JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. (2010)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.1218

A BLOCKING AND REGULARIZATION APPROACH TO HIGH-DIMENSIONAL REALIZED COVARIANCE ESTIMATION

NIKOLAUS HAUTSCH,[a]* LADA M. KYJ[b] AND ROEL C. A. OOMEN[c]

[a] *Institute for Statistics and Econometrics and Center for Applied Statistics and Economics, Humboldt-Universität, Berlin, Germany; Quantitative Products Laboratory, Berlin, Germany; and Center for Financial Studies, Frankfurt, Germany*
[b] *Barclays Capital, New York, NY, USA*
[c] *Deutsche Bank, London, UK; and affiliated with Department of Quantitative Economics, University of Amsterdam, The Netherlands*

SUMMARY

We introduce a blocking and regularization approach to estimate high-dimensional covariances using high-frequency data. Assets are first grouped according to liquidity. Using the multivariate realized kernel estimator of Barndorff-Nielsen *et al*. (2010), the covariance matrix is estimated block-wise and then regularized. The performance of the resulting blocking and regularization ('RnB') estimator is analyzed in an extensive simulation study mimicking the liquidity and market microstructure features of the S&P 1500 universe. The RnB estimator yields efficiency gains for varying liquidity settings, noise-to-signal ratios and dimensions. An empirical application of estimating daily covariances of the S&P 500 index confirms the simulation results.

1. INTRODUCTION

Estimating asset return covariances is indispensable in many areas in financial practice, such as portfolio management, risk management and asset pricing (e.g., Michaud, 1989; Duffie and Pan, 1997; Chan *et al*., 1999; Jagannathan and Ma, 2003). The dimension of the underlying return process is often vast and spans a comprehensive universe of assets, such as that of the S&P 500 index. Producing precise covariance estimates in high dimensions is a substantial challenge: as the number of dimensions increases, an increasing horizon is needed or more structure has to be imposed merely to ensure positive definiteness of the sampling covariance matrix. For instance, Jagannathan and Ma (2003) and Ledoit and Wolf (2003), using daily-level data, show that conditioning of the covariance estimate is important in stabilizing estimates and in providing better out-of-sample portfolio risk management for monthly investment horizons. However, today's practitioners often need to manage their risk with measures that accurately reflect the risk of trading portfolios over comparably short horizons, e.g., a day. Here, the availability of high-frequency asset price data opens up the possibility of efficiently estimating high-dimensional short-term covariances (see, for example, Andersen *et al*., 2001; Barndorff-Nielsen and Shephard, 2004a; Barndorff-Nielsen *et al*., 2010). To produce efficient and positive definite estimates the sampling frequency must increase with the dimension of the underlying asset universe. Increasing the sampling frequency, however, induces severe biases due to market microstructure noise and asynchronous trading effects. As a result, existing high-frequency-based covariance estimators

* Correspondence to: Nikolaus Hautsch, Institute for Statistics and Econometrics, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10099 Berlin, Germany. E-mail: nikolaus.hautsch@wiwi.hu-berlin.de

are either inefficient or are not well conditioned and numerically unstable when applied to vast-dimensional asset universes.

In this paper, we introduce an estimator for vast-dimensional covariances which is consistent, positive definite and well conditioned, while it exploits high-frequency data in a more efficient way. The fundamental idea is to construct one large covariance matrix from a series of smaller covariance matrices, each based on a different sampling time frequency. Grouping together assets trading at similar frequencies offers efficiency gains with respect to data synchronization. In a second step, the resulting covariance estimate is regularized to ensure a positive definite and well-conditioned matrix. Based on an extensive simulation study mimicking the empirical features of the S&P 1500 index, we evaluate the effects of (i) blocking and regularization, (ii) the number of clusters and (iii) cluster size determination in relation to observation frequency distributions and underlying noise properties. It is shown that even for a small number of clusters the RnB estimator can significantly reduce estimation errors. This is particularly true if the underlying cross-section of trading frequencies is heterogeneous.

The traditional high-frequency-based covariance estimator is the realized covariance estimator, defined as the cumulative sum of the cross-products of multivariate returns synchronized in calendar time (e.g., every 5 minutes). This estimator becomes ill conditioned (in the extreme case not positive definite) when the cross-sectional dimension is high relative to the number of intra-day sampling intervals. However, if the sampling frequency is increased, covariance estimates are dominated by market microstructure effects such as the bid–ask bounce, price discreteness, and non-synchroneity of price observations (see, for instance, Epps, 1979; Zhang *et al*., 2005; Bandi and Russell, 2006; Hansen and Lunde, 2006). A number of recent papers have offered alternative covariance estimators that address the above-mentioned complications. Hayashi and Yoshida (2005) introduced an estimator based on the cumulative sum of the cross-product of all fully and partially overlapping transaction returns. This estimator explicitly accounts for asynchroneity of the processes and is free of any biases. Bandi *et al*. (2008), Griffin and Oomen (2010), Martens (2006), Sheppard (2006), and Voev and Lunde (2007) study numerous alternative estimators in a bivariate setting via optimal sampling or lead-lag estimation to obtain substantial efficiency gains. More recently, Barndorff-Nielsen *et al*. (2010, hereafter BNHLS) introduce the multivariate realized kernel (RK) estimator which is applicable in high dimensions, is consistent in the presence of market microstructure noise and is guaranteed to be positive semi-definite. However, a substantial drawback is that synchronization is achieved by 'refresh time sampling' (RTS); i.e., the cross-section of asset returns is sampled whenever all assets have been traded. RTS implies a considerable loss of information if both the cross-sectional dimension and the cross-sectional variation in asset observation frequencies are high making the estimator quite inefficient.

