical investor. Therefore, applying our candlestick-based estimation approach highlights its advantages for a broader set of investors and also shows the robustness across datasets with different frequencies.

The same portfolio construction and evaluation framework is employed with the following exceptions: First, we collect the daily price (open, close, high, and low) data from CRSP. We keep the stocks studied in the previous high-frequency section. Instead of using the most recent month's data, we estimate covariances using the past year's daily observations, assuming 250 days constitute a year. To avoid non-invertible covariance matrices, we apply the linear shrinkage method proposed by Ledoit and Wolf (2004) to the sample covariance matrices[8]. It is important to emphasize that the objective of this empirical analysis is not to run a "horse race" among various

[8]To implement this method, we utilize `LedoitWolf()` function provided in `scikit-leran` library in `python` language.

methods—such as candlestick-based estimators and shrinkage estimators—in order to declare a single winner. Instead, the primary goal is to show that, relative to a given benchmark, our novel methodology, which integrates candlestick information, improves the quality of the covariance matrix estimator. This improvement, in turn, leads to lower out-of-sample portfolio risk.

Table 2: **Performance evaluation of Global Minimum Variance portfolios based on various covariance matrix estimators in daily frequency**: The columns indicate the minimum variance portfolios using equal weights, $\widehat{C}$ and $\widehat{\Sigma}$ with different number of stocks $N \in \{100, 250, 500\}$. Panel A reports performance metrics pertaining to out-of-sample returns and Panel B analyzes portfolio weights. The bold numbers indicates 5% level significance

	$N = 100$			$N = 250$			$N = 500$		
	EW	$\widehat{C}$	$\widehat{\Sigma}$	EW	$\widehat{C}$	$\widehat{\Sigma}$	EW	$\widehat{C}$	$\widehat{\Sigma}$
Panel A: Returns:									
AR	9.41	7.49	7.68	10.26	8.29	9.62	12.03	9.32	9.31
SD	19.51	21.66	**20.70**	19.55	21.77	**21.13**	20.99	12.84	**12.63**
SR	0.48	0.35	0.37	0.53	0.38	0.46	0.57	0.73	0.74
Panel B: Weights:									
TO	0.24	3.30	2.54	0.23	5.28	4.03	0.21	2.05	1.74
CO		0.34	0.27		0.31	0.25		0.22	0.20
SP		-0.81	-0.55		-1.44	-1.01		-1.43	-1.24

Table 2 shows the out-of-sample results of minimum variance portfolios constructed using different covariance matrix estimators. Consistent with Table 1, Panel A evaluates the performance of investment strategies, and Panel B reports measures about portfolio weights. Starting from the standard deviations, our results show that candlestick-based optimal estimator $\widehat{\Sigma}$ systematically yields lower portfolio risk relative to benchmark $\widehat{C}$ across all values of N. Further, we test this difference using Diebold and Mariano (1995) and find that all are significantly different from 0 at 5% and 10% level. Similar to high-frequency cases, GMV portfolios based on sample covariance matrix estimator and candlestick estimator have similar Sharpe ratios. Moreover, they outperform naively diversified portfolios for $N = 500$ and underperform for $N = 100, 250$.

Panel B of Table 2 reinforces the findings above and shows that portfolio weights

estimated using our new method consistently show lower turnover, more stable weights, and smaller short positions across all portfolio sizes. These outcomes support our argument that integrating candlestick information into covariance estimation improves the quality of the resulting estimates.

3.4 Cross-sectional pricing implications: Candlestick-SDF

In this section, we investigate the cross-sectional pricing implications of our novel estimator, $\widehat{\boldsymbol{\Sigma}}$, in the context of the stochastic discount factor (SDF). The asset pricing theory establishes that the covariance matrix is a critical input in SDF construction. Specifically, SDF can be calculated directly using the formula:

$$\boldsymbol{M}_{t+1} = 1 - \boldsymbol{\omega}_{t+1}^{*}\mathbf{r}_{t+1} \tag{33}$$

where $\boldsymbol{M}_{t+1}$ is the SDF, $\mathbf{r}_{t+1}$ is the vector of returns at time $t+1$ and $\boldsymbol{\omega}_{t+1}$ is the tangency portfolio weights[9]. Tangency portfolio weights are given by efficient portfolio weights:

$$\boldsymbol{\omega}_{t+1}^{*} = \boldsymbol{\Sigma}_{t}^{-1}\boldsymbol{\mu}_{t} \tag{34}$$

where $\boldsymbol{\mu}_t$ is the vector of expected returns and $\boldsymbol{\Sigma}_t$ is the covariance matrix of returns. As seen in this equation, the covariance matrix plays a direct role in the computation of the SDF, making its accurate estimation crucial for pricing performance.

