Bayesian Nonparametric Covariance Estimation with Noisy and Nonsynchronous Asset Prices

Abstract

This paper proposes a Bayesian nonparametric approach to estimate the ex-post covariance matrix of asset returns from high-frequency data in the presence of market microstructure noise and nonsynchronous trading. Several contributions are made. First, pooling is used to group returns with similar covariance matrices to improve estimation accuracy. Second, an improved method for synchronizing the observations based on data augmentation is introduced. Third, the covariance matrix estimator is guaranteed to be positive definite. Finally, the proposed approach delivers exact finite sample inference without relying on asymptotic assumptions. All of these benefits lead to a more accurate estimator, as confirmed by Monte Carlo simulation results. In real data applications, the proposed covariance estimator results in improved portfolio choice outcomes.

Key words: realized covariance, high-frequency data, Bayesian nonparametrics, nonsynchronous trading, data augmentation

JEL: C11, C58, C80
1 Introduction

The covariance matrix of asset returns is the key input for many financial problems, such as portfolio allocation and risk management. Estimation of the covariance matrix using intraperiod returns has become a very active area of research since high-frequency data became available. An important first step is the extension of realized variance to realized covariance (Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004)). Under the assumption that observations are free of measurement error and nonsynchronous trading, the realized covariance is a consistent estimator of the integrated covariance. However, in reality, asset prices are contaminated with market microstructure noise and arrive nonsynchronously, which leads to poor statistical performance of the realized covariance estimator.

Several approaches have been adopted for the covariance estimation of noisy and nonsynchronously spaced prices. Adjusting realized covariance using lead and lag autocovariance terms reduces the bias caused by nonsynchronous trading (Scholes & Williams (1977) and Dimson (1979)). Barndorff-Nielsen et al. (2011) introduced the multivariate realized kernel based on refresh time synchronization. Zhang (2011) discussed the optimal sampling frequency in constructing realized covariance and proposed the two-scale estimator. Christensen et al. (2010) applied the pre-averaging method to covariance estimation. A¨ıt-Sahalia et al. (2010) introduced the quasi-maximum likelihood covariance estimator and the generalized synchronization method. Lunde et al. (2016) designed the composite realized kernel for estimating vast covariance matrices. Hayashi & Yoshida (2005) exemplify another branch of the literature. The cumulative covariance estimator they developed can be applied directly to raw observations without synchronization. Voev & Lunde (2007) proposed a bias correction to make the cumulative covariance estimator suitable for noisy prices. Griffin & Oomen (2011) evaluated the statistical properties of realized covariance, realized covariance with lead-lag adjustments and the cumulative covariance estimator. Another way to construct a covariance matrix estimator using high-frequency data relies on parametric models. Hansen et al. (2008) analyzed ex-post variance and covariance estimators based on moving average
models. Peluso et al. (2014) were the first to link Bayesian approaches to covariance estimation. Corsi et al. (2015) proposed a covariance estimator based on a Kalman smoother. Other covariance estimation approaches include Malliavin & Mancino (2002), Renó (2003), Bandi & Russell (2005) and Large (2007).

Instead of treating intraperiod data independently, the proposed approach exploits pooling among return vectors with similar covariance matrices. The Bayesian nonparametric variance estimation method proposed in Griffin et al. (2016) is extended to the multivariate version to allow pooling in covariance estimation. Mykland & Zhang (2009) considered the univariate case of holding the return variance constant over blocks of consecutive returns to improve estimation efficiency. The pooling method adopted in this paper is more general than the blocking approach in Mykland & Zhang (2009). Under the Bayesian nonparametric framework, observations are not required to be consecutive in time and the number of clusters can be determined endogenously. To adjust for bias caused by microstructure noise and nonsynchronous trading, the model incorporates a vector moving average parameterization for high-frequency data. By adopting the adjustment for moving average-based estimator in Hansen et al. (2008), a covariance estimator that corrects for market microstructure noise and nonsynchronous trading is derived from the Bayesian nonparametric model.

In related work, Peluso et al. (2014) used data augmentation based on a dynamic linear model to synchronize observations. The synchronization method proposed in this paper also uses data augmentation but exploits pooling to increase estimation accuracy. Built upon the previous-tick method defined in Hansen & Lunde (2006), the proposed method eliminates the zero returns caused by nonsynchronous trading. The non-updated prices on grid points are treated as missing observations and estimated conditional on observed data and model structure. Conditional on sampled observations without zero returns, the Bayesian nonparametric covariance estimator with moving average adjustment is robust to the nonsynchronous bias.

Another desirable feature is that the proposed covariance matrix estimator is guaranteed
to be positive definite, while almost all model-free high-frequency covariance estimators are positive semidefinite at most. Using an inverse Wishart distribution as the prior, the intraperiod covariance matrix sampled from its posterior is always positive definite. In addition, data augmentation removes the zero returns caused by missing observations, which further eliminates the chance of obtaining singular matrices.

Monte Carlo simulation is conducted to compare the Bayesian nonparametric covariance estimator with the realized covariance and multivariate realized kernel in circumstances with different data frequencies, dimensions and microstructure noise levels. The proposed estimator yields lower root-mean-square-errors when estimating the covariance matrix, especially the off-diagonal elements, than the two classical estimators. Empirical applications to equity data show that the Bayesian nonparametric covariance estimator captures similar time series dynamics of correlation and realized beta as the multivariate realized kernel. Using a volatility-timing strategy, the minimum variance portfolio based on the proposed covariance estimator outperforms those based on realized covariance or multivariate realized kernel in terms of Sharpe ratio and investor’s utility level.

This paper is organized as follows. Section 2 defines the estimation target and briefly reviews two benchmark estimators. In Section 3, the Bayesian nonparametric model, the proposed covariance estimator and the synchronization method with data augmentation are discussed. Section 4 conducts data simulation and compares the proposed estimator with competing alternatives. Empirical applications are presented in Section 5. Section 6 concludes followed by an appendix.

2 Ex-post Covariance and Benchmarks

Suppose the log prices of $d$ assets are generated from

$$dp(t) = m(t)dt + \Pi(t)dw(t),$$  

(1)
where \( \mathbf{m}(t) \) is a vector of drift terms, \( \mathbf{\Pi}(t) \) is the instantaneous volatility matrix, and \( \mathbf{w}(t) \) is a vector of standard Brownian motions. As the true measure of the ex-post covariance, the integrated covariance is the quantity of interest and is defined as

\[
\mathbf{V} = \int_0^1 \mathbf{\Pi}(\tau)\mathbf{\Pi}(\tau)'d\tau.
\] (2)

The realized covariance (RC) formalized in Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004) is defined as

\[
\mathbf{RC} = \sum_{i=1}^n \mathbf{r}_i\mathbf{r}_i'.
\] (3)

where \( \mathbf{r}_i = \mathbf{p}_i - \mathbf{p}_{i-1} \), \( \mathbf{p}_i = (p_{i}^{(1)}, \ldots, p_{i}^{(d)}) \) denotes the \( i \)th intraperiod log price vector of \( d \) assets and \( n \) is the number of regularly spaced returns. Without the presence of microstructure noise and nonsynchronous trading, RC converges to the integrated covariance as \( n \to \infty \). However, in finite samples, RC is a noisy estimator of the ex-post covariance matrix, and the finite sample distribution of RC can only be approximated from the asymptotic result derived in Barndorff-Nielsen & Shephard (2004).

In reality, due to the bid-ask bounce, the discreteness of price changes and measurement error, the observed prices are inevitably contaminated with market microstructure noise, which makes return series autocorrelated. Moreover, the prices of different assets are not updated simultaneously. As a result, the sampled returns have a lead-lag dependence, which leads to underestimation of the covariances, especially when the sampling frequency is high. This phenomenon was documented as the “Epps effect” by Epps (1979). These two challenges restrict the use of data with high frequency to form RC estimates.

A commonly used approach to mitigate the influences of noise and nonsynchronous trading is to adjust the RC estimator using autocovariance terms. Barndorff-Nielsen et al. (2011) integrated the lead-lag autocovariance adjustment, kernel-based weight function and refresh-
time sampling to propose the multivariate realized kernel (RK), which is defined as

$$\text{RK} = \sum_{j=-h}^{h} \left( k\left(\frac{j}{h}\right) \sum_{i=j+1}^{\hat{n}} \hat{r}_i \hat{r}_i^t \right),$$

(4)

where $k(\cdot)$ is the Parzen kernel function\(^1\), and the bandwidth $h$ is determined as $h = c_0 \hat{n}^{3/5} d^{-1} \sum_{i=1}^{d} \xi_i^{4/5}$ . $\xi_i$ is the noise-to-signal ratio, which can be estimated following Section 3.4 in Barndorff-Nielsen et al. (2011). $\hat{r}_i$ is based on prices synchronized using the refresh time scheme, under which prices are sampled when all asset prices are updated and return series are irregularly spaced.