As shown in this paper, blocking the covariance matrix retains a greater amount of data post-synchronization and significantly increases the precision of the corresponding estimates compared to an 'all-in-one approach'. The asset clusters are chosen in a data-driven way minimizing the cross-sectional variation of observation frequencies within each cluster. This leads to blocks of assets implying different RTS timescales. While the individual covariance blocks are positive semi-definite, the whole covariance matrix does not necessarily fulfill this requirement. Thus, a second-stage regularization technique is employed drawing upon results from random matrix theory to generate a positive definite and well-conditioned matrix. In the proposed procedure, the number of blocks controls the trade-off between using more data in the first stage but requiring more regularization in the second stage. Applying the RnB estimator to the estimation of the (realized) variance of randomized portfolios based on the S&P 500 index from January 2007 to April 2009, we show that blocking reduces estimation errors and clearly increases the estimator's efficiency.

DOI: 10.1002/jae

The remainder of the paper is organized as follows. In Section 2, we present the underlying theoretical setting. Section 3 introduces the used blocking and regularization techniques, whereas Section 4 illustrates the simulation study. In Section 5, empirical results and corresponding discussions are given. Finally, Section 6 concludes.

2. BACKGROUND

2.1. Notation and Underlying Assumptions

Consider a p-dimensional log price process $X = (X^{(1)}, X^{(2)}, \ldots, X^{(p)})'$, which is observed over the interval $[0, T]$. For ease of exposition we set $T = 1$ throughout the remainder of this paper. The observation times for the ith asset are written as $t_1^{(i)}, t_2^{(i)}, \ldots$, and are assumed to be strictly increasing. Hence the realizations of $X^{(i)}$ at the observation times are given by $X^{(i)}(t_j)$, for $j = 1, 2, \ldots, N^{(i)}$, and $i = 1, 2, \ldots, p$. The observed price process, X, is assumed to be driven by the efficient price process, Y, which is modeled as a Brownian semi-martingale defined as

$$Y(t) = \int_0^t a(u)\mathrm{d}u + \int_0^t \sigma(u)\mathrm{d}W(u) \tag{1}$$

where a is a predictable locally bounded drift process, σ is a càdlàg volatility matrix process, and W is a vector of independent Brownian motions. Market microstructure frictions are modeled through an additive noise component as

$$X^{(i)}(t_j) = Y^{(i)}(t_j) + U_j^{(i)}, \qquad j = 0, 1, \ldots, N^{(i)} \tag{2}$$

where $U_j^{(i)}$ is covariance stationary and satisfies the conditions (i) $\mathrm{E}[U_j^{(i)}] = 0$, and (ii) $\sum_h |h\Omega_h| < \infty$ with $\Omega_h = \mathrm{Cov}[U_j, U_{j-h}]$.

The object of econometric interest in this study is the quadratic variation of Y, i.e., $[Y] = \int_0^1 \Sigma(u)du$ with $\Sigma = \sigma\sigma'$, which is to be estimated from discretely sampled, non-synchronous and noisy price observations.

2.2. The Multivariate Realized Kernel Estimator

The multivariate realized kernel estimator of BNHLS is the first to simultaneously address market microstructure effects and asynchronous price observations while guaranteeing consistency and positive semi-definiteness. RK estimation is a two-part process. As in Harris *et al.* (1995), the observations are synchronized via refresh time sampling (RTS). Refresh times are defined as the time it takes for all the assets in a set to trade or refresh posted prices. Once all the assets have traded, the most recent new price is used to form the RTS timescale. More formally, the first refresh time sampling point can be defined as $\mathrm{RFT}_1 = \max(t_1^{(1)}, \ldots, t_1^{(p)})$ and $\mathrm{RFT}_{j+1} = \mathrm{argmin}(t_k^{(i)} | t_k^{(i)} > \mathrm{RFT}_j, \forall i)$. Refresh time synchronization allows us to define high-frequency vector returns as $x_j = X_{\mathrm{RFT}_j} - X_{\mathrm{RFT}_{j-1}}$, where $j = 1, 2, \ldots, n$, and n is the number of refresh time observations.

The multivariate realized kernel is defined as

$$K(X) = \sum_{h=-H}^{H} k\left(\frac{h}{H+1}\right)\Gamma_h \tag{3}$$

DOI: 10.1002/jae

where $k(x)$ is the weight function of a Parzen kernel, and Γ_h is a matrix of autocovariances given by

$$\Gamma_h = \begin{cases} \sum_{j=|h|+1}^{n} x_j x'_{j-h}, & h \geq 0, \\ \sum_{j=|h|+1}^{n} x_{j-h} x'_j, & h < 0 \end{cases} \tag{4}$$

The bandwidth parameter H is optimized with respect to the mean squared error criterion by setting $H = c^* \xi^{4/5} n^{3/5}$, where $c^* = 3.5134$, $\xi^2 = \omega^2 / \sqrt{IQ}$ denotes the noise-to-signal ratio, ω^2 is a measure of microstructure noise variance, and IQ is the integrated quarticity as defined in Barndorff-Nielsen and Shephard (2002). The bandwidth parameter H is computed for each individual asset and then a global bandwidth is selected for the entire set of assets considered. In this study the global bandwidth is set as the mean of the bandwidths for the assets within each corresponding block. The resulting global bandwidth may be suboptimal for a very diverse set of bandwidths, providing another motivation for grouping similar assets together. For a more detailed discussion of bandwidth selection, see the web appendix of BNHLS.