In our empirical analysis, we replace $\boldsymbol{\Sigma}_t$ with our novel candlestick-based estimator $\widehat{\boldsymbol{\Sigma}}_t$ and realized covariance estimator $\widehat{\boldsymbol{C}}_t$ constructed using month t observations of S&P 500 stocks. While the former yields the $Candlestick - SDF$, the latter results in the $Realized\ Covariance - SDF$. Following prior literature, we estimate expected returns using a rolling average of the past 5- and 10-year monthly returns.

To test the pricing ability of these SDFs, we study two different data sets: JKP

[9]With a slight abuse of notation, the unit of time may be considered as 1 month.

factors by Jensen et al. (2023) and CZ factors by Chen and Zimmermann (2021). These recently developed data sets have gained significant attention in the asset pricing literature. They include a wide range of factors previously studied in the literature, such as size, value, momentum, profitability, and investment. In total, JKP provides 153 factors and CZ has 212 factors[10]. For each test asset, we define the monthly pricing error as $\boldsymbol{M}_{t+1}\mathbf{f}_{t+1}$ where $\mathbf{f}_{t+1}$ is the vector of test asset returns at time $t+1$. We then calculate the cross-sectional R^2 of the pricing errors across different covariance estimators.

Table 3: **Cross-sectional Pricing Performance of Candlestick-SDF**. This table reports the cross-sectional R^2 values of various Stochastic Discount Factors (SDFs) evaluated in data sets constructed by Jensen et al. (2023) and Chen and Zimmermann (2021), denoted as JKP and CZ factors, respectively. The SDF is estimated using covariance matrix estimates of $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{\Sigma}}$ which are called Realized Covariance-SDF and Candlestick-SDF, respectively. Diebold-Mariano test statistics for comparisons against Candlestick-SDF are presented in parentheses. The very first column shows the estimation window used for returns' mean prediction. The last row indicates the number of test assets in each data set.

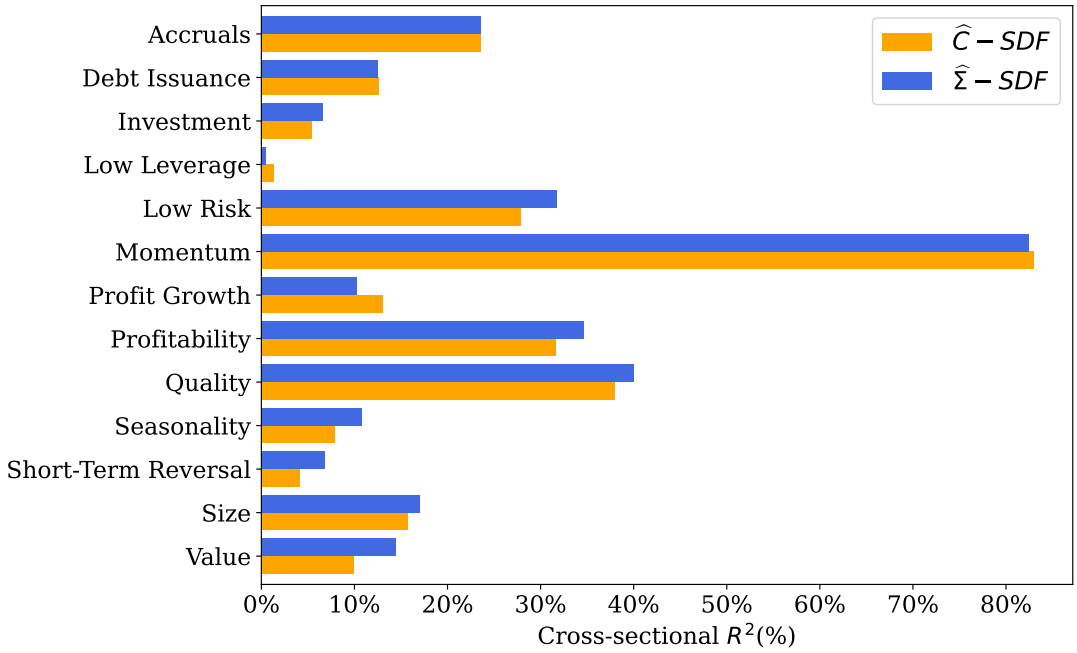
	JKP Factors		**CZ Factors**	
	$\widehat{\boldsymbol{C}}$	$\widehat{\boldsymbol{\Sigma}}$	$\widehat{\boldsymbol{C}}$	$\widehat{\boldsymbol{\Sigma}}$
60 months window	18.03 (5.74)	22.02	20.59 (2.38)	23.76
120 months window	23.27 (5.06)	25.72	18.47 (3.52)	21.06
Number of assets	153		212	