### 3 Bayesian Nonparametric Covariance Estimation

This section introduces a Bayesian nonparametric model of multivariate high-frequency returns and an improved synchronization method with data augmentation. A covariance matrix estimator that is robust to microstructure noise and nonsynchronous trading is derived from the proposed model.

#### 3.1 Model of High-Frequency Data

Pooling intraperiod observations with common covariation pattern could improve the estimation accuracy of the ex-post covariance matrix. However, the optimal number of clusters is difficult to determine. The modeling high-frequency data using parametric models may suffer from misspecification. By contrast, Bayesian nonparametric models provide a flexible framework to exploit pooling. The univariate Dirichlet process mixture (DPM) model used in Griffin et al. (2016) is extended to its multivariate version to allow pooling in the

\(^1\)Parzen kernel function:

$$k(x) = \begin{cases} 
1 - 6x^2 + 6x^3, & 0 \leq x \leq 1/2 \\
2(1 - x)^3, & 1/2 < x \leq 1 \\
0, & x > 1 
\end{cases}$$

For the Parzen kernel function, $c_0 = 3.5143$. 

6
covariance estimation. As a nonparametric version of the mixture model, DPM allows the number of clusters to be nonfixed and inferred based on data. To address autocorrelation and lead-lag dependence in returns, a first-order vector moving average parameterization is incorporated into DPM. The DPM model with vector moving average (DPM-VMA) for extracting covariation information from return vectors \( r_1, \ldots, r_n \) is given as

\[
\begin{align*}
  r_i &= \mu + \Theta \eta_{i-1} + \eta_i, \quad \eta_i \sim N(0, \Sigma_i), \quad i = 1, \ldots, n, \\
  \Sigma_i | G &\overset{iid}{\sim} G, \\
  G | G_0, \alpha &\sim \text{DP}(\alpha, G_0),
\end{align*}
\]

where \( \eta_i = r_i - \mu - \Theta \eta_{i-1} \) is the vector of error terms and \( \Theta \) is the moving average coefficient matrix. \( \Sigma_i \) is a state-dependent intraperiod covariance matrix, and data within a cluster share the same covariance. The distribution of \( \Sigma_i \), denoted \( G \), is a discrete distribution with a varying number of groups. Such a flexible structure is achieved by setting the Dirichlet process \( \text{DP}(\alpha, G_0) \) as the prior of \( G \). A draw from the Dirichlet process is a discrete distribution centered around the base measure \( G_0 \). \( G_0 \) is set to be an inverse Wishart distribution denoted as \( \text{IW}(\Psi, \nu) \), which is the conjugate prior and guarantees the positive definiteness of \( \Sigma_i \). The precision parameter \( \alpha \) influences the degree of pooling. As \( \alpha \) increases, the pooling process is more likely to have more clusters, and the effect of pooling diminishes.

The base measure \( \text{IW}(\Psi, \nu) \) must be calibrated based on the volatility and covariation pattern within the target period. I assume that the base function is centered on the true covariance matrix. The scale matrix \( \Psi \) and the degree of freedom \( \nu \) are set to\(^2\)

\[E(X) = \frac{\Psi}{\nu - d - 1},\]

and variance of diagonal elements of \( X \) are

\[
\text{var}(X^{(j)}) = \frac{2(\Psi^{(jj)})^2}{(\nu - d - 1)^2(\nu - d - 3)}.
\]

\(^2\)For \( X \sim \text{IW}(\Psi, \nu) \), the mean is
\[ \Psi = \frac{\nu - d - 1}{n} \text{RC}, \]  
\[ \nu = \frac{1}{d} \sum_{j=1}^{d} \frac{2(\text{RC}^{(jj)})^2}{\text{var}(r_i^{(j)})^2} n + d + 3, \]  

where the RC based on low-frequency data approximates the true covariance matrix. To add flexibility, the precision parameter \( \alpha \) is treated as unknown with a hierarchical prior Gamma(\( a, b \)). A shrinkage prior N(0, \( \Lambda \)), where \( \Lambda \) is a diagonal matrix with small variance values, is used as the prior of \( \mu \).

### 3.2 Synchronization with Data Augmentation

Synchronization is the process of placing observed prices on grid points. The commonly used previous-tick method yields equally spaced data: given grid length \( h \), the prices are sampled as

\[ p_i^{(j)} = p_{\max(\tau_j | \tau_j \leq ih)}, \quad j = 1, \ldots, d. \]

Although the previous tick method synchronizes observations, it causes two problems. First, the durations of returns are mismatched. As shown in Figure 1, the \( i^{th} \) asset 1 return \( r_i^{(1)} \) and asset 2 return \( r_i^{(2)} \) do not cover the same time interval, and dependence exists between adjacent returns, such as \( r_i^{(1)} \) and \( r_{i+1}^{(2)} \). Second, the sampled returns contain zeros caused by the absence of transactions in one or more interval(s). Zero returns not only influence the estimation of the model parameters but also make the data have lead-lag dependence with more than one period. To correct the bias caused by zero returns, high-frequency covariance estimators must be adjusted using more lead-lag terms at the cost of higher estimation
This paper proposes a synchronization method with data augmentation to improve the previous-tick scheme. In the example shown in Figure 2, no observation exists in the interval \((i + 2, i + 3]\) for both asset 1 and 2 and interval \((i, i + 1]\) for asset 3. In other words, \(p_{i+3}^{(1)}\), \(p_{i+3}^{(2)}\) and \(p_{i+1}^{(3)}\) can be seen as missing. Under the Bayesian framework, missing observations on common grid points can be treated as unknown and estimated along with other model parameters. See Appendix 7.2 for detailed data augmentation steps. In contrast to the previous-tick method, the proposed approach eliminates the zero returns by filling in the missing observation gaps. As a result, the synchronized return series are correlated only with adjacent returns, which can be filtered out through the first order vector moving average parameterization illustrated in Section 3.1.

As in the previous-tick method, the proposed synchronization method requires a common time grid. There is a tradeoff between the length of the grid and the quality of data augmentation. The larger the grid length is, the fewer the missing prices that need to be augmented. However, this leads to more information loss because of the low sampling frequency. Increasing the grid frequency results in a diminished data augmentation value because more grid points contain missing observations, which requires inference. Simulation results show that the root-mean-square-error (RMSE) of the proposed estimator exhibits a U-shaped pattern as the grid length increases. It is recommended that the grid length be set to four times the average price change durations of all assets.

3.3 Model Estimation

The DPM-VMA model is estimated using Markov chain Monte Carlo (MCMC) techniques. Expressing the DP prior as the stick-breaking representation by Sethuraman (1994), the
model is written in the following form of a mixture model with infinite states.

\[
p(\mathbf{r} | \boldsymbol{\mu}, \{\Phi_j\}_{j=1}^{\infty}, \{w_j\}_{j=1}^{\infty}) = \sum_{j=1}^{\infty} w_j N(\mathbf{r} | \boldsymbol{\mu} + \Theta \eta_{i-1}, \Phi_j), \tag{10}
\]

\[
w_1 = v_1, \quad w_j = v_j \prod_{l=1}^{j-1} (1 - w_l), \quad v_j \overset{iid}{\sim} \text{Beta}(1, \alpha), \tag{11}
\]

where \(w_j\) is the weight associated with the \(j^{th}\) component and \(\Phi_j\) denotes the unique covariance matrix in cluster \(j\).

The infinite state space challenges the estimation of model parameters. The slice sampler of Kalli et al. (2011) is applied to the stick-breaking representation of the DPM-VMA model. By introducing a set of auxiliary variables \(u_{1:n} = \{u_1, \ldots, u_n\}\), the infinite state space can be truncated to a finite space. Conditional on \(u_{1:n}\), the model is a finite mixture model, which makes the estimation of model parameters feasible.