3. THE RNB ESTIMATOR

3.1. Motivation

RTS may make inefficient use of data and in high-dimensional covariance estimation contributes to the so-called 'curse of dimensionality' problem where the number of observations is not much greater than the number of dimensions. To illustrate this point consider a universe of p assets, each independently traded with equal Poisson arrival rate β. Define $\mathcal{M}(p) = \mathrm{E}[\max(t_1^{(1)}, t_1^{(2)}, \ldots, t_1^{(p)})]$ as the expected maximum waiting time for all assets to have traded at least once. Then, using the fact that $\Pr[\max(t_1^{(1)}, t_1^{(2)}, \ldots, t_1^{(p)}) < u] = (1 - e^{-\beta u})^p$, $\mathcal{M}$ can be computed as

$$\mathcal{M}(p) = \int_0^\infty \beta p (1 - e^{-\beta u})^{p-1} e^{-\beta u} u \, \mathrm{d}u \tag{5}$$

and can be approximated by $\mathcal{M}(p) \simeq \frac{1}{\beta} \log(0.9 + 1.8p)$. Thus, the implied data loss fraction of the RTS scheme is

$$\mathcal{L}(p) = 1 - (\beta \mathcal{M}(p))^{-1} \tag{6}$$

The solid line in Figure 1(A) plots the relationship between $\mathcal{L}(p)$ and p, implying, for example, data losses of 33%, 66%, and 81% for $p = 2$, 10, 100, respectively. We should emphasize that this is a conservative illustration: the data loss with unequal arrival rates is substantially higher as the sampling points are determined by the slowest trading asset. Consider, for instance, a scenario where p_1 assets have an arrival rate of β_1 and p_2 assets have an arrival rate of β_2, with $\beta_1 \neq \beta_2$. The expected maximum waiting time for all assets to have traded at least once can be derived from $\Pr[\max(t_1^{(1)}, t_1^{(2)}, \ldots, t_1^{(p_1)}, t_1^{(p_1+1)}, \ldots, t_1^{(p_1+p_2)}) < u] = (1 - e^{-\beta_1 u})^{p_1} (1 - e^{-\beta_2 u})^{p_2}$. The dashed gray line in Figure 1(A) represents the data loss for the most active asset in the scenario where $p_1 = p_2$ and $\beta_2 = 5\beta_1$ and shows that variation in arrival rates further increases the implied data loss.

3.2. Blocking Strategy

The blocking strategy starts by ordering the assets in the covariance matrix according to observation frequencies, with the most liquid asset in the top left corner and the least liquid asset in the bottom

DOI: 10.1002/jae

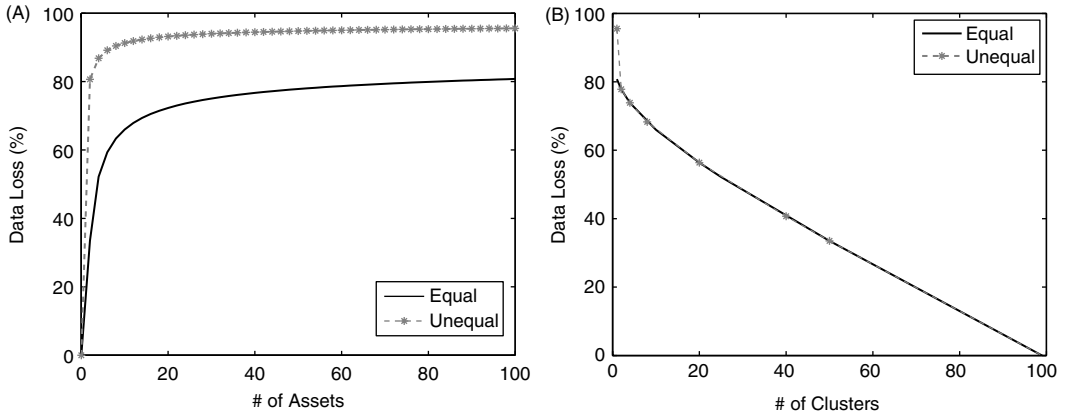

Figure 1. Illustration of the refresh time sampling scheme implied data loss: (A) data loss by number of assets; (B) data loss by number of clusters. Part A reports the percentage of data loss as the number of assets increases and part B reports the percentage of data loss as the number of clusters increases. Portfolios composed of equal and unequal arrival rates are presented. The unequal arrival rates are set at $\beta_2 = 5\beta_1$ with an equal number of assets from each group. For part B, the number of assets is equal to 100

