Bayesian Nonparametric Covariance Estimation with Noisy and Nonsynchronous Asset Prices*

Jia Liu†

March 2019

Abstract

This paper proposes a Bayesian nonparametric approach to estimating 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, a new synchronization method of observations based on data augmentation is introduced. Third, the estimator is guaranteed to be positive definite. Finally, the new approach delivers exact finite sample inference without relying on asymptotic assumptions. All of those benefits lead to a more accurate estimator, which is confirmed by Monte Carlo simulation results. In real data applications, the proposed covariance estimator results in better portfolio choice outcomes.

*I am grateful for comments from John Maheu, Ronald Balvers, Peter Miu and seminar participants at the 2017 RCEA Bayesian Econometrics Workshop.

†Sobey School of Business, Saint Mary’s University, Canada, Jia.Liu@smu.ca
1 Introduction

This paper proposes a Bayesian nonparametric method of estimating the ex-post covariance matrix of asset returns, in the presence of nonsynchronous trading and market microstructure noise. The univariate Bayesian nonparametric variance estimator introduced by Griffin et al. (2016) is extended to its multivariate version to allow pooling in covariance estimation. The method delivers exact finite sample inference and the estimated covariance matrix is guaranteed to be positive definite. In addition, a new way of synchronizing observations based on data augmentation is introduced.

The covariance matrix of asset returns is the key input for many finance problems, such as portfolio allocation and asset pricing. Since the availability of high frequency data, estimation of covariance using intraday returns has become a very active area of research. An important first step is Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004). They formalized the realized covariance estimator and showed it is an asymptotic consistent estimator of the integrated covariance, under the assumption that observations are free of measurement error. However, in reality, prices are contaminated with market microstructure noise and transactions arrive nonsynchronously, which lead to poor statistical performance of the realized covariance estimator.

Several different approaches have been used to pave the way for covariation estimation of noisy and nonsynchronously spaced prices. One branch of the literature constructs the covariance estimator using synchronized return and corrects the bias. Zhang (2011) suggested the optimal sampling frequency in constructing realized covariance and proposed the two scales estimator. Griffin & Oomen (2011) formally studied the realized covariance with lead-lag adjustments. Barndorff-Nielsen et al. (2011) introduced the multivariate realized kernel based on refresh time synchronization. Lunde et al. (2016) proposed the composite realized kernel for vast covariance estimation. Aït-Sahalia et al. (2010) proposed the Quasi-maximum likelihood estimator of covariance as well as the generalized synchronization method. The cumulative covariance estimator proposed by Hayashi & Yoshida (2005) exemplifies another branch of the literature. Their estimator can be applied directly to raw observations and is unbiased under a no noise assumption. Voev & Lunde (2007) proposed a bias correction to make the cumulative covariance estimator suitable for noisy prices. Another way of constructing high frequency covariance matrix estimator is model based. Hansen et al. (2008) discussed a multivariate moving average can be used to filter out the microstructure noise. Peluso et al. (2014) introduced a Bayesian estimator of the covariance of noisy and asynchronous returns based on a parametric model. Corsi et al. (2015) proposed a covariance estimation method based on Kalman smoother and expectation maximization algorithm.

This paper proposes a Bayesian nonparametric approach to estimate the ex-post covariance matrix of a vector of asset returns. Instead of using the data independently, I exploit pooling among observations with a common covariance matrix. This is achieved with a use of Dirichlet process mixture model. The Bayesian nonparametric framework allows the number of groups or clusters of covariance matrices to vary flexibly and to be determined endogenously. To adjust for bias, I use a vector moving average model for high-frequency data and introduce pooling in this setting. From this model, a covariance estimator that corrects for market microstructure noise and nonsynchronous trading is derived based on Hansen et al. (2008).
In related work Peluso et al. (2014) introduced data augmentation based on dynamic linear model to synchronize the high-frequency prices. This paper also uses data augmentation to synchronize the data but exploits pooling to increase estimation accuracy. The proposed synchronization method is built on top of the previous-tick method defined in Hansen & Lunde (2006) but eliminates the zero-return problem caused by stale prices. Missing observations are augmented as unknown variables and are estimated conditional on observed data and model structure. With the proposed synchronization method, the Bayesian nonparametric covariance estimator with moving average adjustment fully accounts for the nonsynchronous bias.

Another advantage of the Bayesian nonparametric covariance estimator is that it is guaranteed to be positive definite. Using an inverse Wishart distribution as the prior guarantees the sampled intraday covariance is always positive definite. In addition, with synchronization based on data augmentation, the zero returns caused by stale prices are removed, which ensures non-singular matrices.

Monte Carlo simulation is conducted to compare the Bayesian nonparametric covariance estimator with realized covariance and multivariate realized kernel given nonsynchronously spaced data, with microstructure noise. The results show the proposed estimator yields lower root-mean-square-errors in estimating the covariance matrix, especially the off-diagonal elements. Empirical applications to equity data show the Bayesian 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 ones based on realized covariance or multivariate realized kernel in terms of Sharpe ratio and utility level of an investor. Moreover, the Bayesian approach provides the exact distribution of the covariance, which allows users to analyze how the optimal weights and return of a portfolio are influenced by the covariance uncertainty.

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 found 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

\[
    dP(t) = \mu(t)dt + \Pi(t)dW(t),
\]

where $\mu(t)$ is the drift term, $\Pi(t)$ is the instantaneous volatility matrix and $W(t)$ stands for a standard Brownian motion.

As the true measure of the ex-post daily covariance, the integrated covariance is the quantity of interest and is defined as

\[
    V_t = \int_{t-1}^t \Pi(\tau)\Pi(\tau)'d\tau.
\]
Given intraday return \( R_i = P_i - P_{i-1}, \ i = 1, \ldots, n \), where \( P_i = (p_i^{(1)}, \ldots, p_i^{(d)}) \) denotes the regularly spaced intraday log-price vector and \( p_i^{(j)} \) denotes the \( i^{th} \) log-price of asset \( j \), the realized covariance (RC) formalized in Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004) is defined as

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

Under no microstructure noise setting, RC converges to the integrated covariance as \( n \to \infty \). In finite samples, the summation of cross products of intraday returns provides a noisy estimator of the integrated covariance matrix and the finite sample distribution of RC is unknown and must be approximated from the asymptotic distribution derived by Barndorff-Nielsen & Shephard (2004).

In reality, due to bid-ask bounce, the discreteness of price change and measurement error, the observed prices are contaminated with microstructure noise, which turns return series to be autocorrelated. Moreover, the prices of different assets are not observed simultaneously. The returns based on nonsynchronous data have the lead-lag dependence, which makes the summation of cross products of intraday returns underestimates the covariance. This phenomenon was documented as “Epps effect” by Epps (1979). Those two challenges restricts the use of high-frequency data to form a realized covariance estimator.

A commonly-used way of mitigating the influences of microstructure noise and nonsynchronous trading is to add lead-lag adjustment or autocovariance adjustment to the realized covariance estimator. For example, the multivariate realized kernel (RK_\( t \)) proposed by Barndorff-Nielsen et al. (2011) is defined as

\[
RK = H \sum_{h=-H}^{H} \left( K\left( \frac{h}{H} \right) \sum_{i=h+1}^{n} \tilde{R}_i \tilde{R}_{i-h}' \right),
\] (4)

where \( K(\cdot) \) is the Parzen kernel function\(^1\), the optimal bandwidth \( H = c_0\bar{n}^{3/5} \) and \( \tilde{R}_i \) is based on prices with microstructure noise and is synchronized using refresh time scheme\(^2\).