A set of latent state variables \(s_{1:n} = \{s_1, \ldots, s_n\}\), where \(s_i \in \{1, 2, \ldots K\}\), is introduced to label each observation’s cluster. Given \(s_i = j\), \(\Sigma_i = \Phi_j\) and all the return vectors in cluster \(j\) share \(\Phi_j\). Note that the number of clusters \(K\) is adjusted over the MCMC iterations. A new cluster with a covariance matrix \(\Phi_{K+1} \sim \text{IW}(\Psi, \nu)\) can be created and redundant clusters can be merged.

Combining the priors and data information, the joint posterior of the model parameters and \(u_{1:n}\) is

\[
\pi(\boldsymbol{\mu})\pi(\Theta) \prod_{j=1}^{K} \pi(\Psi_j)\pi(\alpha) \prod_{i=1}^{n} \mathbb{1}(u_i < w_{s_i}) N(\mathbf{r}_i | \boldsymbol{\mu} + \Theta \eta_{i-1}, \Phi_{s_i}). \tag{12}
\]

The posterior sampling contains the following steps.

1. Sample \(\boldsymbol{\mu}|\mathbf{r}_{1:n}, \Phi_{1:K}, \Theta, s_{1:n}\).
2. Sample \(\Theta|\mathbf{r}_{1:n}, \boldsymbol{\mu}, \Phi_{1:K}, s_{1:n}\).
3. Sample \(\Phi_j|\mathbf{r}_{1:n}, \boldsymbol{\mu}, \Theta, s_{1:n}\) for \(j = 1, \ldots, K\).
4. Sample $v_j|s_{1:n}$ for $j = 1, \ldots, K$.

5. Sample $u_i|w_i, s_{1:n}$ for $i = 1, \ldots, n$.

6. Update $K$ based on $u_{1:n}$ and $w_{1:n}$.

7. Sample $s_i|r_{1:n}, \mu, \Theta, \Phi_{1:K}, s_{-i}, u_{1:n}$ for $i = 1, \ldots, n$.

8. Sample $\alpha|K$.

Sampling high-dimensional parameters $\mu$ and $\Theta$ using Metropolis-Hasting results in low mixing and good proposals are difficult to find. The Hamiltonian Monte Carlo introduced in Neal (2011) is applied to sample $\mu$ and $\Theta$. Hamilton dynamics, rather than a probability distribution, is adopted to propose draws in the Markov chain. Unlike the random walk proposal, Hamilton dynamics produces distant proposals that explore the target distribution more efficiently. The Gibbs sampler handles the estimation of covariance matrices $\{\Phi_1, \Phi_2, \ldots, \Phi_K\}$. The concentration parameter $\alpha$ is sampled using the method in Escobar & West (1994). The estimation details can be found in Appendix 7.1.

### 3.4 Covariance Matrix Estimator

Incorporating the adjustment for the moving average-based estimator in Hansen et al. (2008), the target quantity $V$ is estimated by the posterior mean of the summation of the intraperiod covariance matrix adjusted by the moving average coefficient matrix $\Theta$.

$$E[V|r_{1:n}] = E\left[ (I + \Theta) \sum_{i=1}^{n} \Sigma_i (I + \Theta)' r_{1:n} \right],$$

where $I$ is the identity matrix.

$E[V|r_{1:n}]$ can be estimated by integrating out the parameters and distributional uncertainties. Based on $G$ MCMC outputs, the Bayesian nonparametric covariance estimator
(BNC) is defined as

\[
\text{BNC} = \frac{1}{G} \sum_{g=1}^{G} (I + \Theta^{(g)}) \left( \sum_{i=1}^{n} \Sigma_{i}^{(g)} \right) (I + \Theta^{(g)})' = \frac{1}{G} \sum_{g=1}^{G} (I + \Theta^{(g)}) \left( \sum_{i=1}^{n} \Phi_{i_{g}}^{(g)} \right) (I + \Theta^{(g)})'.
\] (14)

Appendix 7.3 proves that the proposed covariance estimator unbiasedly estimates the ex-post covariance in the presence of independent microstructure noise and nonsynchronous trading if no zero-return problem exists.

The proposed covariance estimation method does not rely on asymptotic assumptions, and the exact finite sample results can be obtained directly. For instance, the posterior distributions of any functions of the covariance matrix, such as realized beta or correlation, are readily available from MCMC outputs, while the classical estimator relies on the asymptotic distribution or delta method to approximate those results.

The proposed method has two mechanisms to avoid obtaining singular covariance matrices. First, the inverse Wishart base function serves as the conjugate prior, and all draws from the conditional posterior of \( \Sigma_{i} \) are positive definite, which ensures the positive definiteness of the proposed estimator. Second, the estimation of a large covariance matrix may face the challenge that the number of sampled observations is less than the number of dimensions. For example, the number of returns synchronized using the refresh time scheme is limited to the most inactive asset. Under the previous-tick scheme, the data may contain many zeros due to the slowly updated price. The proposed synchronization approach fills missing price gaps using data augmentation, so the grid length can be adjusted to make the number of nonzero return vectors greater than the data dimension.

Compared with the multivariate MA-based covariance estimator proposed in Hansen et al. (2008), the BNC estimator has three improvements. First, the proposed method exploits pooling. Second, the Bayesian nonparametric estimator is positive definite while the MA-based estimator is positive semidefinite. Finally, the MA-based estimator requires more
moving average coefficient matrices to handle nonsynchronously observed data. Estimating multiple moving average coefficient matrices is challenging and less precise estimates jeopardize the quality of the covariance estimator. Since the sampled data synchronized using the proposed method have only first-order dependence, only one moving average coefficient matrix needs to be estimated when calculating the proposed covariance estimator.

4 Simulation Results

The data generating process used in Barndorff-Nielsen et al. (2011) is adapted to simulate the data. The fundamental log prices are generated from the following multivariate factor stochastic volatility model

\[
dp^{(j)} = m^{(j)} dt + \rho^{(j)} \sigma^{(j)} dB^{(j)} + \sqrt{1 - \rho^{(j)2}} \sigma^{(j)} dw, \tag{15}
\]

\[
\sigma^{(j)} = \exp(\beta_0^{(j)} + \beta_1^{(j)} v^{(j)}), \tag{16}
\]

\[
dv^{(j)} = \alpha^{(j)} v^{(j)} dt + dB^{(j)}, \tag{17}
\]

where \(w\) and \(B^{(j)}\) are standard Brownian motions, \(\text{cor}(dw, dB^{(j)}) = 0\) and \((m^{(j)}, \beta_0^{(j)}, \beta_1^{(j)}, \alpha^{(j)}, \rho^{(j)}) = (0.04, -0.3125, 0.125, -0.025, -0.3)\) for \(j = 1, \ldots, d\). Prices are simulated every one second \((N = 23400)\). Error terms are added to the fundamental prices to obtain data with microstructure noise.

\[
\tilde{p}^{(j)}_i = p^{(j)}_i + \epsilon^{(j)}_i, \quad \epsilon^{(j)}_i \sim N(0, \sigma_{\epsilon}^{(j)}), \quad \sigma_{\epsilon}^{(j)} = \xi^2 \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\epsilon^{(j)}_i)^4}, \tag{18}
\]

where \(\xi^2\) represents the noise-to-signal ratio that governs the size of microstructure noise. The error term are assumed to be independent of one another, and the variance of the error increases with volatility. The arrival times of observed prices are simulated from independent Poisson processes. Parameter \(\lambda\) in the Poisson process governs the trading frequency of the simulated data. For example, \(\lambda = 5\) means that the asset prices arrive every 5 seconds on
average.

The estimation target is the daily ex-post covariance matrix $V = \sum_{i=1}^{N} \Sigma_{i}$, where $\Sigma_{i}^{(jk)} = \sqrt{1 - \rho_{ij}^{2}}\sigma_{i}^{(j)}\sqrt{1 - \rho_{ik}^{2}}\sigma_{i}^{(k)}$. The RC is formed using 5-minute previous-tick returns to reduce the influence of noise and nonsynchronous trading. Returns synchronized by the refresh time scheme are used to calculate RK. Synchronization with data augmentation is used for BNC, and the grid length is set to be four times the average duration of price changes. The estimation is based on 5,000 MCMC runs, after 10,000 burnin draws. The prior for $\mu$ is $N(0, \Lambda)$, where $\Lambda$ is a diagonal matrix with a diagonal value of $0.01/n$. The prior of the elements of $\Theta$ is assumed to be $N(0, 0.5)$, and the hierarchical prior on $\alpha$ is $\text{Gamma}(4, 8)$. The hyperparameters $\Psi$ and $\nu$ are calculated following equation (8) and equation (9).