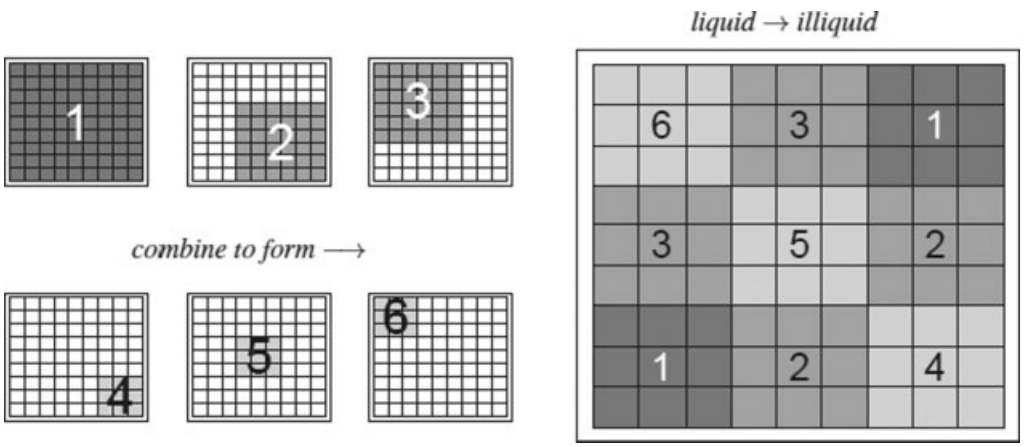

Figure 2. Visualization of the blocking strategy. Assets are ordered according to liquidity, with the most liquid asset in the top left corner of the covariance matrix and the least liquid asset in the bottom right corner. Covariance estimates are computed on a series of blocks and then combined to form a multi-block estimator

right corner. This initial step ensures that subsequent blocks will group together assets with similar arrival rates. Proceed by dividing the assets into liquidity-based clusters. Asset clusters are then combined to form a series of blocks of the covariance matrix, where each block is itself a covariance matrix.

Figure 2 illustrates the construction of the BLOCK estimator with three equal-sized asset clusters. The six resulting covariance blocks, each with a different RTS timescale, combine to form this multi-block estimator. Block 1 implies estimating the multivariate RK for the entire set of assets. This serves as a baseline covariance estimate for the BLOCK estimator. In the next step, the covariances of the six least liquid assets are replaced by the kernel estimate of block 2. Similarly, the covariances of the six most liquid assets are replaced by estimates of block 3.

DOI: 10.1002/jae

Finally, estimates for blocks 4, 5, and 6, composed of the three slowest assets, three middle assets, and three fastest assets, respectively, replace the corresponding elements in the BLOCK estimator. In the end, the farthest off-diagonal blocks (1) are from the original nine-asset RK, the middle off-diagonal blocks (2) and (3) stem from the six-asset RKs, and the diagonal blocks (4), (5), and (6) are from the corresponding three-asset RKs.

The elements in the diagonal blocks of this estimator are more precisely estimated than the original RK. The off-diagonal blocks are no worse in terms of RTS than the original RK. The precision gains are driven by the fact that this multi-timescale design substantially increases the effective number of observations used and in turn speeds up convergence. The blocking procedure increases the sample size without imposing any additional structure on the covariance estimate.[1]

Grouping assets according to their trading frequency directly addresses the data reduction problem. As a result, each resulting block has an individual RTS timescale, allowing for liquid sets to include more observations than before. Referring back to the above illustration, the data loss fraction in case of K (equal-sized) clusters is

$$\mathcal{L}(p, K) = 1 - (\beta \mathcal{M}(p/K))^{-1} \tag{7}$$

Figure 1(B) shows that blocking yields significant efficiency gains, e.g., with $p = 100$ and $K = 10$, the data loss is 66% instead of 81% without blocking. Moreover, the data loss decreases as the number of clusters increases. The data loss is driven by the sizes of the individual clusters where $p = 100$ with $K = 5$ has the same data loss as $p = 500$ with $K = 25$ and $p = 1000$ with $K = 50$. The first difference of the data loss function with respect to the number of clusters suggests that the greatest gains in data loss improvement are accomplished with a relatively modest number of clusters, e.g., four or five. Finally, the impact of blocking is greater in the presence of unequal arrival rates. By separating the illiquid and liquid assets into two clusters, the maximum data loss moves to the lower data loss curve of equal arrival rates.

Our approach is fundamentally different from other covariance 'blocking' estimators, as our strategy is due to observation frequency and is exclusively focused on estimation efficiency. Bonato *et al*. (2009) use blocks to group assets with high dependence together according to predetermined economic criteria (i.e., industry or market capitalization). In contrast to our blocking strategy, these methods by no means guarantee efficient use of data.

The individual covariance blocks can in principle be estimated also using alternative estimators to the realized kernel. For instance, pooling together assets with similar frequency reduces the Epps effect and would make (block-wise) lower-dimensional realized covariances as underlying estimators also applicable as long as block sizes (and thus block-specific dimensions) are not chosen too large. In this study, we explicitly focus on the realized kernel which allows us to employ arbitrary block sizes and thus provides maximal flexibility also in very high dimensions.

As discussed by Barndorff-Nielsen and Shephard (2004b), estimators which include contemporaneous returns are not robust in the presence of large jumps. Consequently, the employed multivariate realized kernel estimator and thus the resulting blocking estimator estimate the total quadratic covariation but do not allow for disentangling between jump and diffusion components. The occurrence of jumps results in larger bandwidths but, as shown in Barndorff-Nielsen *et al*. (2008), overestimating the bandwidth is more tolerable than underestimating. These effects are expected to be stronger in small asset clusters, but will be averaged out in large asset clusters. However, given the block sizes employed in this paper we do not expect the estimator's performance to be systematically affected by jumps. Alternatively, replacing realized kernels by (block-wise)

[1] Alternatively, correlations could be used instead of covariances. This paper presents covariances as efficiency gains in the variance elements of the estimates are of interest.