### 3 Bayesian Nonparametric Covariance Estimation

Unlike existing approaches treat intraperiod returns independently, the proposed approach allows data to cluster. Returns with similar intraday covariance can be grouped flexibly under the Bayesian nonparametric framework. Mykland & Zhang (2009) considers the univariate case of holding the return variance constant over blocks of consecutive returns to help improve estimation efficiency. The proposed method considers pooling in a multivariate setting and

\[^1\text{Parzen kernel function:}\]

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

\[^2\text{In refresh time scheme, the prices is sampled at the time that all assets haven been traded. The return series is irregularly spaced.}\]
is more flexible than the blocking method in Mykland & Zhang (2009). It does not require observations to be consecutive in time and the size of cluster can be endogenously determined for each separate cluster.

### 3.1 DPM-VMA Model

The univariate Dirichlet process mixture (DPM) model used in Griffin et al. (2016) is extended to its multivariate version. The DPM is a nonparametric version of mixture model. The number of clusters in a DPM is not fixed and the number of necessary clusters can be inferred. The vector moving average parametrization is incorporated to correct the bias caused by microstructure noise and nonsynchronous trading. The DPM with vector moving average (DPM-VMA) is applied to extract covariation information from return vector \( \tilde{R}_i \).

\[
\tilde{R}_i = \mu + \Theta \eta_{i-1} + \eta_i, \quad \eta_i \sim N(0, \Sigma_i), \quad i = 1, \ldots, n, 
\]

where the mean of \( \tilde{R}_i \) has a moving average structure with the error term, \( \eta_i = \tilde{R}_i - \mu - \Theta \eta_{i-1} \) and the coefficient matrix \( \Theta \) of dimension \( d \times d \). The covariance matrix \( \Sigma_i \) is state-dependent and the data within a cluster shares the same intraday covariance.

As the distribution of \( \Sigma_i \), \( G \) is a discrete distribution with varying number of clusters. This is achieved with the help of Dirichlet process DP(\( \alpha, G_0 \)), which is set to be the prior of \( G \). A draw from DP(\( \alpha, G_0 \)) is centred around the base distribution \( G_0 \) which is an inverse Wishart distribution denoted as IW(\( \Psi, \nu \)). The base measure guarantees the positive definiteness of \( \Sigma_i \) as any draw from an inverse Wishart distribution is positive definite. The number of clusters is influenced by the precision parameter \( \alpha \) in the Dirichlet process. As \( \alpha \) increases, the number of cluster increases and the effect of pooling get diminished. The DPM-VMA model on different days are independent. \( \mu \) and \( \Theta \) are constant for \( i \) but will change with the day \( t \).

The DPM-VMA model implies \( \text{cov}(R_i, R_{i-1}) = \Theta \Sigma_{i-1} \) which could capture both the autocorrelation and the cross sectional dependence in returns. Note that the DPM-VMA model allows error terms to be heteroskedastic, so it does not require the covariance of microstructure noise to be identical.

The base function \( G_0 = \text{IW}(\Psi, \nu) \) in DP need to be calibrated day by day as the dynamics of asset volatility changes across time. Based on return observations, the scaled matrix \( \Psi \)
and the degree of freedom \( \nu \) are set to be\(^3\)
\[
\Psi = \frac{\nu - d - 1}{n},
\]
\[
\nu = \frac{1}{d} \sum_{j=1}^{d} \frac{2(\text{RC}(jj))^2}{\hat{\text{var}}(r_i^{(j)})} + d + 3,
\]
which makes the center of the base function to be the realized covariance based on low frequency data.

To add flexibility, \( \alpha \) is treated as a parameter and a hierarchical prior \( \text{Ga}(a, b) \) is placed on it. The prior \( \text{N}(0, \Lambda) \), where \( \Lambda \) is a diagonal matrix with small variance values, is used for \( \mu \). The prior of elements of \( \Theta \) is assumed to be \( \Theta^{(jk)} \sim \text{N}(0, 0.5) \) for \( j = 1, \ldots, d \) and \( k = 1, \ldots, d \).

### 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 price series is sampled as
\[
\tilde{p}_i^{(j)} = \max(p_{\text{max}(\tau_j|\tau_j \leq ih)}, j = 1, \ldots, d.
\]

Figure 1 provides one example of three assets to illustrate the mechanism. The summation of cross products of previous tick returns underestimates the daily covariance because of two problems. The first one is the trading times of assets are mismatched. As shown in Figure 2, \( r_i^{(1)} \) and \( r_i^{(2)} \) do not cover the same time interval and there exists dependence between \( r_i^{(1)} \) and \( r_{i+1}^{(2)} \). The second problem is caused by no transaction in one or multiple interval(s), which is called the zero return bias. As pointed by Kanatani & Renó (2007), if there is no zero return, one lead-lag adjustment can account for nonsynchronous bias. To control the zero return bias, more lead-lag adjustment terms are required. Another data synchronization method is the refresh time scheme proposed by Barndorf-Nielsen et al. (2011). It provides irregularly-spaced returns and the number of data synchronized using the refresh time method is limited by the least liquid asset, as shown in Lunde et al. (2016).

This paper proposes a synchronization method based on data augmentation to improve the previous-tick scheme. The missing observations on common grid points are treated as unknown variables and are estimated along with other model parameters under the Bayesian

---

\(^3\)For matrix \( X \sim \text{IW}(\Psi, \nu) \), the mean and variance of diagonal elements of \( X \) are
\[
E(X) = \frac{\Psi}{\nu - d - 1} \quad \text{and} \quad \text{var}(X^{(j)}) = \frac{2(\Psi^{(jj)})^2}{(\nu - d - 1)^2(\nu - d - 3)}.
\]

Expressing \( \Psi \) and \( \nu \) in terms of \( E(X) \) and \( \text{var}(X^{(j)}) \) yields
\[
\Psi = E(X)(\nu - d - 1) \quad \text{and} \quad \nu = \frac{2(E(X^{(j)}))^2}{\text{var}(X^{(j)})} + d + 3.
\]
Substituting \( E(X) \) and \( \text{var}(X^{(j)}) \) with \( \frac{1}{n} \text{RC} \) and \( \hat{\text{var}}(r_i^{(j)}) \) yields one estimate of \( \nu \). \( \nu \) is set to be the average of \( \nu^{(j)} \) based on all \( d \) assets.
framework. In this example shown in Figure 1, there is no transaction in interval \((i+2, i+3]\) for both asset 1 and 2 and in interval \((i, i+1]\) for asset 3. In other words, \(p_{i+3}^{(1)}, p_{i+3}^{(2)}\) and \(p_{i+1}^{(3)}\) are missing and need to be augmented. Using observed prices and knowing the model structure, the missing prices can be inferred. Compared with the previous-tick method, the proposed approach eliminates the zero return by filling the missing observation gaps. The return series only have first order lead and lag dependence, which can be filtered out through the first order vector moving average process illustrated in Section 3.1.