The estimation accuracies of RC, RK and BNC are compared under different data frequencies, microstructure noise levels and data dimensions. Both three-asset and ten-asset cases are considered. The norm of the RMSE matrix and average RMSE in estimating both diagonal and off-diagonal elements are reported in Table 1. Panels A and B show the results under noise levels $\xi^{2} = 0.001$ and $\xi^{2} = 0.003$, respectively. As expected, the 5-minute RC provides less precise estimates because only 78 observations are used. RK and BNC benefit from high-frequency data and the estimation errors decrease as the sampling frequency increases. The top performer is the BNC estimator, which yields the lowest RMSE in estimating the ex-post covariance matrix, especially the off-diagonal elements, in all 16 cases. For instance, in the ten-asset case with frequency $(\lambda^{(j)} \in \{5, 6, 8\})$, switching from RK to BNC reduces the RMSE norm from 1.1174 to 1.0180, and the average RMSE of covariance estimation from 0.6584 to 0.5876. The improvement is greater than 10%.

5 Empirical Applications

The tick prices and national best bid and offer (NBBO) prices of 10 equities (stock symbols: BAC, CAT, DD, F, GIS, JNJ, KO, T, WMT, and XOM) and the Standard & Poor’s Depos-
itory Receipt (SPY) from July 1, 2014 to June 29, 2016 are obtained from Tickdata. The clearing procedure used by Barndorff-Nielsen et al. (2009) is applied to clear the data. The sample contains 503 trading days.

The 5-minute RC, RK and BNC estimator are applied to estimate the daily covariance matrix of the ten assets. The synchronization methods and estimation details are the same as in the simulation section.

5.1 Covariance, Correlation and Realized Beta

Daily returns are not influenced by microstructure noise and nonsynchronous trading, so the sample covariance matrix calculated using open-to-close returns provides a benchmark to assess the unbiasedness of the covariance estimators. Table 2 reports the open-to-close daily covariance and averages of RC$^{5m}$, RK and BNC over the sample period. BNC is closer to the sample covariance of open-to-close returns than are RC$^{5m}$ and RK. The average bias$^3$ of BNC is 0.009, while RC$^{5m}$ and RK have higher biases of 0.053 and 0.020, respectively.

Figure 3 provides the dynamic correlation series between BAC and CAT based on RC$^{5m}$, RK and BNC. The correlation estimates based on RK and BNC share similar dynamics, while the correlation implied from the 5-minute RC is more volatile.

The realized beta of asset $i$ is defined as $\beta_t = \frac{\hat{\mathbf{V}}_{ij}^t}{\hat{\mathbf{V}}_{jj}^t}$, where $\hat{\mathbf{V}}_t$ is the ex-post covariance estimator on day $t$ and the $j^{th}$ asset is the market index. Figure 4 plots the realized beta of BAC based on RC$^{5m}$, RK and BNC in a $2 \times 2$ case of BAC and SPY. $\beta_{t}^{\text{RK}}$ and $\beta_{t}^{\text{BNC}}$ have very similar paths, while RC provides more volatile realized beta estimates. Table 3 shows the estimation results of the ARMA(1,1) model for the three versions of the realized beta, which also confirms that RK and BNC share similar time series dynamics.

$^3$The bias is measured as $\sum_{i=1}^{D} \sum_{j=1}^{D} e_{ij}$, where $e_{ij}$ is the difference between the $(i,j)$ elements of daily covariance and the average of high-frequency covariance estimator.
5.2 Portfolio Allocation Evaluation

Whether the proposed estimator improves portfolio allocation is worth exploring. Following Fleming et al. (2003), the performance of minimum-variance portfolios based on high-frequency covariance estimators are compared. Consider an investor who constructs her portfolio using ten stocks (BAC, CAT, DD, F, GIS, JNJ, KO, T, WMT and XOM) and a risk-free asset. She applies the volatility-timing strategy to adjust the portfolio weights each day by solving the following minimization problem given the desired portfolio return $\mu_0$.

$$
\min_{w_t} \mathbf{w}_t' \hat{\Sigma}_t \mathbf{w}_t \quad \text{s.t.} \quad \mathbf{w}_t' \bar{\mu} = \mu_0 \quad \text{and} \quad \mathbf{w}_t' \mathbf{1} = 1,
$$

where $\mathbf{w}_t$ stands for portfolio weights on day $t$, $\hat{\Sigma}_t$ is the covariance matrix, and $\bar{\mu}$ is the return mean vector. The solution to the minimization problem is

$$
\mathbf{w}_t = \frac{\hat{\Sigma}_t^{-1} \bar{\mu}}{\bar{\mu}' \hat{\Sigma}_t^{-1} \bar{\mu}} \mu_0. \tag{20}
$$

Based on ex-post covariance estimates, the next period covariance is predicted using an exponential smoother

$$
\hat{\Sigma}_t = \exp(-\kappa) \hat{\Sigma}_{t-1} + \kappa \exp(-\kappa) \hat{V}_{t-1}, \tag{21}
$$

where $\kappa$ is the decay rate and $\hat{V}_{t-1}$ is the ex-post covariance estimator, which can be RC$^{5m}$, RK or BNC. The return mean $\bar{\mu}$ is set to the sample mean. If the summation of $\mathbf{w}_t$ differs from 100%, the remaining portion of capital is allocated to the risk-free asset, so that the portfolio return on day $t$ equals $r_t^p = \mathbf{w}_t' \mathbf{r}_t + (1 - \mathbf{w}_t' \mathbf{1}) r_f^p$.

Table 4 shows the mean, variance and Sharpe ratio of portfolios under various decay rates ($\kappa = 0.03, 0.06$ and $0.09$) and required annual rates of return (10%, 20% and 30%). The portfolio based on the BNC estimator has the highest Sharpe ratio in all 9 cases. For example, given $\kappa = 0.06$ and a 20% required return, the Sharpe ratio of the BNC portfolio is 0.319, while the ratios of RC$^{5m}$ and RK-based portfolios are 0.314 and 0.312, respectively.
The utility-based approach used in Fleming et al. (2003) is applied. Assume that the investor has the following quadratic utility function.

\[
U(r^p_t) = W_0 \left[ (1 + r^f_t + r^p_t) - \frac{\gamma}{2(1 + \gamma)} (1 + r^f_t + r^p_t)^2 \right],
\]

(22)

where \(r^p_t\) is the portfolio return on day \(t\), \(r^f_t\) is the daily risk-free rate and \(\gamma\) stands for the risk aversion coefficient. The performance fee \(\Delta\) is the cost that an investor would pay to switch from the benchmark portfolio to an alternative. \(\Delta\) is the value that equates the utility levels under two competing portfolios.

\[
\sum_{t=1}^{T} U(r^p_1) = \sum_{t=1}^{T} U(r^p_2 - \Delta).
\]

(23)

The portfolio based on the 5-minute RC serves as the benchmark portfolio and its return is \(r^p_1\).

Table 5 lists the annualized basis point fees that an investor with quadratic utility would like to pay to switch from the benchmark portfolio to portfolios using RK or BNC. Under different decay rates and required portfolio returns, both a less risk-averse investor (\(\gamma = 1\)) and a more conservative investor (\(\gamma = 10\)) would be willing to pay higher performance fees to obtain BNC estimates to guide the portfolio allocation decision. For example, in the case with \(\kappa = 0.06\) and \(\mu_0 = 20\%\), an investor with \(\gamma = 10\) would pay over 30 extra basis points to choose the portfolio based on BNC, instead of 5-minute RC.

6 Conclusion

This paper proposes a Bayesian nonparametric method for estimating the covariance matrix for nonsynchronous prices contaminated with microstructure noise. The proposed method improves the covariance matrix estimation in four aspects. First, pooling observations with similar covariance increases the precision of the ex-post covariance estimation. Second, the
covariance estimator is guaranteed to be positive definite. Third, the Bayesian approach delivers exact finite sample results without relying on any infill asymptotic assumptions. Finally, a new synchronization method with data augmentation is introduced to improve the previous-tick synchronization method.

Monte Carlo simulations confirm that the Bayesian nonparametric approach provides more precise ex-post covariance matrix estimates. Empirical applications to equity returns show that the correlation and realized beta implied by the BNC estimator have similar time series dynamics as that obtained with the RK. The minimum-variance portfolio based on the proposed estimator outperforms the portfolio formed using 5-minute RC or RK in terms of the Sharpe ratio and utility level.
References


Large, J. (2007), Accounting for the epps effect: Realized covariation, cointegration and common factors, Manuscript, University of Oxford.