DOI: 10.1002/jae

bipower covariation estimators as proposed by Barndorff-Nielsen and Shephard (2005) would allow for an explicit separation between jumps and continuous components. This would open another interesting line of research which we leave for future studies.

3.3. Regularization

While our proposed blocking estimator improves estimation precision, it is done at the expense of positive semi-definiteness and well-conditioning. Ill-conditioned matrices are characterized by eigenvalues vanishing to zero, behave similar to numerically singular matrices and result in unstable matrix inversions. Guaranteeing the covariance matrix to be both positive definite and well-conditioned necessitates the consideration of regularization techniques which recover both of these properties. There are many regularization techniques that can be applied to covariance estimates (see, for example, Ledoit and Wolf, 2004; Qi and Sun, 2006; Bickel and Levina, 2008) and the choice of technique depends on the manner in which the covariance matrix will be applied. This study focuses on a signal–noise decomposing technique. 'Eigenvalue cleaning' is a regularization technique developed from random matrix theory by Laloux *et al*. (1999) and further developed by Tola *et al*. (2008). It follows the same intuition as the original (Stein, 1977) eigenvalue shrinkage approach. Applications of random matrix theory have emerged as a common regularization technique for high-dimensional realized covariance matrices. Other applications include, for instance, Onatski (2009), Wang and Zou (2010), and Zumbach (2009).

Eigenvalue cleaning draws upon random matrix theory to determine the distribution of eigenvalues as a function of the ratio of N observations relative to p dimensions $q = N/p$. The regularization focuses on the correlation matrix R with spectral decomposition $R = Q\Lambda Q'$, where Q is the matrix of eigenvectors and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of eigenvalues. Under the null hypothesis of independent assets, the correlation matrix R is the identity matrix, and the distribution of eigenvalues is given by the Marchenko–Pastur distribution with maximum eigenvalue given by $\lambda_{\max} = \sigma^2 \left(1 + \frac{1}{q} + 2\sqrt{\frac{1}{q}}\right)$, where σ^2 is the variance of the entire portfolio.

The principle of eigenvalue cleaning is to compare the empirical eigenvalues with those arising under the null hypothesis of independent assets and to identify those eigenvalues which deviate from those driven by noise. In particular, the largest estimated eigenvalue $\hat{\lambda}_1$ clearly violates the 'pure noise' hypothesis and can be seen as a 'market signal'. Removing this eigenvalue and recomputing $\sigma^2 = 1 - \hat{\lambda}_1/p$ (and correspondingly $\lambda_{\max}$) as the market-neutral variance has the effect of 'tightening' the Marchenko–Pastur density and allowing for smaller signals to be better identified. Then, large positive eigenvalues greater than (the re-scaled) $\lambda_{\max}$ are identified as further 'signals'. Eigenvalues smaller than this threshold are identified as noise-driven eigenvalues and are transformed to take a value away from zero. In particular:

$$\tilde{\lambda}_i = \begin{cases} \hat{\lambda}_i & \text{if} \quad \hat{\lambda}_i > \lambda_{\max} \\ \delta & \text{otherwise} \end{cases} \tag{8}$$

where the parameter δ is chosen such that the trace of the correlation matrix is preserved. To ensure that the resulting matrix is positive definite, the trace of the positive semi-definite projection of the correlation matrix is used. In particular,

$$\delta = \frac{\text{trace}(R_+) - \sum_{(\hat{\lambda}_i > \lambda_{\max})} \hat{\lambda}_i}{p - (\text{no. of } \hat{\lambda}_i > \lambda_{\max})} \tag{9}$$

DOI: 10.1002/jae

This results in a matrix $\hat{R} = Q\hat{L}Q'$, where $\hat{L} = \text{diag}(\tilde{\lambda}_i)$. Finally, the RnB estimator is defined as the corresponding covariance constructed from $\hat{R}$.

In applications one may conservatively set the number of observations used in the eigenvalue cleaning procedure to be equal to the minimum observation in any block of the multi-block estimator. Moreover, in the analysis that follows, matrices are regularized only if they are either non-positive definite or ill conditioned. A matrix is defined to be ill conditioned when the condition number of the matrix, $\kappa(A) = \left|\frac{\lambda_{\max}}{\lambda_{\min}}\right|$, is greater than $10 \times p$.

4. MONTE CARLO STUDY

The objective of the simulations below is to examine the performance of the RnB estimator in the context of three challenges: (i) non-synchronous price observations; (ii) price distortions due to market microstructure effects; and (iii) high dimensions relative to the number of observations. To evaluate the estimator in a realistic setting, the simulation study is designed in an empirically driven way mimicking the market microstructure effects and non-synchroneity of price observations of the S&P 1500 index. This setting allows us also to study the impact of the ratio of observations to dimensions by holding intra-day observations fixed and changing the ratio by expanding the number of dimensions towards a high-dimensional setting. This provides insight into the performance of the proposed estimator in realistic financial settings where the investment universe considered may easily be in the range of hundreds of assets.