As the previous-tick method, the proposed synchronization method requires setting a common-time grid. There is a tradeoff between the length of grid and the quality of data augmentation. The larger the grid length, the fewer missing prices need to be augmented. But that leads to more information loss as the sampling frequency is lower. Conditional on data, increasing grid frequency results in diminished value in data augmentation since more grid points contain missing observations and data augmentation is necessary at more points. Based on Monte carlo simulation results, setting \(2D\) as the grid length minimizes error in estimating both diagonal and off-diagonal elements of the covariance, where \(D\) denotes the average of price change duration of all assets.

### 3.3 Model Estimation

The model is estimated using Markov chain Monte Carlo (MCMC) techniques. I apply the slice sampler of Kalli et al. (2011) to the stick-breaking representation of the DPM. Expressing the DP prior as the stick-breaking representation by Sethuraman (1994), the DPM model can be written as

\[
p(\tilde{R}_i, \mu, \{\Phi_j\}_{j=1}^\infty, \{w_j\}_{j=1}^\infty) = \sum_{j=1}^\infty w_j N(\tilde{R}_i|\mu, \Phi_j),
\]

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

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

In the slice sampling, a set of auxiliary variables \(u_{1:n} = \{u_1, \ldots, u_n\}\) is introduced to slice the infinite state space to a finite one so that the sampling of model parameters is feasible. \(u_{1:n}\) is sampled along with other parameters and randomly truncates the state space to \(K = \sum_{j=1}^\infty \mathbb{1}(u_i < w_j)\) at each MCMC iteration. The joint distribution of \(\tilde{R}_i\) and \(u_i\) is given by

\[
f \left( \tilde{R}_i, u_i | \mu, \{\Phi_j\}_{j=1}^\infty, \{w_j\}_{j=1}^\infty \right) = \sum_{j=1}^\infty \mathbb{1}(u_i < w_j) N \left( \tilde{R}_i | \mu, \Phi_j \right) .
\]

The original model (10) is recovered by integrating out \(u_i\).

Next, introduce a set of latent state variables \(s_{1:n} = \{s_1, \ldots, s_n\}\) that label each observation’s cluster. Given \(s_i \in \{1, 2, \ldots K\}\), then \(\Sigma_i = \Phi_{s_i}\). Note that the number of clusters \(K\) is adjusted over MCMC iterations. A new cluster with covariance \(\Phi_{K+1} \sim \text{IW}(\Psi, \nu)\) can be opened and clusters with similar covariances can be merged.
Because of the moving average structure, the Gibbs sampler is not feasible to sample $\mu$ and $\Theta$. Sampling high-dimensional parameters such as $\mu$ and $\Theta$ using Metropolis-Hasting results in very low mixing and finding good proposals is also challenging. I apply the Hamiltonian Monte Carlo method by Neal (2011) to sample $\mu$ and $\Theta$ as blocks. The Hamiltonian Monte Carlo adopts the Hamilton dynamics, rather than a probability distribution, to propose draws in Markov chain. Unlike the random walk proposal, the Hamilton dynamics produces distant proposals which explore the target distribution more efficiently.

The Gibbs sampler handles the estimation of $\Phi_j$, which represents the unique values of $\Sigma_i$. The concentration parameter $\alpha$ is sampled using the method in Escobar & West (1994). Each MCMC run contains the following steps. The estimation contains the following steps and details can be found in Appendix 7.1.

1. Sample $\mu | \tilde{R}_{1:n}, \Phi_{1:K}, \Theta, s_{1:n}$.
2. Sample $\Theta | \tilde{R}_{1:n}, \mu, \Phi_{1:K}, s_{1:n}$.
3. Sample $\Phi_j | \tilde{R}_{1:n}, \mu, \Theta, s_{1:n}$ 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. Sample $s_i | \tilde{R}_{1:n}, s_{1:n}, \mu, \Theta, \Phi_{1:K}, u_{1:n}, K$ for $i = 1, \ldots, n$.
7. Sample $\alpha | K$.
8. Sample missing prices and update return vector.

### 3.4 Covariance Matrix Estimator

In the Bayesian nonparametric framework, the estimator of $V_t$ is the posterior mean of covariance matrix. Hansen et al. (2008) pointed that the covariance estimator based on moving average model requires an adjustment in order to obtain an unbiased estimator. Incorporating their adjustment, the ex-post covariance $V_t$ is given as

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

(13)

where $I$ is the identity matrix.

Integrating out all parameter and distributional uncertainty and using $M$ MCMC outputs, $E[V_t | \tilde{R}_{1:n}]$ can be estimated. The Bayesian nonparametric covariance estimator (BNC) is defined as

$$\text{BNC} = \frac{1}{M} \sum_{m=1}^{M} (I + \Theta^{(m)}) \left( \sum_{i=1}^{n} \Sigma_i^{(m)} \right) (I + \Theta^{(m)})'$$

(14)

$$= \frac{1}{M} \sum_{m=1}^{M} (I + \Theta^{(m)}) \left( \sum_{i=1}^{n} \Phi_{s_i}^{(m)} \right) (I + \Theta^{(m)})'.$$
Appendix 7.2 proves the Bayesian estimator provided in equation (14) correctly recovers the ex-post covariance in the presence of independent microstructure noise and nonsynchronous trading, if assuming no zero-return bias.

The posterior distributions of any functions of the covariance matrix, such as realized beta or correlation, are readily available from MCMC outputs. For instance, a $(1 - \alpha)\%$ probability density interval for the covariance between asset $j$ and $k$ is interval between the $\alpha/2\%$ and $(1 - \alpha/2\%)$ quantiles of $\sum_{i=1}^{n} \Sigma_{i}^{(jk)}$. Note that the exact finite sample estimates can be obtained directly, while the classical estimator relies on asymptotic distribution to derive the confidence intervals.

The proposed Bayesian nonparametric covariance estimator is guaranteed to be positive definite, while classical approaches deliver positive semi-definite results. In the estimation of large covariance matrices, it is possible that the number of available observations is lower than the number of assets. For example, the number of data points synchronized using refresh time depends on the most inactive asset and can be driven to a number below data dimension. The previous-tick approach can increase the number of observations by shrinking grid length, but it may result in many zero returns. In those cases, the covariance matrix calculated using traditional methods can be singular and not positive definite. The proposed Bayesian nonparametric approach can solve the difficulty. As the missing price gaps can all be filled using data augmentation, the grid length can be adjusted to make the number of non-zero return vectors above the data dimension. In addition, using an inverse Wishart distribution as the base distribution of intraday covariance ensures the positive definiteness of the covariance estimator.

4 Simulation Results

Following Barndorff-Nielsen et al. (2011), the fundamental log prices are generated from a multivariate factor stochastic volatility model of the form

$$dp^{(j)} = \mu^{(j)} dt + \rho^{(j)} \sigma^{(j)} dB^{(j)} + \sqrt{1 - \rho^{(j)}^2} \sigma^{(j)} dW,$$

$$\sigma^{(j)} = \exp(\beta_0^{(j)} + \beta_1^{(j)} v^{(j)}),$$

$$dv^{(j)} = \alpha^{(j)} v^{(j)} dt + dB^{(j)},$$

where $W$ and $B^{(j)}$ are standard Brownian motions, $\text{cor}(dW, dB^{(j)}) = 0$ and the values of parameters are $(\mu^{(j)}, \beta_0^{(j)}, \beta_1^{(j)}, \alpha^{(j)}, \rho^{(j)}) = (0.04, -0.3125, 0.125, -0.025, -0.3)$ for $j = 1, \ldots, d$.