Table 3 reports the cross-sectional R^2 results for both JKP and CZ factors. While the columns show the type of covariance matrix estimates, the rows indicate different estimation methods of expected returns. These results suggest that the cross-sectional R^2 values are significantly higher for the Candlestick-SDF compared to the Realized Covariance-SDF, offering a roughly 3% improvement. This finding is consistent across different data sets and expected return estimates. To further analyze the performance

[10]The JKP factors are obtained from the website www.jkpfactors.com and CZ factors are available in www.openassetpricing.com

of the Candlestick-SDF, we group the JKP factors into different themes such as momentum, quality, investment, and report the cross-sectional R^2 values for each theme in Figure 5[11]. This figure shows that the Candlestick-SDF outperforms the Realized Covariance-SDF across majority factor themes with a few exceptions. Consequently, our findings suggest that the novel estimator, $\widehat{\boldsymbol{\Sigma}}$, enhances the pricing ability of the SDF, potentially leading to more accurate asset pricing models.

Figure 5: **Cross-sectional R^2 across factor themes in JKP data set.** This figure shows the cross-sectional R^2 values for different factor themes in the JKP factors. he factor themes are listed on the y-axis in alphabetical order. While blue bars indicate the results of Candlestick-SDF ($\widehat{\boldsymbol{\Sigma}}$-SDF)), orange bars reperesent SDF constructed using realized covariance ($\widehat{\boldsymbol{C}}$-SDF).



4 Conclusion

This paper introduces a novel framework that leverages high-frequency candlestick data to enhance covariance matrix estimation. The proposed methodology utilizes the entire

[11]This classification is directly obtained from the original paper by Jensen et al. (2023)

price path captured in candlestick observations, including the highest and lowest prices within intervals, to improve the precision of covariance estimates. Particularly, we offer a feasible linear combination of different integrated covariance estimators, which brings significant efficiency gains through a reduction in the variance of estimation errors. Moreover, our simulations show that the feasible estimator that we develop here closely approximates the true optimal linear combination, which is inherently infeasible since it depends on unobservables.

Empirical applications, particularly global minimum variance (GMV) portfolio construction, highlight the practical advantages of the proposed method. Portfolios constructed using candlestick-based estimators consistently exhibit lower risk, reduced turnover, and less extreme asset positions compared to benchmarks, including realized covariance estimators. Notably, the benefits are more pronounced in larger and more diversified portfolios. Furthermore, the robustness of the approach is validated through applications in daily CRSP price data, showing its adaptability to different datasets and investment horizons.

Finally, we explore the cross-sectional pricing implications of our novel estimator in the context of the stochastic discount factor (SDF). By mitigating the estimation error in the covariances, our approach might have potential to improve the pricing ability of corresponding SDF. Consequently, we find that the Candlestick-SDF yields smaller cross-sectional pricing errors relative to the Realized Covariance-SDF. This finding underscores the broader implications of our approach in asset pricing as well as portfolio management.

References

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

Aleti, S. (2022). The high-frequency factor zoo. *Available at SSRN 4021620*.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *Review of Economics and Statistics*, 89(4):701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96(453):42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Bandi, F. M., Russell, J. R., and Zhu, Y. (2008). Using high-frequency data in dynamic portfolio choice. *Econometric Reviews*, 27(1-3):163–198.

Bannouh, K., Van Dijk, D., and Martens, M. (2009). Range-based covariance estimation using high-frequency data: The realized co-range. *Journal of Financial Econometrics*, 7(4):341–372.

Barndorff-Nielsen, O., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *The Econometrics Journal*, 12(3):C1–C32.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and 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, O. E., Hansen, P. R., Lunde, A., and 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, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B*, 64(2):253–280.

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

Bibinger, M., Hautsch, N., Malec, P., and Reiß, M. (2014). Estimating the quadratic covariation matrix from noisy observations: Local method of moments and efficiency. *The Annals of Statistics*, 42(4):1312–1346.

Bollerslev, T. (2022). Realized semi (co) variation: Signs that all volatilities are not created equal. *Journal of Financial Econometrics*, 20(2):219–252.

Bollerslev, T., Hood, B., Huss, J., and Pedersen, L. H. (2018a). Risk everywhere: Modeling and managing volatility. *Review of Financial Studies*, 31(7):2729–2773.