7 Appendix

7.1 Estimation Steps

1. Sampling $\mu_{1:n}, \Phi_{1:K}, \Theta, s_{1:n}$:

   Given prior: $\mu \sim N(m, H)$, the conditional posterior of $\mu$ is

   \[
   p(\mu | r_{1:n}, \Theta, \Phi_{1:K}) \propto p(\mu) \prod_{i=1}^{n} p(r_i | \mu + \Theta \eta_{i-1}, \Phi_{s_i})
   \]

   \[
   \propto \exp \left\{ -\frac{1}{2} \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} + H^{-1} \right) \mu - \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} (r_i - \Theta \eta_{i-1}) + H^{-1} m \right) \right\}. \tag{24}
   \]

   where $\eta_{i-1} = r_{i-1} - \mu - \Theta \eta_{i-2}$.

   Define $U_1(\mu) = -\log [p(\mu | r_{1:n}, \Theta, \Phi_{1:K})]$, we have

   \[
   U_1(\mu) = \frac{1}{2} \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} + H^{-1} \right) \mu - \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} (r_i - \Theta \eta_{i-1}) + H^{-1} m \right). \tag{25}
   \]

   \[
   \frac{\partial U_1(\mu)}{\partial \mu} = \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} + H^{-1} \right) \mu - \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} (r_i - \Theta \eta_{i-1}) + H^{-1} m \right). \tag{26}
   \]

   The Hamiltonian dynamics is used as the proposal and is approximated by the leapfrog method with step $L$ and stepsize $\epsilon$. Define a function $K_1(B_1) = \sum_{j=1}^{d} \frac{(B_1^{(j)})^2}{2}$, where $B_1$ is an auxiliary $d$-dimensional vector. A draw from the Hamilton dynamics is generated using the following steps.

   (1) Initialize $B_1$ as $B_1^{(j)} \sim N(0, 1)$.

   (2) $B_1' = B_1 - \frac{\epsilon}{2} \frac{\partial U_1(\mu_{m-1})}{\partial \mu}$ and $\mu' = \mu^{(m-1)}$, where $\mu^{(m-1)}$ is the value of $\mu$ in the previous iteration.

   (3) For $l$ from $1$ to $L$, $\mu' = \mu' + \epsilon B_1'$.

      a. $B_1' = B_1' - \frac{\epsilon}{2} \frac{\partial U_1(\mu_{m-1})}{\partial \mu}$, if $l < L$.

      b. $B_1' = B_1' - \frac{\epsilon}{2} \frac{\partial U_1(\mu_{m-1})}{\partial \mu}$, if $l = L$. 

   22
Update $\mu = \mu'$ with acceptance rate $\min \{1, \exp(U_1(\mu_0) + K_1(B_1) - U_1(\mu') - K_1(B_1'))\}$. 

$\epsilon$ is adjusted every 10 MCMC iterations. If the average acceptance rate is zero, set $\epsilon = 0.9\epsilon$. If the average acceptance rate is above 0.8, adjust $\epsilon = 1.1\epsilon$. $L = 20$.

2. Sampling $\Theta | r_{1:n}, \mu, \Phi_{1:K}, s_{1:n}$:

Given prior $\Theta_{jk} \sim N(m_{jk}, v_{jk}^2)$, the conditional posterior of $\Theta$ is

$$p(\Theta | r_{1:n}, \Theta, \Phi_{1:K}) \propto \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} \left[ \eta_i' \Theta' \Phi_{s_i}^{-1} (\Theta \eta_i - 2(r_i - \mu)) \right]\right\} \cdot \prod_{j=1}^{d} \prod_{k=1}^{d} \exp\left[-\frac{(\Theta_{jk} - m_{jk})^2}{2v_{jk}^2}\right]$$  \hspace{1cm} (27)

Define $U_2(\Theta) = -\log [p(\Theta | r_{1:n}, \Theta, \Phi_{1:K})]$, we have

$$U_2(\Theta) = -\frac{1}{2} \sum_{i=1}^{n} \left[ \eta_i' \Theta' \Phi_{s_i}^{-1} (2(r_i - \mu) - \Theta \eta_i) \right] + \sum_{j=1}^{d} \sum_{k=1}^{d} \left[ \frac{(\Theta_{jk} - m_{jk})^2}{2v_{jk}^2} \right]$$  \hspace{1cm} (28)

$$\frac{\partial U_2(\Theta)}{\partial \Theta} = -\sum_{i=1}^{n} \Phi_{s_i}^{-1} (r_i - \mu - \Theta \eta_i) \eta_i' - \sum_{j=1}^{d} \sum_{k=1}^{d} \frac{(\Theta_{jk} - m_{jk})^2}{v_{jk}^2} \Phi_{jk}$$  \hspace{1cm} (29)

Define function $K_2(B_2) = \sum_{j=1}^{d} \sum_{k=1}^{d} \frac{(B_{jk}^2)^2}{2}$, where $B_2$ is an auxiliary $d \times d$ matrix.

(1) Initialize each element of $B_2$ from $N(0, 1)$.

(2) $B_2' = B_2 - \frac{\epsilon}{2} \frac{\partial U_2(\Theta_0)}{\partial \Theta}$ and $\Theta' = \Theta^{(m-1)}$, where $\Theta^{(m-1)}$ is the result of the previous iteration.

(3) For $l$ from 1 to $L$, $\Theta' = \Theta' + \epsilon B_2'$.

a. $B_2' = B_2' - \frac{\epsilon}{2} \frac{\partial U_2(\Theta')}{\partial \Theta}$, if $l < L$.

b. $B_2' = B_2' - \frac{\epsilon}{2} \frac{\partial U_2(\Theta')}{\partial \Theta}$, if $l = L$.

Accept $\Theta'$ with acceptance rate $\min \{1, \exp(U_2(\Theta_0) + K_2(B_2) - U_2(\Theta') - K_2(B_2'))\}$. 

23
3. Sampling $\Phi_j | r_{1:n}, \mu, \Theta, s_{1:n}$ for $j = 1, \ldots, K$:

Given prior $\Phi_j \sim \text{IW}(\Psi, \nu)$, the conditional posterior of $\Phi_j$ is

$$p(\Phi_j | r_{1:n}, s_{1:n}, \mu, \Theta) \propto p(\Phi_j) \prod_{s_i = j} p(r_i | \mu + \Theta \eta_{i-1}, \Phi_j)$$

$$\propto |\Phi_j|^{-\frac{1}{2}(n_j + \nu + d + 1)} \exp \left[ -\frac{1}{2} \text{tr}((\Psi + Q_j)\Phi_j^{-1}) \right] \sim \text{IW}(\Psi + Q_j, n_j + \nu),$$ (30)

where $Q_j = \sum_{s_i = j} (r_i - \mu - \Theta \eta_{i-1})(r_i - \mu - \Theta \eta_{i-1})'$ and $n_j = \sum_{i=1}^n \mathbb{1}(s_i = j)$.

4. Sampling $v_j | s_{1:n}$ and calculate $w_j$ for $j = 1, \ldots, K$:

$$p(v_j | s_{1:n}, \alpha) \sim \text{Beta} \left( 1 + \sum_{i=1}^n \mathbb{1}(s_i = j), \alpha + \sum_{i=1}^n \mathbb{1}(s_i > j) \right).$$ (31)

$w_j$ are computed as $w_1 = v_1$, and $w_j = v_j \prod_{l=1}^{j-1} (1 - v_l)$.

5. Sampling $u_i | w_i, s_{1:n}$ for $i = 1, \ldots, n$:

$$p(u_i | s_{1:n}, w_{1:K}) \sim \text{Unif}(0, w_{s_i}).$$ (32)

6. Find the smallest $K$ such that $\sum_{j=1}^K w_j > 1 - \min(u_{1:n})$.

7. Sampling $s_i | r_{1:n}, \mu, \Theta, \Phi_{1:K}, s_{-i}, u_{1:n}$ for $i = 1, \ldots, n$:

$$p(s_i = j | r_i, \mu, \Phi_{1:K}, w_{1:K}, u_i) \propto \sum_{j=1}^K \mathbb{1}(w_j > u_i) N(r_i | \mu + \Theta \eta_{i-1}, \Phi_j).$$ (33)

8. Sampling $\alpha | K$:

Given prior $\alpha \sim \text{Gamma}(a, b)$,

$$p(\alpha | K) \sim q \cdot \text{Gamma}(a + K, b - \log \zeta) + (1 - q) \cdot \text{Gamma}(a + K - 1, b - \log \zeta).$$ (34)
where \( q = \frac{a + K - 1}{a + K - 1 + n(b - \log \zeta)} \) and \( \zeta \sim \text{Beta}(\alpha + 1, n) \).