The arrival times of observed prices are simulated from independent Poisson process. Parameter $\lambda$ in Poisson process governs the trading frequency of simulated data. For example, $(\lambda^{(1)}, \lambda^{(2)}) = (5, 8)$ means the transactions of asset 1 and 2 arrives every 5 seconds and 8 seconds on average, respectively.

Following Barndorff-Nielsen et al. (2011), the prices with microstructure noise are simulated by adding the following error term.

$$\hat{p}_{t,l}^{(j)} = p_{t,l}^{(j)} + \epsilon_{t,l}^{(j)}, \quad \epsilon_{t,l}^{(j)} \sim \text{N}(0, \sigma_{\omega}^{(j)^2}), \quad l = 1, \ldots, N_t,$$

$$\sigma_{\omega}^{(j)^2} = \xi^2 \sqrt{\frac{1}{N_t} \sum_{l=1}^{N_t} (\sigma_{t,l}^{(j)})^4},$$
where $\xi^2$ stands for the noise-signal ratio which governs the size of microstructure noise. The error term is assumed to be independent of each other and the variance of error increases with the data volatility.

The estimation target is the daily ex-post covariance matrix $V_t = \sum_{l=1}^{N_t} \Sigma_{t,l}$, where $\Sigma_{t,l}^{(j,k)} = \sqrt{1 - \rho^{(j)} \sigma_{t,l}^{(j)}} \sqrt{1 - \rho^{(k)} \sigma_{t,l}^{(k)}}$. To reduce the influence of nonsynchronous trading and microstructure noise, RC is formed using 5-minute return synchronized by previous tick. RK uses returns based on refresh time synchronization. The synchronization with data augmentation is used for BNC and the grid length is set to be 2 times the average duration of price changes\(^4\).

RC, RK and BNC are compared in different dimension, microstructure noise circumstances. The norm of RMSE matrix, average RMSE in estimating both diagonal and off-diagonal elements are reported in Table 1. Both three assets case and ten assets case are considered. Panel A and B shows the results in two cases with different microstructure noise levels ($\xi^2 = 0.001$ and $\xi^2 = 0.003$). Four frequencies ranging from the low frequency case with $\lambda^{(j)} \in \{120, 80, 90\}$ to the high frequency case with $\lambda^{(j)} \in \{5, 6, 8\}$ are considered.

5-minute RC always has higher RMSE compared with the other two estimators since the small number of observations (78 per day) makes estimation less precise. RK and BNC are benefited from high frequency data and RMSEs get lowers as the data frequency increases. The top performer is the BNC estimator, which yields the lowest RMSE norm and has the lowest error in estimating covariance elements in all the 16 cases. For instance, in the highest frequency case ($\lambda^{(j)} \in \{5, 6, 8\}$) with 10 assets, the average RMSEs of diagonal and off-diagonal elements of BNC are 0.0820 and 0.0541, while the values are 0.1109 and 0.0739 if RK serves as the estimator. The improvement is greater than 20%.

5 Empirical Applications

The tick prices and national best bid and offer (NBBO) prices of 10 equities (stock symbol: BAC, CAT, DD, F, GIS, JNJ, KO, T, WMT, XOM) and the Standard & Poor’s Depository 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 data. The sample contains data on 503 trading days.

The 5-minute realized covariance, multivariate realized kernel and the Bayesian nonparametric covariance estimator based on DPM-VMA model are applied to estimate the daily covariance matrix of returns of the ten assets. The synchronization methods for RC, RK and BNC are previous tick with 5-minute interval, refresh-time and the proposed synchronization with data augmentation. Because of the discreteness of transaction prices, real data contain more zero returns than simulated data. The grid length in proposed synchronization is set to be $4\bar{D}_t$, where $\bar{D}_t$ is the average duration of price changes of $d$ assets on day $t$.

\(^4\)If the calculated length is not the divisor of 23400, then round it to the smallest number that makes 23400 divisible.
5.1 Covariance Matrix, Correlation and Realized Beta

The sample covariance matrix calculated using open-to-close daily returns is not influenced by microstructure noise and nonsynchronous trading and provide a benchmark to assess the unbiasedness of the three estimator. Table 2 reports the open-to-close daily covariance, averages of RC$^{5m}$, RK and BNC over the sample period. BNC is more close to the sample covariance of daily returns, compared with RC$^{5m}$ and RK, which can be confirmed by the average bias reported in the last column of Table 2. For example, ...

Figure 3 provides an example of the dynamic correlation series between BAC and CAT based on RC$^{5m}$, RK and BNC. Overall, the three versions of correlation estimates share similar dynamics. The correlations based on RK and BNC are close to each other and the correlation implied by 5-minute RC seems more volatile.

Given the covariance matrix between asset returns and market returns, the realized beta is defined as $\beta_t = \frac{\text{Cov}_{ij}^t}{\text{Cov}_{jj}^t}$, where 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. $\beta_{t}^{RK}$ and $\beta_{t}^{BNC}$ have very similar path, while the realized beta estimates based on RC$^{5m}$ is larger and more volatile than the other two. Table 3 shows the fitted values of ARMA(1,1) model for the three versions of realized beta. The estimation result confirms the strong persistency of realized beta. Comparison of parameter estimates implies $\beta_{t}^{RK}$ and $\beta_{t}^{BNC}$ share similar time series dynamics.

5.2 Portfolio Allocation Evaluation

It is worthy to explore whether the proposed covariance matrix estimator improves portfolio allocation. Following Fleming et al. (2003), the evaluation is based on the performance of minimum-variance portfolios formed using the BNC estimator, along with RC$^{5m}$ and RK. Suppose an investor would like to hold a portfolio formed by ten stocks (BAC, CAT, DD, F, GIS JNJ, KO, T, WMT and XOM) and a risk-free asset from July 2, 2014 to June 29, 2016. She applies the volatility-timing strategy to adjust the portfolio weights each day by solving the following risk minimization problem given a desired portfolio return $\mu_0$.

Min $w_t'\hat{\Sigma}_t w_t$ s.t. $w_t' \mu = \mu_0$ and $w_t' 1 = 1,$ \hspace{1cm} (20)

where $w_t$ stands for portfolio weights on day $t$, $\hat{\Sigma}_t$ is the covariance matrix, $\mu$ is the daily return mean of assets and $\mu_0$ is the required return of portfolio. The solution of the minimization problem is

$$w_t = \frac{\hat{\Sigma}_t^{-1} \mu}{\mu' \hat{\Sigma}_t^{-1} \mu} \mu_0.$$ \hspace{1cm} (21)

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

$$\hat{\Sigma}_t = \exp(-\kappa)\hat{\Sigma}_{t-1} + \kappa \exp(-\kappa)\hat{S}_{t-1},$$ \hspace{1cm} (22)

where $\kappa$ is the decay rate and $\hat{S}_{t-1}$ is the ex-post covariance estimator, which can be RC$^{5m}$, RK or BNC. The only difference between each portfolio is the estimates of $\hat{S}_{t-1}$ used. The
initial covariance matrix \( \hat{S}_0 \) are set to be \( \text{RC}_0^{5m} \). The return mean \( \mu \) is assumed to be a constant and set to be the sample mean.