4.1. Simulation Design

The underlying efficient price process Y is a simple diffusion with a constant covariance, i.e.,

$$Y_t = \Theta Z_t \tag{10}$$

where Θ' is the Cholesky factorization of the covariance matrix such that $\Theta\Theta' = \Sigma$, and Z is a $(p \times 1)$ vector of independent standard Brownian motions. To simulate the process, we use a Euler discretization scheme with step size $\Delta = 1/23400$. The covariance structure is generated from an ad hoc statistical three-factor model that closely mimics the cross-sectional distribution of correlations for the S&P 1500 universe (see the Appendix for details). The results reported below are based on 1000 simulation replications.

Non-synchronous price observations and the accompanying Epps effect are major obstacles in covariance estimation. The simulation is designed to include this feature by considering asset liquidity as a measure of the non-synchroneity of observations. Specifically, the asset liquidity represented by the number of trades per day is used as a proxy for observation frequency. By drawing annual average numbers of daily trades from the S&P 1500, three liquidity classes can be identified: the 500 most liquid assets ('Liquid 500'), the next 400 liquid assets ('Middle 400'), and the remaining assets ('Illiquid 600'). These categories are chosen to be liquidity counterparts to the large, mid, and small cap S&P 500, S&P 400 and S&P 600 indices. Arrival times are then modeled by uniformly sampling $M^{(i)}$ observations from [0, 1]. Figure 3 illustrates the liquidity scenarios considered: (i) a liquid set of assets, where the number of observations is sampled from the Liquid 500; (ii) a heterogeneous (S&P 1500 mimicking) set of assets where the number of observations is sampled from the Illiquid 600, Middle 400, and Liquid 500; and (iii) an illiquid set of assets where the number of observations is sampled from the Illiquid 600.

DOI: 10.1002/jae

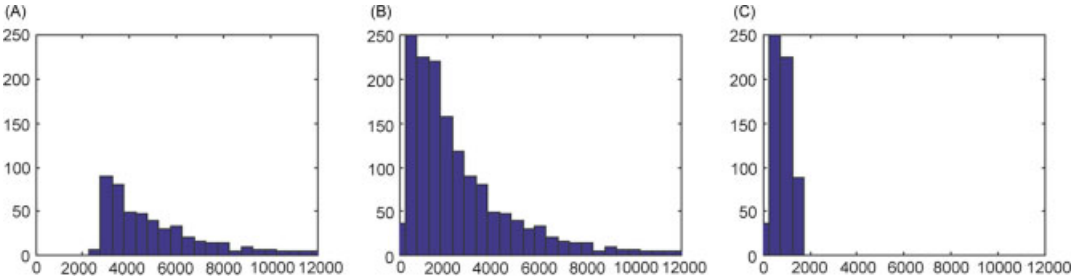

Figure 3. Liquidity classification by observation frequency: (A) liquid; (B) heterogeneous; (C) illiquid. Parts A, B, and C show the distribution of number of observations from the top 500 assets, entire sample, and bottom 600 assets of the S&P 1500 universe when ordered by number of observations. This figure is available in color online at wileyonlinelibrary.com/journal/jae

To allow for market microstructure effects, additive noise is introduced to the simulated efficient price process for asset i at time j as: $X_j^{(i)} = Y_j^{(i)} + U_j^{(i)}$ for $j = 0, \ldots, N$, where the market microstructure effect for each asset i is given as $U_j^{(i)} \sim N(0, \omega_{(i)}^2)$.

The choice of $\omega_{(i)}^2$ in the simulation is calibrated to the S&P 1500 universe to ensure a realistic setup. Table I reports the percentiles of the noise ratio of Oomen (2006) defined as $\gamma^2 = M\omega^2/\sigma^2$. Interestingly, the distribution of this normalized noise-to-signal ratio is similar across the different groups, with the liquid group showing the greatest variation (see Oomen, 2009, for further discussion). Motivated by this, a spectrum of microstructure noise levels is considered where $\gamma^2 = (0.25, 0.375, 0.50, 1.0)$, corresponding to low noise, medium noise, high noise, and very extreme noise, respectively.

Finally, portfolio dimensions are set to realistic investment sizes of dimension $p = 64$ and 256.[2] Note that portfolios of this high-dimension size have rarely been studied in the realized covariance literature. A notable exception is Wang and Zou (2010), who consider a very high-dimensional setting, $p = 512$, with asynchronously observed assets each observed only 200 times per day. Their analysis focuses on the performance of threshold regularization of realized covariance, where the underlying realized covariance estimator is synchronized via previous-tick interpolation and does not directly address the asynchroneity or data reduction issues.

Table I. Microstructure noise statistics of S&P 1500 trade data (2008)

	$\gamma^2 = M\omega^2/\sigma^2$				
	Q5	Q25	Q50	Q75	Q95
Illiquid 600	0.22	0.27	0.34	0.41	0.63
Middle 400	0.23	0.31	0.38	0.46	0.76
Liquid 500	0.20	0.29	0.36	0.46	0.94

Note: This table reports the 5th, 25th, 50th, 75th, and 95th percentile of the noise ratio $\gamma^2 = M\omega^2/\sigma^2$ computed across all stocks in each group and all days over the period 2 January 2008 to 31 December 2008. The index constituent lists are from January 2009. Assets are grouped according to liquidity characteristics into Illiquid 600, Middle 400, and Liquid 500.

[2] The dimensions are chosen to be powers of 2, which in turn allows examination of sequentially smaller cluster sizes, while still maintaining equal cluster size.