### 7.2 Estimation Steps for Missing Observations

If \( b \) of \( d \) elements in price vector \( \mathbf{p}_i \) are missing, the sampling of missing price records is conditional on the \( q = d - b \) observed prices, the adjacent prices \( \mathbf{p}_{i-1} \) and \( \mathbf{p}_{i+1} \), the mean vector \( \mathbf{\mu} \) and the covariance matrices \( \Sigma_{i} \) and \( \Sigma_{i+1} \). \( \mathbf{r}_i \) and \( \mathbf{r}_{i+1} \) provide a link between the model and the missing observations. First, \( \mathbf{r}_i, \mathbf{r}_{i+1}, \mathbf{p}_i, \mathbf{p}_{i+1}, \mathbf{\mu}, \Sigma_i \) and \( \Sigma_{i+1} \) are split into two groups, one corresponding to the \( b \) missing observations in \( \mathbf{p}_i \), and the other corresponding to the \( q \) observed prices.

\[
\begin{align*}
\mathbf{p}_{i-1} &= \begin{bmatrix} p_{i-1}^b \\ p_{i-1}^q \end{bmatrix}, \quad \mathbf{p}_i = \begin{bmatrix} p_i^b \\ p_i^q \end{bmatrix}, \quad \mathbf{p}_{i+1} = \begin{bmatrix} p_{i+1}^b \\ p_{i+1}^q \end{bmatrix}, \quad \mathbf{r}_i = \begin{bmatrix} r_i^b \\ r_i^q \end{bmatrix}, \quad \mathbf{r}_{i+1} = \begin{bmatrix} r_{i+1}^b \\ r_{i+1}^q \end{bmatrix} \\
\mathbf{\mu} &= \begin{bmatrix} \mu^b \\ \mu^q \end{bmatrix}, \quad \Sigma_i = \begin{bmatrix} \Sigma_{bb}^i & \Sigma_{bq}^i \\ \Sigma_{qb}^i & \Sigma_{qq}^i \end{bmatrix} \quad \text{and} \quad \Sigma_{i+1} = \begin{bmatrix} \Sigma_{b+1}^{bb} & \Sigma_{b+1}^{bq} \\ \Sigma_{q+1}^{qb} & \Sigma_{q+1}^{qq} \end{bmatrix}.
\end{align*}
\]

The conditional distributions of \( \mathbf{r}_i^b \) and \( \mathbf{r}_{i+1}^b \) given observed \( \mathbf{r}_i^q \) and \( \mathbf{r}_{i+1}^q \) are

\[
\begin{align*}
\mathbf{r}_i^b | \mathbf{r}_i^q &= \mathbf{p}_i^b - \mathbf{p}_{i-1}^b | \mathbf{r}_i^q \sim \mathcal{N} \left( \overline{\mathbf{\mu}}_i, \Sigma_i \right), \\
\mathbf{r}_{i+1}^b | \mathbf{r}_{i+1}^q &= \mathbf{p}_{i+1}^b - \mathbf{p}_{i+1}^b | \mathbf{r}_{i+1}^q \sim \mathcal{N} \left( \overline{\mathbf{\mu}}_{i+1}, \Sigma_{i+1} \right),
\end{align*}
\]

where \( \overline{\mathbf{\mu}}_i \) and \( \overline{\Sigma}_i \) are the mean and covariance of the distribution of \( \mathbf{r}_i^b \) conditional on \( \mathbf{r}_i^q \).

The conditional mean has a moving average dynamics and is derived as

\[
\begin{align*}
\overline{\mathbf{\mu}}_i &= \mu^b + (\Theta \eta_{i-1})^b + \Sigma_{i}^{bq}(\Sigma_{i}^{qq})^{-1}(\mathbf{r}_i^q - \mu^q - (\Theta \eta_{i-1})^q), \\
\Sigma_i &= \Sigma_{i}^{bb} - \Sigma_{i}^{bq}(\Sigma_{i}^{qq})^{-1}(\Sigma_{i}^{qb})'.
\end{align*}
\]

(37) and (38) provide the mean and covariance for the next period, \( \overline{\mathbf{\mu}}_{i+1} \) and \( \overline{\Sigma}_{i+1} \) can be derived similarly.

25
The density of \( p_i^b \) conditional on observed prices and model parameters is given as

\[
\pi (p_i^b | \cdots) \propto \exp \left\{ -\frac{1}{2} \left[ p_i^b \Sigma_i^{-1} p_i^b - 2p_i^b \Sigma_i^{-1} p_{i-1} + \bar{\mu}_i \right] \right. \\
- \frac{1}{2} \left[ p_i^b \Sigma_{i+1}^{-1} p_i^b - 2p_i^b \Sigma_{i+1}^{-1} (p_{i+1} - \bar{\mu}_{i+1}) \right] \right\}
\]

\sim N(m^b, H^b),

(39)

where

\[
m^b = H^b \left[ \Sigma^{-1}_i (p_{i-1} + \bar{\mu}_i) + \Sigma_{i+1}^{-1} (p_{i+1} - \bar{\mu}_{i+1}) \right],
\]

(40)

\[
H^b = \left( \Sigma^{-1}_i + \Sigma^{-1}_{i+1} \right)^{-1}.
\]

(41)

### 7.3 Proof of the Unbiasedness of the BNC Estimator

Let \( r_i \) denotes the \( i^{th} \) regularly spaced return vector constructed from log price with noise term \( \epsilon_i \sim N(0, \Omega_i) \), where \( \Omega_i \) is a diagonal matrix. Assuming there is no zero-return bias, there exists first order dependence between adjacent returns. Let \( \Gamma_i \) be the matrix that captures the lead-lag dependence relationship, where \( \Gamma_i^{jj} = 0 \) and \( \Gamma_i^{jk} = \text{cov}(r_i^{(j)}, r_{i+1}^{(k)}) \).

The covariance and first-order autocovariance of \( r_i \) are

\[
\text{cov}(r_i) = V_i + \Omega_i - \Gamma_i - \Gamma_i = V_i + \Xi_{i-1} + \Xi_i,
\]

(42)

\[
\text{cov}(r_i, r_{i-1}) = -\Omega_i - \Gamma_i - \Gamma_i = -\Xi_{i-1}.
\]

(43)

where \( \Xi_i = \Omega_i - \Gamma_i \) and \( V_i \) is the true intraperiod covariance matrix.

Consider the following heteroskedastic vector moving average model for \( r_i \),

\[
r_i = \mu + \Theta \epsilon_{i-1} + \eta_i, \quad \eta_i \sim N(0, \Sigma_i),
\]

(44)
The model implies that

\[
\text{cov}(r_i) = \Theta \Sigma_{i-1} \Theta' + \Sigma_i, \quad (45)
\]
\[
\text{cov}(r_i, r_{i-1}) = \Theta \Sigma_{i-1}. \quad (46)
\]

Equating (43) and (46), we have

\[
-\Xi_{i-1} = \Theta \Sigma_{i-1} \quad \text{and} \quad -\Xi_i = \Theta \Sigma_i. \quad (47)
\]

Equating (42) and (45) and using the result in (47), we have

\[
V_i = \Theta \Sigma_{i-1} \Theta' + \Sigma_i - \Xi_i - \Xi_{i-1} = -\Xi_{i-1} \Theta' - \Theta^{-1} \Xi_i - \Xi_i - \Xi_{i-1}
\]
\[
= -(I + \Theta') \Xi_{i-1} - (I + \Theta^{-1}) \Xi_i \quad (48)
\]

Using the results in (48) and (47), the summations of \(V_i\) and \(\Sigma_i\), over \(i = 1, \ldots, n\), are

\[
\sum_{i=1}^{n} V_i = -(I + \Theta') \sum_{i=1}^{n} \Xi_{i-1} - (I + \Theta^{-1}) \sum_{i=1}^{n} \Xi_i, \quad (49)
\]
\[
\sum_{i=1}^{n} \Sigma_i = -\Theta^{-1} \sum_{i=1}^{n} \Xi_i. \quad (50)
\]