If the summation of \( \mathbf{w}_t \) does not equal to 100\%, the remaining portion of capital will be allocated to risk-free asset, so that the portfolio return on day \( t \) equals to \( r_p^t = \mathbf{w}_t R_t + (1 - \mathbf{w}_t^t \mathbf{1}) r_f \). Table 4 shows the return mean, variance and Sharpe ratio of portfolios based on RC, RK and BNC under three decay rates (\( \kappa = 0.03, 0.06 \) and 0.09) and three annual required rate of returns (10\%, 20\% and 30\%). The portfolio based on BNC estimator has the highest Sharpe ratio in all of the 9 cases. For example, given \( \kappa = 0.06 \) and 20\% required rate of return, the Sharpe ratio of BNC portfolio is 0.319, while portfolios based on \( \text{RC}_0^{5m} \) and RK are 0.314 and 0.312, respectively.

A utility-based approach is used to assess economic gains. The quadratic utility function used Fleming et al. (2003) is adapted.

\[
U(r_p^t) = W_0 \left[ (1 + r_p^t + r_f^t) - \frac{\gamma}{2(1 + \gamma)} (1 + r_f^t + r_p^t)^2 \right],
\]

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 of two competing strategies can be evaluated through calculating the performance fee \( \Delta \) that an investor would pay to switch from one portfolio to another. \( \Delta \) is a constant that satisfies the following equation

\[
\sum_{t=1}^{T} U(r_p^1) = \sum_{t=1}^{T} U(r_p^2 - \Delta).
\]

The daily return of portfolio based on 5-minute RC is set to be \( r_p^1 \), which is the benchmark portfolio. \( r_p^2 \) is the portfolio return based on RK or BNC.

Table 5 also lists the annualized basis point fees that an investor with quadratic utility would like to pay to switch from 5-minute RC based strategy to portfolio using RK or BNC. Given 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 fee in order to choose portfolio based on BNC, instead of the RK based one. For example, in the case with \( \kappa = 0.06 \) and \( \mu_0 = 20\% \), an investor with \( \gamma = 10 \) would like to pay over 30 extra basis points in order to choose BNC, rather than \( \text{RC}_0^{5m} \) as covariance estimator.

Furthermore, the Bayesian approach provides a method to assess how the estimation uncertainty influences the optimal weights, thereby the return of portfolio. From a Bayesian point of view, covariance matrix estimators provide only point estimates about the center of the covariance distribution. However, the realization of parameter can deviate from the distribution mean. In the equation (22), the uncertainty in ex-post covariance estimation \( (\hat{S}_0, \ldots, \hat{S}_{t-1}) \) leads to different forecasted value of \( \hat{\Sigma}_t \). Since the Bayesian model delivers an exact finite sample distribution of covariance, all the possible outcomes \( \{(\hat{S}_{0}^{(i)}, \ldots, \hat{S}_{t-1}^{(i)})\}_{i=1}^{G} \) and \( \{\hat{\Sigma}_t^{(i)}\}_{i=1}^{G} \) can be obtained. Figure 5 and Figure 6 provide histograms of portfolio returns and the weights of the ten assets on July 31, 2014. Due to the uncertainty in covariance estimation, the weight of T-mobile stock in the portfolio varies from 20\% to 30\% and the possible portfolio returns lies in the range of 0.3\% and 0.47\%. Classical approaches can only provide the asymptotic distribution, which is an approximation in finite sample.
6 Conclusion

This paper proposes a Bayesian nonparametric method of estimating covariance matrix for nonsynchronous price contaminated with independent microstructure noise. The proposed estimator provides at least four benefits in covariance estimation. First, pooling observations with similar covariance increases the precision of ex-post covariance estimation. Second, the Bayesian approach delivers exact finite sample results without relying on any infill asymptotic assumption. Third, the estimated covariance estimator is guaranteed to be positive definite. Last but not least, a new synchronization method with data augmentation is introduced to convert nonsynchronous observations to regularly spaced data series without zero-return problem.

Monte Carlo simulation confirms the Bayesian nonparametric covariance estimator is very competitive with existing estimators. Empirical application to equity returns shows the correlation and realized beta implied on Bayesian nonparametric covariance estimator have similar time series dynamics as the multivariate realized kernel. The minimum variance portfolio based on proposed estimator outperforms the portfolio formed using realized covariance and multivariate realized kernel in terms of Sharpe ratio and utility level.
References


Kanatani, T. & Renó, R. (2007), Unbiased covariance estimation with interpolated data, Department of economics university of siena, Department of Economics, University of Siena.


7 Appendix

7.1 Estimation Steps of DPM-VMA Model

1. Sampling $\mu$:
   Given prior: $\mu \sim N(M_\mu, V_\mu)$. The posterior of $\mu$ is
   \[
   p(\mu | \tilde{R}_{1:n}, \Theta, \Phi_{1:K}) \propto p(\mu) \prod_{i=1}^{n} p(\tilde{R}_i | \mu + \Theta \eta_{i-1}, \Phi_{s_i})
   \]
   \[
   \propto \exp \left\{ -\frac{1}{2} \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} + V_\mu^{-1} \right) \mu - \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} (\tilde{R}_i - \Theta \eta_{i-1}) + V_\mu^{-1} M_\mu \right) \right\}. \tag{25}
   \]
   where $\eta_{i-1} = \tilde{R}_{i-1} - \mu - \Theta \eta_{i-2}$.
   Define $U_1(\mu) = -\log \left[ p(\mu | \tilde{R}_{1:n}, \Theta, \Phi_{1:K}) \right]$, we have
   \[
   U_1(\mu) = \frac{1}{2} \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} + V_\mu^{-1} \right) \mu - \mu' \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} (\tilde{R}_i - \Theta \eta_{i-1}) + V_\mu^{-1} M_\mu \right). \tag{26}
   \]
   \[
   \frac{\partial U_1(\mu)}{\partial \mu} = \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} + V_\mu^{-1} \right) \mu - \left( \sum_{i=1}^{n} \Phi_{s_i}^{-1} (\tilde{R}_i - \Theta \eta_{i-1}) + V_\mu^{-1} M_\mu \right). \tag{27}
   \]
   Introduce an auxiliary $d$-dimensional vector $B_1$ and define function $K_1(B_1) = \sum_{j=1}^{d} \frac{(B_j^{(j)})^2}{2}$.
   
   The leapfrog method is used to approximate the Hamiltonian dynamics.
   
   (1) Set $\mu_0 = \mu^{(m-1)}$, where $\mu^{(m-1)}$ is the value of $\mu$ in previous iteration. Initialize $B_1$ as $B_1^{(j)} \sim N(0, 1)$.
   (2) $B_1' = B_1 - \frac{\epsilon}{2} \frac{\partial U_1(\mu_0)}{\partial \mu}$ and $\mu' = \mu_0$.
   (3) For $l$ from 1 to $L$,
      a. $\mu' = \mu' + \epsilon B_1'$.
      b. If $l < L$, $B_1' = B_1' - \epsilon \frac{\partial U_1(\mu')}{\partial \mu}$.
      c. If $l = L$, $B_1' = B_1' - \frac{\epsilon}{2} \frac{\partial U_1(\mu')}{\partial \mu}$.