DOI: 10.1002/jae

Since the true underlying covariance matrix is known, the estimator's performance is assessed using three statistical criteria. First, the scaled Frobenius norm defined as

$$||A||_{F_p} = \sqrt{\left(\sum_{i=1}^{p}\sum_{j=1}^{p}|a_{ij}|^2\right)/p} = \sqrt{\text{trace}(AA^T)/p}$$

where A is the difference between the estimate and the parameter value. Scaling by the dimension size, p, allows for comparability as the number of assets increases. Second, the scaled Euclidean distance between two vectors,

$$||a||_{E_p} = \sqrt{\frac{a_1^2 + \cdots + a_n^2}{p}}$$

is used to isolate between estimation errors stemming from covariance and variance elements. Finally, as the invertibility of the resulting estimates is of interest, the positive definiteness of a covariance estimate is determined by the smallest estimated eigenvalue being positive, i.e.

$$\text{PD} = \begin{cases} 1 & \text{if } \hat{\lambda}_{\min} > 0 \\ 0 & \text{otherwise} \end{cases}$$

4.2. Results

4.2.1 Simulation 1: Market Microstructure Effects and Liquidity

The first simulation exercise examines the impact of market microstructure effects under different distributions of liquidity. Tables II and III report the scaled Frobenius norm of the covariance matrix (FRB) and inverse covariance matrix (INV) estimates, as well as the fraction of covariance estimates that are positive definite (PD). The estimates considered are the multivariate realized kernel (RK), the blocking estimator based on four clusters of equal size (BLOCK), together with regularized versions using eigenvalue cleaning (RRK and RnB, respectively). All criteria are evaluated under varying noise levels, observation arrival structures, and dimension sizes.

Tables II and III show the results for $p = 64$ and $p = 256$, respectively, and four general findings emerge. First, estimation error increases with market microstructure effects. Holding observation frequency constant and increasing the noise level results in increased estimation errors. This feature is true for both error evaluation criteria FRB and INV. Recalling that market microstructure effects are treated as noise, this is a fully anticipated outcome. Second, holding the noise level fixed and decreasing the observation frequency increases estimation error. Third, blocking reduces estimation error as well as positive definiteness. It is shown that for each noise and liquidity scenario the estimation error of the blocked estimator is smaller than that of the corresponding realized kernel. This result validates our expectations that grouping similar assets together into clusters reduces estimation error. However, this is accomplished at the cost of positive definiteness. Fourth, estimation precision gains realized due to blocking are preserved and sometimes even further improved after regularization.

Table III shows that for higher dimensions the PD statistic is now virtually zero for all RK estimates in the heterogeneous and illiquid settings. The illiquid setting has few observations and the heterogeneous setting suffers the greatest data reduction due to RTS. Although the RK estimator is positive semi-definite by construction, it does require at least p observations to maintain this property. The additional reduction between the unregularized and regularized statistics suggests that by imposing structure via regularization reduces estimation error in high-dimensional systems.

DOI: 10.1002/jae

Table II. Performance of RnB estimator for $p = 64$ and 4 asset clusters

	Unregularized				Regularized			
	RK		BLOCK		RRK		RnB	
	FRB	PD	FRB	PD	FRB	INV	FRB	INV
Panel A: Low noise ($\gamma^2 = 0.250$)								
Liquid	0.528	1.000	0.496	0.590	0.532	1.258	0.499	1.256
Heterogeneous	1.062	1.000	0.902	0.000	1.021	1.444	0.862	1.401
Illiquid	1.289	1.000	1.156	0.000	1.242	1.637	1.097	1.467
Panel B: Medium noise ($\gamma^2 = 0.375$)								
Liquid	0.555	1.000	0.523	0.507	0.557	1.269	0.522	1.265
Heterogeneous	1.100	1.000	0.938	0.000	1.058	1.475	0.890	1.405
Illiquid	1.343	1.000	1.207	0.000	1.295	1.728	1.142	1.498
Panel C: High noise ($\gamma^2 = 0.500$)								
Liquid	0.578	1.000	0.545	0.458	0.578	1.280	0.541	1.274
Heterogeneous	1.132	1.000	0.969	0.000	1.089	1.502	0.915	1.412
Illiquid	1.386	1.000	1.249	0.000	1.338	1.796	1.178	1.524
Panel D: Extreme noise ($\gamma^2 = 1.000$)								
Liquid	0.643	1.000	0.607	0.319	0.638	1.318	0.598	1.310
Heterogeneous	1.224	1.000	1.056	0.000	1.179	1.585	0.988	1.438
Illiquid	1.508	1.000	1.364	0.000	1.458	1.973	1.279	1.587

Note: This table reports the scaled Frobenius norm of the covariance matrix (FRB) and inverse covariance matrix (INV) estimates as well as the fraction of covariance estimates that are positive definite (PD). The estimates considered are the multivariate realized kernel (RK), blocking estimator based on four clusters of equal size (BLOCK) together with regularized versions using eigenvalue cleaning (RRK and RnB).