Bollerslev, T., Li, J., and Li, Q. (2024a). Optimal nonparametric range-based volatility estimation. *Journal of Econometrics*, 238(1):105548.

Bollerslev, T., Li, J., Li, Q., and Li, Y. (2024b). Optimal candlestick-based spot volatility estimation: New tricks and feasible inference procedures. *Available at SSRN 5046917*.

Bollerslev, T., Li, J., Li, Y., and Zhang, Q. (2024c). Illuminating important economic news by candlesticks: Optimal testing meets technical analysis. *Working Paper*.

Bollerslev, T., Li, S. Z., and Todorov, V. (2016). Roughing up beta: Continuous versus discontinuous betas and the cross section of expected stock returns. *Journal of Financial Economics*, 120(3):464–490.

Bollerslev, T., Meddahi, N., and Nyawa, S. (2019). High-dimensional multivariate realized volatility estimation. *Journal of Econometrics*, 212(1):116–136.

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2018b). Modeling and forecasting (un) reliable realized covariances for more reliable financial decisions. *Journal of Econometrics*, 207(1):71–91.

Brandt, M. W. and Diebold, F. X. (2006). A no-arbitrage approach to range-based estimation of return covariances and correlations. *Journal of Business*, 79(1):61–74.

Cai, T. T., Hu, J., Li, Y., and Zheng, X. (2020). High-dimensional minimum variance portfolio estimation based on high-frequency data. *Journal of Econometrics*, 214(2):482–494.

Chen, A. Y. and Zimmermann, T. (2021). Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*.

Christensen, K., Kinnebrock, S., and Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 159(1):116–133.

Christensen, K. and Podolskij, M. (2007). Realized range-based estimation of integrated variance. *Journal of Econometrics*, 141(2):323–349.

Christensen, K. and Podolskij, M. (2012). Asymptotic theory of range-based multipower variation. *Journal of Financial Econometrics*, 10(3):417–456.

Christensen, K., Podolskij, M., and Vetter, M. (2009). Bias-correcting the realized range-based variance in the presence of market microstructure noise. *Finance and Stochastics*, 13:239–268.

Dai, C., Lu, K., and Xiu, D. (2019). Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data. *Journal of Econometrics*, 208(1):43–79.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial studies*, 22(5):1915–1953.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

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

Fleming, J., Kirby, C., and Ostdiek, B. (2003). The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, 67(3):473–509.

Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, pages 67–78.

Gospodinov, N., Kan, R., and Robotti, C. (2014). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *The Review of Financial Studies*, 27(7):2139–2170.

Hautsch, N., Kyj, L. M., and Malec, P. (2015). Do high-frequency data improve high-dimensional portfolio allocations? *Journal of Applied Econometrics*, 30(2):263–290.

Jacod, J. and Protter, P. (2012). *Discretization of Processes*. Springer.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58(4):1651–1683.

Jagannathan, R. and Wang, Z. (1996). The conditional capm and the cross-section of expected returns. *The Journal of finance*, 51(1):3–53.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Kan, R. and Zhang, C. (1999). Two-pass tests of asset pricing models with useless factors. *the Journal of Finance*, 54(1):203–235.

Kelly, B. T., Malamud, S., Pourmohammadi, M., and Trojani, F. (2024). Universal portfolio shrinkage. Technical report, National Bureau of Economic Research.

Kim, D. and Oh, M. (2024). Dynamic realized minimum variance portfolio models. *Journal of Business & Economic Statistics*, pages 1–12.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of Econometrics*, 149(2):149–173.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Lee, S. S. and Mykland, P. A. (2008). Jumps in financial markets: A new nonparametric test and jump dynamics. *Review of Financial Studies*, 21(6):2535–2563.

Li, J. (2015). Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business & Economic Statistics*, 33(3):381–392.

Li, J., Wang, D., and Zhang, Q. (2024). Reading the candlesticks: An ok estimator for volatility. *Review of Economics and Statistics*, 106(4):1114–1128.

Li, Y., Nolte, I., Yu, S., et al. (2023). Nonparametric range-based estimation of integrated variance with episodic extreme return persistence. *Sandra and Yu, Shifan, Nonparametric Range-Based Estimation of Integrated Variance with Episodic Extreme Return Persistence (July 12, 2023)*.

Liu, Q. (2009). On portfolio optimization: How and when do we benefit from high-frequency data? *Journal of Applied Econometrics*, 24(4):560–582.

Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36(2):270–296.

Martens, M. and Van Dijk, D. (2007). Measuring volatility with the realized range. *Journal of Econometrics*, 138(1):181–207.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, pages 61–65.

Patton, A. J. and Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3):683–697.