   Update $\mu = \mu'$ with acceptance rate $\min(1, \exp(U_1(\mu_0) + K_1(B_1) - U_1(\mu') - K_1(B_1')))$.
   
   $L$ is the leapfrog step and $\epsilon$ is the stepsize. I set $L = 20$ and adjust $\epsilon$ very 10 MCMC iterations. If the average acceptance rate in previous 10 runs is zero, set $\epsilon = 0.9 \epsilon$. If the average acceptance rate is above 0.8, adjust $\epsilon = 1.1 \epsilon$.

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

$$p(\Theta|\tilde{R}_{1:n}, \Theta, \Phi_{1:K}) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left[ \eta_{i-1}' \Theta' \Phi^{-1}_{s_i} \left( \Theta \eta_{i-1} - 2(\tilde{R}_i - \mu) \right) \right] \right\}$$

$$\cdot \prod_{j=1}^{d} \prod_{k=1}^{d} \exp \left[ -\frac{(\Theta(jk) - m_{jk})^2}{2v_{jk}^2} \right]$$

(28)

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

$$U_2(\Theta) = -\frac{1}{2} \sum_{i=1}^{n} \left[ \eta_{i-1}' \Theta' \Phi^{-1}_{s_i} \left( 2(\tilde{R}_i - \mu) - \Theta \eta_{i-1} \right) \right] + \sum_{j=1}^{d} \sum_{k=1}^{d} \left[ \frac{(\Theta(jk) - m_{jk})^2}{2v_{jk}^2} \right]$$

(29)

$$\frac{\partial U_2(\Theta)}{\partial \Theta} = -\sum_{i=1}^{n} \Phi^{-1}_{s_i} \left( \tilde{R}_i - \mu - \Theta \eta_{i-1} \right) \eta_{i-1}' + \frac{(\Theta(jk) - m_{jk})}{v_{jk}^2} \mathbb{I}_{jk}$$

(30)

Introduce an auxiliary $d \times d$ matrix $B_2$ and define function $K_2(B_2) = \sum_{j=1}^{d} \sum_{k=1}^{d} \frac{(B_{2(jk)})^2}{2}$.

(1) Set $\Theta_0 = \Theta^{(m-1)}$, where $\Theta^{(m-1)}$ is the result from previous iteration. Initialize $B_2$ that each element is drawn from $N(0, 1)$.

(2) $B'_2 = B_2 - \epsilon \frac{\partial U_2(\Theta_0)}{\partial \Theta}$ and $\Theta' = \Theta_0$.

(3) For $l$ from 1 to $L$,

a. $\Theta' = \Theta' + \epsilon B'_2$.

b. If $l < L$, $B'_2 = B'_2 - \epsilon \frac{\partial U_2(\Theta')}{\partial \Theta}$.

c. If $l = L$, $B'_2 = B'_2 - \frac{\epsilon}{2} \frac{\partial U_2(\Theta')}{\partial \Theta}$.

Accept $\Theta'$ with acceptance rate $\min(1, \exp(U_2(\Theta_0) + K_2(B_2) - U_2(\Theta') - K_2(B'_2)))$.

3. Sampling $\Phi_j$ for $j = 1, \ldots, K$:

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

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

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

(31)

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

4. Sampling $s_i$ for $i = 1, \ldots, n$:

$$P(s_i = j|R_i, \mu, \Phi_{1:K}, w_{1:K}, u_i) \propto \sum_{j=1}^{K} \mathbb{I}(w_j > u_i) N \left( \tilde{R}_i|\mu + \Theta \eta_{i-1}, \Phi_j \right).$$

(32)
5. Sampling $v_j$ and calculate $w_j$ for $j = 1, \ldots, K$:

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

(33)

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

6. Sampling $u_i$ for $i = 1, \ldots, n$:

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

(34)

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

8. Sampling $\alpha$:

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

$$p(\alpha \mid K) \sim q \cdot \text{Ga}(a + K, b - \log \xi) + (1 - q) \cdot \text{Ga}(a + K - 1, b - \log \xi),$$

(35)

where $q = \frac{a + K - 1}{a + K - 1 + n(b - \log \xi)}$ and $\xi \sim \text{Beta}(\alpha + 1, n)$.

7.2 Proof of the Unbiasedness of the BNC Estimator

The regularly spaced return vector of the $j$ assets is denoted as $\tilde{R}_i = \tilde{P}_i - \tilde{P}_{i-1}$, where $\tilde{P}_i = \left( \tilde{p}_i^{(1)}, \tilde{p}_i^{(2)}, \ldots, \tilde{p}_i^{(d)} \right)'$ represents price with noise.

Assuming there is no zero-return bias, the variance and first autocovariance of $\tilde{R}_i$ are

$$\text{cov}(\tilde{R}_i) = V_i + \Omega_{i-1} + \Omega_i - \Gamma_{i-1} - \Gamma_i = V_i + \Xi_{i-1} + \Xi_i,$$

$$\text{cov}(\tilde{R}_i, \tilde{R}_{i-1}) = -\Gamma_{i-1} + \Omega_{i-1} = -\Xi_{i-1}.$$

(36)  (37)

where $\Omega_i = \text{diag} \left( (\omega_i^{(1)})^2, (\omega_i^{(2)})^2, \ldots, (\omega_i^{(d)})^2 \right)$, $\Gamma_i$ is matrix with zero diagonals and measures the lead-lag dependence between every two assets and $\Xi_i = \Omega_i - \Gamma_i$.

Consider the following heteroskedastic VMA model for $\tilde{R}_i$,

$$\tilde{R}_i = \mu + \Theta \eta_{i-1} + \eta_i, \quad \eta_i \sim N(0, \Sigma_i),$$

(38)

which will be used to recover an estimate of ex-post variance for the fundamental return process, $V = \sum_{i=1}^{n} V_i$.

The corresponding moments of this process are

$$\text{cov}(\tilde{R}_i) = \Theta \Sigma_{i-1} \Theta' + \Sigma_i,$$

$$\text{cov}(\tilde{R}_i, \tilde{R}_{i-1}) = \Theta \Sigma_{i-1}.$$

(39)  (40)
Equating (37) and (40), we have
\[- \Xi_{i-1} = \Theta \Sigma_{i-1} \quad \text{and} \quad - \Xi_i = \Theta \Sigma_i. \tag{41}\]

Equating (36) and (39) and using the result in (41), we have
\[V_i + \Xi_i + \Xi_{i-1} = \Theta \Sigma_{i-1} \Theta' + \Sigma_i = - \Xi_{i-1} \Theta' - \Theta^{-1} \Xi_i\]
\[V_i = -(I + \Theta') \Xi_{i-1} - (I + \Theta^{-1}) \Xi_i. \tag{42}\]

Using the results in (42) and (41), the summation of \(V_i\) and \(\Sigma_i\), over \(i = 1, \ldots, n\), are
\[\sum_{i=1}^{n} V_i = -(I + \Theta') \sum_{i=1}^{n} \Xi_{i-1} - (I + \Theta^{-1}) \sum_{i=1}^{n} \Xi_i, \tag{43}\]
\[\sum_{i=1}^{n} \Sigma_i = - \Theta^{-1} \sum_{i=1}^{n} \Xi_i. \tag{44}\]