The ratio between (49) and (50) is

\[
\sum_{i=1}^{n} V_i \left(\sum_{i=1}^{n} \Sigma_i\right)^{-1} = (I + \Theta) \sum_{i=1}^{n} \Xi_{i-1} \left(\sum_{i=1}^{n} \Xi_i\right)^{-1} \Theta + (I + \Theta^{-1}) \sum_{i=1}^{n} \Xi_i \left(\sum_{i=1}^{n} \Xi_i\right)^{-1} \Theta
\]
\[
= (I + \Theta) \Theta + (I + \Theta^{-1}) \Theta
\]
\[
= (I + \Theta)(I + \Theta)'. \quad (51)
\]

Finally, we have

\[
\sum_{i=1}^{n} V_i = (I + \Theta) \sum_{i=1}^{n} \Sigma_i (I + \Theta), \quad \text{if} \quad \Xi_0 = \Xi_n. \quad (52)
\]
Table 1: RMSEs of RC, RK and BNC

Panel A: $\xi^2 = 0.001$

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Frequency $\lambda$</th>
<th>RC $^{5m}$</th>
<th>RK</th>
<th>BNC</th>
<th>RC $^{5m}$</th>
<th>RK</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Assets</td>
<td>(80, 90, 120)</td>
<td>0.6113</td>
<td>0.4465</td>
<td><strong>0.4388</strong></td>
<td>0.3399</td>
<td>0.3317</td>
<td><strong>0.3344</strong></td>
</tr>
<tr>
<td></td>
<td>(30, 40, 60)</td>
<td>0.5150</td>
<td>0.4044</td>
<td><strong>0.3562</strong></td>
<td>0.3499</td>
<td>0.2993</td>
<td><strong>0.2697</strong></td>
</tr>
<tr>
<td></td>
<td>(10, 15, 20)</td>
<td>0.4112</td>
<td>0.2663</td>
<td><strong>0.2212</strong></td>
<td>0.3187</td>
<td>0.2001</td>
<td><strong>0.1666</strong></td>
</tr>
<tr>
<td></td>
<td>(5, 6, 8)</td>
<td>0.4853</td>
<td>0.2658</td>
<td><strong>0.2020</strong></td>
<td>0.3759</td>
<td>0.2006</td>
<td><strong>0.1542</strong></td>
</tr>
<tr>
<td>10 Assets</td>
<td>(80, 90, 120)</td>
<td>2.6436</td>
<td>2.1159</td>
<td><strong>1.9344</strong></td>
<td>0.9073</td>
<td>1.0384</td>
<td><strong>1.0326</strong></td>
</tr>
<tr>
<td></td>
<td>(30, 40, 60)</td>
<td>1.8699</td>
<td>1.6864</td>
<td><strong>1.3644</strong></td>
<td>0.9043</td>
<td>0.8668</td>
<td><strong>0.7725</strong></td>
</tr>
<tr>
<td></td>
<td>(10, 15, 20)</td>
<td>1.4427</td>
<td>1.1661</td>
<td><strong>0.9173</strong></td>
<td>0.7818</td>
<td>0.5819</td>
<td><strong>0.4897</strong></td>
</tr>
<tr>
<td></td>
<td>(5, 6, 8)</td>
<td>1.4337</td>
<td>0.9244</td>
<td><strong>0.7377</strong></td>
<td>0.8061</td>
<td>0.4572</td>
<td><strong>0.3538</strong></td>
</tr>
</tbody>
</table>

Panel B: $\xi^2 = 0.003$

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Frequency $\lambda$</th>
<th>RC $^{5m}$</th>
<th>RK</th>
<th>BNC</th>
<th>RC $^{5m}$</th>
<th>RK</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Assets</td>
<td>(80, 90, 120)</td>
<td>0.9435</td>
<td>0.4836</td>
<td><strong>0.4826</strong></td>
<td>0.7530</td>
<td>0.3609</td>
<td>0.3707</td>
</tr>
<tr>
<td></td>
<td>(30, 40, 60)</td>
<td>0.9549</td>
<td>0.4600</td>
<td><strong>0.4144</strong></td>
<td>0.8424</td>
<td>0.3426</td>
<td><strong>0.3193</strong></td>
</tr>
<tr>
<td></td>
<td>(10, 15, 20)</td>
<td>0.8317</td>
<td>0.3244</td>
<td><strong>0.2754</strong></td>
<td>0.7650</td>
<td>0.2479</td>
<td><strong>0.2104</strong></td>
</tr>
<tr>
<td></td>
<td>(5, 6, 8)</td>
<td>0.9940</td>
<td>0.3274</td>
<td><strong>0.2982</strong></td>
<td>0.9145</td>
<td>0.2499</td>
<td><strong>0.2377</strong></td>
</tr>
<tr>
<td>10 Assets</td>
<td>(80, 90, 120)</td>
<td>3.3068</td>
<td>2.2263</td>
<td><strong>2.1770</strong></td>
<td>1.9305</td>
<td>1.1007</td>
<td>1.2609</td>
</tr>
<tr>
<td></td>
<td>(30, 40, 60)</td>
<td>2.8757</td>
<td>1.8408</td>
<td><strong>1.7201</strong></td>
<td>2.1235</td>
<td>0.9575</td>
<td>1.0747</td>
</tr>
<tr>
<td></td>
<td>(10, 15, 20)</td>
<td>2.4215</td>
<td>1.3586</td>
<td><strong>1.2227</strong></td>
<td>1.8692</td>
<td>0.6926</td>
<td>0.7102</td>
</tr>
<tr>
<td></td>
<td>(5, 6, 8)</td>
<td>2.4180</td>
<td>1.1174</td>
<td><strong>1.0180</strong></td>
<td>1.8902</td>
<td>0.5685</td>
<td><strong>0.5156</strong></td>
</tr>
</tbody>
</table>