Pooter, M. d., Martens, M., and Dijk, D. v. (2008). Predicting the daily covariance matrix for s&p 100 stocks using intraday data—but which frequency to use? *Econometric Reviews*, 27(1-3):199–229.

Rogers, L. and Shepp, L. (2006). The correlation of the maxima of correlated brownian motions. *Journal of Applied Probability*, 43(3):880–883.

Rogers, L. C. and Zhou, F. (2008). Estimating correlation from high, low, opening and closing prices. *Annals of Applied Probability*, 18(2):813–823.

Shanken, J. (1992). On the estimation of beta-pricing models. *The review of financial studies*, 5(1):1–33.

Shephard, N. and 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.

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

Appendix A Data Collection

Our dataset consists of the stocks listed in S&P 500 between January 1996 and December 2023. We closely follow the high-frequency data collection steps outlined in Aleti (2022), which mostly draws on Barndorff-Nielsen et al. (2009) and Patton and Sheppard (2015). Specifically, we obtain historical constituents of S&P 500 index from WRDS. The data construction step takes three main steps: first, we get high-frequency prices from NYSE TAQ; second, daily open and close prices and adjusted returns are downloaded from CRSP; and finally, we merge these two datasets.

To begin with, NYSE TAQ has two different datasets "TAQ Monthly" and "TAQ Daily". The TAQ Monthly dataset was utilized for data prior to 2015, while the TAQ Daily dataset was used for data starting from 2015. We extensively use the WRDS cloud for collection and cleaning prices. In particular, for each stock, I consider trades with the following exclusion/inclusion rules:

- Include only trades occurring between 09:30:00 and 16:00:00
- Exclude trades with non-positive prices or sizes
- Retain trades from the exchange with the highest volume for the stock
- Include trades with condition codes E, F, or blank
- Restrict trades to those with transaction prices within the CRSP daily bid-low and ask-high range
- Exclude trades with five or more immediate reversals within a 50-sample moving window
- Keep the trade with the median price when multiple trades share the same timestamp
- Retain only trades with correction numbers 0 or 1

- Exclude trades from exchange "D."

Using cleaned trade data, we construct 5-minute prices for times 09:30:00, 09:35:00, 09:40:00, ..., 16:00:00. For each time interval (like 09:35:00−09:40:00), we obtain the highest, the lowest, and the last transacted prices and match them to the right end point of time. Additionally, the first trade in 09:30:00-09:35:00 is assigned to 09:30:00. This makes, in total, 79 observations per day and per stock.

The second step involves extracting daily open and close prices from the CRSP database to properly handle overnight returns, accounting for stock splits, dividends, and delistings. The CRSP database provides key data, including open prices, close prices, adjusted close-to-close returns, delisting returns, and shares outstanding.

- First, the close-to-open return (or overnight return) is calculated using the open price, close price, and close-to-close return. This overnight return is assigned to 09:30:00.

- Additionally, the 16:00:00 return is adjusted to include delisting returns.

The final step is to merge CRSP open-close prices with TAQ high-frequency intraday prices using `permno` links, found in the TAQ-CRSP link table in WRDS. Intuitively, our dataset can be considered as CRSP daily prices filled by TAQ intraday prices.

Appendix B Proofs

The detailed proofs for the theoretical results presented in this paper are currently under development. These will be finalized and included in the full version of the paper before the conference. We appreciate your understanding.

Assumption B1. *The log-price vector X_t as an Ito semi-martingale process defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$:*

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{b}_s ds + \int_0^t \boldsymbol{\sigma}_s d\boldsymbol{W}_s, \quad 0 \leq t \leq T, \tag{B.1}$$

where $\boldsymbol{b} \in \mathbb{R}^N$ is locally bounded drift process, $\boldsymbol{\sigma}$, the stochastic volatility, is cádlág and $\boldsymbol{W}$ is a $N-$ dimensional standard Brownian motion.

Assumption B2. *The stochastic volatility process $\boldsymbol{\sigma}$ follows the following dynamics:*

$$\boldsymbol{\sigma}_t = \boldsymbol{\sigma}_0 + \int_0^t \tilde{\boldsymbol{b}}_s ds + \int_0^t \tilde{\boldsymbol{\sigma}}_s d\boldsymbol{W}_s + \widetilde{M}_t \tag{B.2}$$

where $\tilde{\boldsymbol{b}}$ is a locally bounded drift process, $\tilde{\boldsymbol{\sigma}}$ is a $d \times d \times d$ tensor-valued cádlág process, and $\widetilde{M}$ is a local martingale orthogonal to Brownian motion $\boldsymbol{W}_t$.