The ratio between (43) and (44) is
\[\frac{\sum_{i=1}^{n} V_i}{\sum_{i=1}^{n} \Sigma_i} = (I + \Theta) \sum_{i=1}^{n} \Xi_{i-1} (I + \Theta^{-1}) \sum_{i=1}^{n} \Xi_i (I - \Theta)^{-1} \Theta - \Theta (I - \Theta) \Theta = (I + \Theta) (I + \Theta)^{'} \tag{45}\]

Finally, we have
\[\sum_{i=1}^{n} V_i = (I + \Theta) \sum_{i=1}^{n} \Sigma_i (I + \Theta), \quad \text{if} \quad \Xi_n = \Xi_0. \tag{46}\]

### 7.3 Estimation Steps of Missing Observations

If \(b\) out of \(d\) elements in price vector \(P_{i,t}\) are missing, where \(1 \leq b \leq d\), the sampling of missing price records are conditional on the \(q = d - b\) observed prices, the adjacent prices \(P_{i-1}\) and \(P_{i+1}\), the mean vector \(\mu\) and the covariance matrices \(\Sigma_{i-1}\) and \(\Sigma_{i+1}\). The return vectors \(R_i\) and \(R_{i+1}\) provide the linkage between the model and the missing observations. First splitting \(P_i, P_{i+1}, P_{i+2}, R_i, R_{i+1}, \mu, \Sigma_i\) and \(\Sigma_{i+1}\) into two groups, one corresponds to the \(b\) missing observations in \(P_i\), the other matches to the \(q\) observed prices.

\[P_{i-1} = \begin{bmatrix} P_{i-1}^b \\ P_{i-1}^q \end{bmatrix}, \quad P_i = \begin{bmatrix} P_i^b \\ P_i^q \end{bmatrix}, \quad P_{i+1} = \begin{bmatrix} P_{i+1}^b \\ P_{i+1}^q \end{bmatrix}, \quad R_i = \begin{bmatrix} R_i^b \\ R_i^q \end{bmatrix}, \quad R_{i+1} = \begin{bmatrix} R_{i+1}^b \\ R_{i+1}^q \end{bmatrix}\]
\[\mu = \begin{bmatrix} \mu^b \\ \mu^q \end{bmatrix}, \quad \Sigma_i = \begin{bmatrix} \Sigma_{i}^{bb} & \Sigma_{i}^{bq} \\ \Sigma_{i}^{qb} & \Sigma_{i}^{qq} \end{bmatrix} \quad \text{and} \quad \Sigma_{i+1} = \begin{bmatrix} \Sigma_{i+1}^{bb} & \Sigma_{i+1}^{bq} \\ \Sigma_{i+1}^{qb} & \Sigma_{i+1}^{qq} \end{bmatrix}.\]

The conditional distribution of \(R_i^b\) and \(R_{i+1}^b\) given observed \(R_i^q\) and \(R_{i+1}^q\) are
\[R_i^b | R_i^q = P_i^b - P_{i-1}^b | R_i^q \sim N \left( \mu_i, \Sigma_i \right), \tag{47}\]
\[R_{i+1}^b | R_{i+1}^q = P_{i+1}^b - P_i^b | R_{i+1}^q \sim N \left( \mu_{i+1}, \Sigma_{i+1} \right), \tag{48}\]
where \( \overline{\mu}_i \) and \( \Sigma_i \) are the mean and covariance of distribution of \( R^b_i \) conditional on \( R^q_i \).

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

\[
\overline{\mu}_i = \mu^b + (\Theta \eta_{i-1})^b + \Sigma_i^b (\Sigma_i^{qq})^{-1} (R^q_i - \mu^q - (\Theta \eta_{i-1})^q), \tag{49}
\]

\[
\Sigma_i = \Sigma_i^b - \Sigma_i^b (\Sigma_i^{qq})^{-1} (\Sigma_i^b)', \tag{50}
\]

\( \overline{\mu}_{i+1} \) and \( \Sigma_{i+1} \) can be derived similarly.

If the prices of \( d \) assets are all missing, then \( \overline{\mu}_i = \mu_i \) for DPM model, \( \overline{\mu}_i = \mu + \Theta \eta_{i-1} \) for DPM-VMA model and \( \Sigma_i = \Sigma_i \).

The density of \( P^b_i \) conditional on observed prices and model parameters is given as

\[
\pi \left( P^b_i \mid \cdots \right) \propto \exp \left\{ -\frac{1}{2} \left[ P^b_{t,i} \Sigma_i^{-1} P^b_i - 2 P^b_{t,i} \Sigma_i^{-1} P^b_{i-1} + \overline{\mu}_i \right] -\frac{1}{2} \left[ P^b_{i+1} \Sigma_i^{-1} P^b_i - 2 P^b_{i+1} \Sigma_i^{-1} (P^b_{i+1} - \overline{\mu}_{i+1}) \right] \right\} \tag{51}
\]

\(~ N(M^b, V^b), \)

where

\[
M^b = V^b \left[ \Sigma_i^{-1} (P^b_{i-1} + \overline{\mu}_i) + \Sigma_{i+1}^{-1} (P^b_{i+1} - \overline{\mu}_{i+1}) \right], \tag{52}
\]

\[
V^b = \left( \Sigma_i^{-1} + \Sigma_{i+1}^{-1} \right)^{-1}. \tag{53}
\]
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>$\xi^2 = 0.001$</th>
<th>$\xi^2 = 0.003$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Assets</td>
<td>(80, 90, 120)</td>
<td>0.6113 0.4465 0.4388</td>
<td>0.9435 0.4836 0.4826</td>
</tr>
<tr>
<td></td>
<td>(30, 40, 60)</td>
<td>0.5150 0.4044 0.3562</td>
<td>0.9549 0.4600 0.4144</td>
</tr>
<tr>
<td></td>
<td>(10, 15, 20)</td>
<td>0.4112 0.2663 0.2212</td>
<td>0.8317 0.3244 0.2754</td>
</tr>
<tr>
<td></td>
<td>(5, 6, 8)</td>
<td>0.4853 0.2658 0.2020</td>
<td>1.4337 0.9244 0.7377</td>
</tr>
<tr>
<td>10 Assets</td>
<td>(80, 90, 120)</td>
<td>2.6436 2.1159 1.9344</td>
<td>0.9073 1.0384 1.0326</td>
</tr>
<tr>
<td></td>
<td>(30, 40, 60)</td>
<td>1.8699 1.6864 1.3644</td>
<td>0.9043 0.8668 0.7725</td>
</tr>
<tr>
<td></td>
<td>(10, 15, 20)</td>
<td>1.4427 1.1661 0.9173</td>
<td>0.7818 0.5819 0.4897</td>
</tr>
<tr>
<td></td>
<td>(5, 6, 8)</td>
<td>1.4337 0.9244 0.7377</td>
<td>0.8061 0.4572 0.3538</td>
</tr>
</tbody>
</table>

Results are based on 2,000 days’ simulation. $||\text{RMSE}|| = \frac{1}{T-t_0} \sum_{t=t_0}^{T} ||\hat{\text{Cov}}_t - \text{Cov}_t||$, 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 5,000 MCMC runs, after 10,000 burn-in.
Table 2: Summary of Global Minimum Variance Portfolio Performance