Results are based on 2,000 days’ simulation. $||X|| = \frac{1}{T-t_{0}} \sum_{t=t_{0}}^{T} ||\hat{V}_{t} - V_{t}||$, where $||X|| = \sqrt{\sum_{i} \sum_{j} x_{ij}^{2}}$. RMSE(diag) stands for the average of RMSEs for diagonal elements of covariance matrices. RMSE(off-diag) is the average of RMSEs for off-diagonal elements. Bayesian nonparametric estimator BNC is estimated based on on 5,000 MCMC runs, after 10,000 burn-in.
<table>
<thead>
<tr>
<th></th>
<th>BAC</th>
<th>CAT</th>
<th>DD</th>
<th>F</th>
<th>GIS</th>
<th>JNJ</th>
<th>KO</th>
<th>T</th>
<th>WMT</th>
<th>XOM</th>
<th>Average</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Daily Covariance (Open-to-close)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.4604</td>
<td>-</td>
</tr>
<tr>
<td>BAC</td>
<td>1.798</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>0.775</td>
<td>1.613</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD</td>
<td>0.622</td>
<td>0.708</td>
<td>1.515</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.850</td>
<td>0.824</td>
<td>0.542</td>
<td>1.735</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GIS</td>
<td>0.201</td>
<td>0.222</td>
<td>0.239</td>
<td>0.297</td>
<td>0.700</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JNJ</td>
<td>0.413</td>
<td>0.359</td>
<td>0.330</td>
<td>0.387</td>
<td>0.271</td>
<td>0.711</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>0.278</td>
<td>0.266</td>
<td>0.258</td>
<td>0.303</td>
<td>0.348</td>
<td>0.302</td>
<td>0.599</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.282</td>
<td>0.350</td>
<td>0.243</td>
<td>0.335</td>
<td>0.273</td>
<td>0.287</td>
<td>0.264</td>
<td>0.619</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT</td>
<td>0.317</td>
<td>0.203</td>
<td>0.265</td>
<td>0.337</td>
<td>0.335</td>
<td>0.319</td>
<td>0.288</td>
<td>0.284</td>
<td>1.051</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XOM</td>
<td>0.575</td>
<td>0.735</td>
<td>0.531</td>
<td>0.600</td>
<td>0.287</td>
<td>0.431</td>
<td>0.291</td>
<td>0.388</td>
<td>0.252</td>
<td>1.178</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average of 5-minute Realized Covariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5134</td>
<td>0.0530</td>
</tr>
<tr>
<td>BAC</td>
<td>2.026</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>0.830</td>
<td>1.857</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD</td>
<td>0.652</td>
<td>0.669</td>
<td>1.412</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.907</td>
<td>0.827</td>
<td>0.610</td>
<td>2.175</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GIS</td>
<td>0.327</td>
<td>0.248</td>
<td>0.290</td>
<td>0.342</td>
<td>0.819</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JNJ</td>
<td>0.490</td>
<td>0.340</td>
<td>0.372</td>
<td>0.558</td>
<td>0.399</td>
<td>1.238</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>0.325</td>
<td>0.251</td>
<td>0.301</td>
<td>0.299</td>
<td>0.422</td>
<td>0.351</td>
<td>0.799</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.365</td>
<td>0.330</td>
<td>0.303</td>
<td>0.342</td>
<td>0.272</td>
<td>0.289</td>
<td>0.274</td>
<td>0.796</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT</td>
<td>0.393</td>
<td>0.311</td>
<td>0.312</td>
<td>0.364</td>
<td>0.335</td>
<td>0.358</td>
<td>0.339</td>
<td>0.268</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XOM</td>
<td>0.668</td>
<td>0.725</td>
<td>0.557</td>
<td>0.616</td>
<td>0.293</td>
<td>0.383</td>
<td>0.311</td>
<td>0.359</td>
<td>0.331</td>
<td>1.407</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average of Multivariate Realized Kernel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.4807</td>
<td>0.0203</td>
</tr>
<tr>
<td>BAC</td>
<td>2.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>0.739</td>
<td>1.839</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD</td>
<td>0.626</td>
<td>0.633</td>
<td>1.409</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.784</td>
<td>0.727</td>
<td>0.547</td>
<td>2.168</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GIS</td>
<td>0.332</td>
<td>0.265</td>
<td>0.301</td>
<td>0.315</td>
<td>0.817</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JNJ</td>
<td>0.416</td>
<td>0.301</td>
<td>0.347</td>
<td>0.349</td>
<td>0.326</td>
<td>0.881</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>0.325</td>
<td>0.257</td>
<td>0.285</td>
<td>0.304</td>
<td>0.413</td>
<td>0.289</td>
<td>0.811</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.355</td>
<td>0.305</td>
<td>0.298</td>
<td>0.335</td>
<td>0.266</td>
<td>0.270</td>
<td>0.265</td>
<td>0.794</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT</td>
<td>0.370</td>
<td>0.298</td>
<td>0.305</td>
<td>0.335</td>
<td>0.322</td>
<td>0.316</td>
<td>0.315</td>
<td>0.255</td>
<td>0.965</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XOM</td>
<td>0.629</td>
<td>0.676</td>
<td>0.556</td>
<td>0.577</td>
<td>0.286</td>
<td>0.322</td>
<td>0.306</td>
<td>0.336</td>
<td>0.298</td>
<td>1.431</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average of Bayesian Nonparametric Covariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.4695</td>
<td>0.0090</td>
</tr>
<tr>
<td>BAC</td>
<td>1.910</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td>0.723</td>
<td>1.765</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD</td>
<td>0.619</td>
<td>0.608</td>
<td>1.412</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.750</td>
<td>0.686</td>
<td>0.550</td>
<td>2.044</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GIS</td>
<td>0.357</td>
<td>0.285</td>
<td>0.309</td>
<td>0.357</td>
<td>0.842</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JNJ</td>
<td>0.424</td>
<td>0.287</td>
<td>0.327</td>
<td>0.370</td>
<td>0.326</td>
<td>1.091</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>0.343</td>
<td>0.272</td>
<td>0.300</td>
<td>0.307</td>
<td>0.415</td>
<td>0.281</td>
<td>0.784</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.367</td>
<td>0.305</td>
<td>0.300</td>
<td>0.326</td>
<td>0.266</td>
<td>0.265</td>
<td>0.269</td>
<td>0.769</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT</td>
<td>0.375</td>
<td>0.292</td>
<td>0.309</td>
<td>0.324</td>
<td>0.318</td>
<td>0.298</td>
<td>0.314</td>
<td>0.253</td>
<td>0.936</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XOM</td>
<td>0.621</td>
<td>0.660</td>
<td>0.552</td>
<td>0.560</td>
<td>0.300</td>
<td>0.317</td>
<td>0.307</td>
<td>0.335</td>
<td>0.295</td>
<td>1.391</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: ARMA(1,1) for Realized Beta

<table>
<thead>
<tr>
<th>Parameter</th>
<th>RC^{5m}</th>
<th>RK</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_0$</td>
<td>1.3513</td>
<td>1.1588</td>
<td>1.1197</td>
</tr>
<tr>
<td></td>
<td>(0.1191)</td>
<td>(0.0836)</td>
<td>(0.0862)</td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>0.9815</td>
<td>0.9701</td>
<td>0.9730</td>
</tr>
<tr>
<td></td>
<td>(0.0145)</td>
<td>(0.0135)</td>
<td>(0.0141)</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>-0.8577</td>
<td>-0.6566</td>
<td>-0.7210</td>
</tr>
<tr>
<td></td>
<td>(0.0523)</td>
<td>(0.0521)</td>
<td>(0.0549)</td>
</tr>
</tbody>
</table>

This table reports the regression results of ARMA model: $\beta_t = \phi_0 + \phi_1 \beta_{t-1} + \rho_1 \epsilon_{t-1} + \epsilon_t$. The values in the bracket are the standard error.

Table 4: Summary of Minimum Variance Portfolio Performance

<table>
<thead>
<tr>
<th>$\mu_0$</th>
<th>RC^{5m}</th>
<th>RK</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean  Stdev  SR</td>
<td>Mean  Stdev  SR</td>
<td>Mean  Stdev  SR</td>
</tr>
<tr>
<td>Panel A: Decay Rate $\kappa = 0.03$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>0.116  0.229  0.507</td>
<td>0.116  0.229  0.506</td>
<td>0.117  0.229  0.509</td>
</tr>
<tr>
<td>20%</td>
<td>0.132  0.422  0.314</td>
<td>0.132  0.421  0.313</td>
<td>0.134  0.421  0.317</td>
</tr>
<tr>
<td>30%</td>
<td>0.149  0.624  0.238</td>
<td>0.148  0.623  0.238</td>
<td>0.151  0.623  0.242</td>
</tr>
<tr>
<td>Panel B: Decay Rate $\kappa = 0.06$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>0.116  0.229  0.507</td>
<td>0.115  0.228  0.507</td>
<td>0.117  0.228  0.512</td>
</tr>
<tr>
<td>20%</td>
<td>0.132  0.421  0.314</td>
<td>0.131  0.419  0.312</td>
<td>0.133  0.418  0.319</td>
</tr>
<tr>
<td>30%</td>
<td>0.148  0.623  0.238</td>
<td>0.147  0.620  0.236</td>
<td>0.150  0.618  0.243</td>
</tr>
<tr>
<td>Panel C: Decay Rate $\kappa = 0.09$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>0.115  0.228  0.505</td>
<td>0.115  0.227  0.505</td>
<td>0.116  0.227  0.512</td>
</tr>
<tr>
<td>20%</td>
<td>0.131  0.422  0.311</td>
<td>0.129  0.418  0.310</td>
<td>0.132  0.417  0.318</td>
</tr>
<tr>
<td>30%</td>
<td>0.147  0.624  0.235</td>
<td>0.144  0.618  0.233</td>
<td>0.149  0.616  0.241</td>
</tr>
</tbody>
</table>

This table provides the mean return, variance and Sharpe ratio of portfolios based on RC^{5m}, RK or BNC. The period is from 07/02/2014 to 06/29/2016.
Table 5: Performance Fees

<table>
<thead>
<tr>
<th></th>
<th>RK</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta(\gamma = 1)$</td>
<td>$\Delta(\gamma = 10)$</td>
</tr>
<tr>
<td>$\mu_0$</td>
<td>$\kappa = 0.03$</td>
<td>$\kappa = 0.06$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>$-9.110$</td>
<td>$33.609$</td>
</tr>
<tr>
<td>30%</td>
<td>$-13.623$</td>
<td>$50.483$</td>
</tr>
</tbody>
</table>

The values listed in this table is the annualized base point fees that an investor with quadratic utility and risk aversion coefficient $\gamma$ is willing to pay for switching portfolio based on RC$^{5m}$ to portfolio based on RK or BNC. Assuming one year contains 252 trading days.
Figure 1: Top: mis-matched return pairs. Bottom: zero return problem

Figure 2: Synchronization with Data Augmentation
Figure 3: Correlation between BAC and CAT based on $RC^{5m}$, RK and BNC

Figure 4: Realized beta of BAC based on $RC^{5m}$, RK and BNC