<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>
</tbody>
</table>

| **Average of 5-minute Realized Covariance** |     |     |    |     |     |     |     |     |     |     | 0.5134  | 0.0530 |
| BAC            | 2.026 |     |    |     |     |     |     |     |     |     |         |      |
| CAT            | 0.830 | 1.857 |     |     |     |     |     |     |     |     |         |      |
| DD             | 0.652 | 0.669 | 1.412 |     |     |     |     |     |     |     |         |      |
| F              | 0.907 | 0.827 | 0.610 | 2.175 |     |     |     |     |     |     |         |      |
| GIS            | 0.327 | 0.248 | 0.290 | 0.342 | 0.819 |     |     |     |     |     |         |      |
| JNJ            | 0.490 | 0.340 | 0.372 | 0.558 | 0.399 | 1.238 |     |     |     |     |         |      |
| KO             | 0.325 | 0.251 | 0.301 | 0.299 | 0.422 | 0.351 | 0.799 |     |     |     |         |      |
| T              | 0.365 | 0.330 | 0.303 | 0.342 | 0.272 | 0.289 | 0.274 | 0.796 |     |     |         |      |
| WMT            | 0.393 | 0.311 | 0.312 | 0.364 | 0.335 | 0.358 | 0.339 | 0.268 | 1.000 |     |         |      |
| XOM            | 0.668 | 0.725 | 0.557 | 0.616 | 0.293 | 0.383 | 0.311 | 0.359 | 0.331 | 1.407 |         |      |

| **Average of Multivariate Realized Kernel** |     |     |    |     |     |     |     |     |     |     | 0.4807  | 0.0203 |
| BAC            | 2.000 |     |    |     |     |     |     |     |     |     |         |      |
| CAT            | 0.739 | 1.839 |     |     |     |     |     |     |     |     |         |      |
| DD             | 0.626 | 0.633 | 1.409 |     |     |     |     |     |     |     |         |      |
| F              | 0.784 | 0.727 | 0.547 | 2.168 |     |     |     |     |     |     |         |      |
| GIS            | 0.332 | 0.265 | 0.301 | 0.315 | 0.817 |     |     |     |     |     |         |      |
| JNJ            | 0.416 | 0.301 | 0.347 | 0.349 | 0.326 | 0.881 |     |     |     |     |         |      |
| KO             | 0.325 | 0.257 | 0.285 | 0.304 | 0.413 | 0.289 | 0.811 |     |     |     |         |      |
| T              | 0.355 | 0.305 | 0.298 | 0.335 | 0.266 | 0.270 | 0.265 | 0.794 |     |     |         |      |
| WMT            | 0.370 | 0.298 | 0.305 | 0.335 | 0.322 | 0.316 | 0.315 | 0.255 | 0.965 |     |         |      |
| XOM            | 0.629 | 0.676 | 0.556 | 0.577 | 0.286 | 0.322 | 0.306 | 0.336 | 0.298 | 1.431 |         |      |

| **Average of Bayesian Nonparametric Covariance** |     |     |    |     |     |     |     |     |     |     | 0.4695  | 0.0090 |
| BAC            | 1.910 |     |    |     |     |     |     |     |     |     |         |      |
| CAT            | 0.723 | 1.765 |     |     |     |     |     |     |     |     |         |      |
| DD             | 0.619 | 0.608 | 1.412 |     |     |     |     |     |     |     |         |      |
| F              | 0.750 | 0.686 | 0.550 | 2.044 |     |     |     |     |     |     |         |      |
| GIS            | 0.357 | 0.285 | 0.309 | 0.357 | 0.842 |     |     |     |     |     |         |      |
| JNJ            | 0.424 | 0.287 | 0.327 | 0.370 | 0.326 | 1.091 |     |     |     |     |         |      |
| KO             | 0.343 | 0.272 | 0.300 | 0.307 | 0.415 | 0.281 | 0.784 |     |     |     |         |      |
| T              | 0.367 | 0.305 | 0.300 | 0.326 | 0.266 | 0.265 | 0.269 | 0.769 |     |     |         |      |
| WMT            | 0.375 | 0.292 | 0.309 | 0.324 | 0.318 | 0.298 | 0.314 | 0.253 | 0.936 |     |         |      |
| XOM            | 0.621 | 0.660 | 0.552 | 0.560 | 0.300 | 0.317 | 0.307 | 0.335 | 0.295 | 1.391 |         |      |
Table 3: ARMA(1,1) Model Estimation Results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>RC$^{5m}$</th>
<th>RK</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<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>
<tr>
<td>$\mu$</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>
</tbody>
</table>

1 This table reports OLS regression results of ARMA model: $\beta_t = \mu + \phi_1 \beta_{t-1} + \rho_1 \epsilon_{t-1} + \epsilon_t$. The value in the bracket are the standard error. Realized beta for AA-SPY combination are calculated using $2 \times 2$ RC, RK and BNC.

Sample period: 2014/07/02 - 2016/06/29.

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</td>
<td>Stdev</td>
<td>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</td>
<td>0.229</td>
<td>0.507</td>
</tr>
<tr>
<td>20%</td>
<td>0.132</td>
<td>0.422</td>
<td>0.314</td>
</tr>
<tr>
<td>30%</td>
<td>0.149</td>
<td>0.624</td>
<td>0.238</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</td>
<td>0.229</td>
<td>0.507</td>
</tr>
<tr>
<td>20%</td>
<td>0.132</td>
<td>0.421</td>
<td>0.314</td>
</tr>
<tr>
<td>30%</td>
<td>0.148</td>
<td>0.623</td>
<td>0.238</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</td>
<td>0.228</td>
<td>0.505</td>
</tr>
<tr>
<td>20%</td>
<td>0.131</td>
<td>0.422</td>
<td>0.311</td>
</tr>
<tr>
<td>30%</td>
<td>0.147</td>
<td>0.624</td>
<td>0.235</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>μ₀</th>
<th>RK</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>∆(γ = 1)</td>
<td>∆(γ = 10)</td>
</tr>
<tr>
<td>Panel A: Decay Rate κ = 0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>-4.569</td>
<td>-4.402</td>
</tr>
<tr>
<td>20%</td>
<td>-9.110</td>
<td>-8.518</td>
</tr>
<tr>
<td>30%</td>
<td>-13.623</td>
<td>-12.351</td>
</tr>
<tr>
<td>Panel B: Decay Rate κ = 0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>-30.822</td>
<td>-28.753</td>
</tr>
<tr>
<td>30%</td>
<td>-46.045</td>
<td>-41.217</td>
</tr>
<tr>
<td>Panel C: Decay Rate κ = 0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>-18.734</td>
<td>-18.104</td>
</tr>
<tr>
<td>20%</td>
<td>-37.261</td>
<td>-34.103</td>
</tr>
<tr>
<td>30%</td>
<td>-55.581</td>
<td>-47.986</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 γ is willing to pay for switching portfolio based on RC to portfolio based on RK or BNC. Assuming one year contains 252 trading days.
Figure 1: Synchronization with Data Augmentation

Figure 2: Top: example of mis-matched return pairs. Bottom: example of zero return
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
Figure 5: Histogram of Portfolio Return Caused by Covariance Uncertainty on July 31, 2014
Figure 6: Histogram of Weights Caused by Covariance Uncertainty on July 31, 2014