The much larger difference in the INV statistic clearly shows the importance of blocking and regularization in estimating the inverse of high-dimensional systems. Moreover, it is shown that regularization alone is not sufficient as blocking *and* regularization result in substantially less estimation error of the inverse than the corresponding regularized (but not blocked) RK estimator. In summary, blocking universally reduces the estimation error relative to RK estimates, and the greatest improvement is achieved in the most heterogeneous observation setting resembling the characteristics of the S&P 1500 universe.[3]

4.2.2 Simulation 2: Number of Asset Clusters

The second simulation exercise examines the performance gains in the RnB estimator as the number of asset clusters increases. In this context, the simulation environment is set as $p = 256$, noise level $\gamma^2 = 0.375$, with heterogeneous observation structure. Again, the asset clusters are restricted to being of equal size, but as the number of clusters increases, the size of individual clusters decreases. Note that the estimator with one cluster has only one RTS timescale and is equivalent to the RK estimator. In addition to the RnB estimator constructed with varying numbers of clusters, results are also reported for the Hayashi and Yoshida (HY) estimator, which is treated as a baseline. The simulation design implies that the market microstructure effects are uncorrelated across assets. As a result, the HY estimator is sensitive to noise accumulation on the variance estimates, but not to noise accumulation on the covariance estimates. Therefore, estimation errors in the diagonal elements are distinguished from errors in the off-diagonal elements as well as in those of the

[3] Robustness of the regularization procedure was evaluated with respect to different choices of number of observations used to determine the maximum eigenvalue threshold. The comparison between regularized RK and BLOCK estimators remains qualitatively the same.

DOI: 10.1002/jae

Table III. Performance of RnB estimator for $p = 256$ and 4 asset clusters

	Unregularized				Regularized			
	RK		BLOCK		RRK		RnB	
	FRB	PD	FRB	PD	FRB	INV	FRB	INV
Panel A: Low noise ($\gamma^2 = 0.250$)								
Liquid	1.120	1.000	1.060	0.000	1.103	1.683	1.051	1.569
Heterogeneous	2.348	0.000	1.991	0.000	2.267	2.710	1.992	1.474
Illiquid	2.841	0.000	2.537	0.000	2.784	5.670	2.526	1.897
Panel B: Medium noise ($\gamma^2 = 0.375$)								
Liquid	1.178	1.000	1.115	0.000	1.158	1.822	1.098	1.666
Heterogeneous	2.439	0.000	2.075	0.000	2.362	3.110	2.050	1.514
Illiquid	2.964	0.000	2.652	0.000	2.911	6.806	2.610	2.058
Panel C: High noise ($\gamma^2 = 0.500$)								
Liquid	1.227	1.000	1.162	0.000	1.206	1.942	1.140	1.750
Heterogeneous	2.514	0.000	2.143	0.000	2.439	3.454	2.101	1.551
Illiquid	3.060	0.000	2.742	0.000	3.010	7.756	2.679	2.178
Panel D: Extreme noise ($\gamma^2 = 1.000$)								
Liquid	1.369	1.000	1.299	0.000	1.345	2.327	1.265	2.002
Heterogeneous	2.721	0.000	2.338	0.000	2.654	4.485	2.250	1.670
Illiquid	3.318	0.000	2.989	0.000	3.273	10.267	2.874	2.412

Note: This table reports the scaled Frobenius norm of the covariance matrix (FRB) and inverse covariance matrix (INV) estimates as well as the fraction of covariance estimates that are positive definite (PD). The estimates considered are the multivariate realized kernel (RK), blocking estimator based on four clusters of equal size (BLOCK) together with regularized versions using eigenvalue cleaning (RRK and RnB).

entire matrix. Accordingly, Table IV reports two new statistics: the scaled Euclidean norm of the diagonal elements of the estimate (DIA) and the scaled Euclidean norm of the vectorized off-diagonal elements of the estimate (OFF).

Panel A of Table IV presents the results for the entire matrix, panel B is associated with the most liquid half of assets, and panel C reports findings for the most liquid quarter of assets. Again, four main results emerge. First, relative to the HY estimator, the RK estimator offers a larger reduction in estimation error of the variance elements (DIA), but performs poorly for the off-diagonal elements (OFF). Second, by increasing the number of asset clusters in the BLOCK estimator, the error in all reported statistics is reduced. Error reduction cannot just be attributed to decreasing the cluster (and by extension block) size, but rather it is due to the exclusion of less liquid assets. Hence segregating illiquid assets from liquid assets is a substantial step in gaining estimation efficiency. Third, while there is swift error reduction for dividing one cluster into two and two into four, after four it slows down substantially. This suggests that the bulk of estimation gains can be achieved with a parsimonious model. Finally, it is shown that due to the error accumulation on the diagonal the HY estimator is a poor estimator of the inverse. In contrast, the RnB estimator with only two asset clusters provides smaller estimation errors of the inverse than the HY estimator, and the improvement increases with additional clusters.

Panel D of Table IV presents the corresponding results for the least liquid half of assets. A comparison with panel B shows that the estimation error is more than doubled for the illiquid set. Furthermore, the off-diagonal estimation error (OFF) is relatively closer to the HY benchmark. It also turns out that blocking reduces error for liquid sub-matrices, but may increase error for illiquid matrices. The latter effect is mainly present in the switch from one to two asset blocks, whereas a further increase of the number of blocks again reduces estimation errors. Hence segregating illiquid

DOI: 10.1002/jae

Table IV. Results for different numbers of asset clusters for $p = 256$

		Unregularized			Regularized			Unregularized			Regularized		
		FRB	DIA	OFF	FRB	OFF	INV	FRB	DIA	OFF	FRB	OFF	INV
		Panel A: Entire matrix (1 : 256)						*Panel B: Upper quarter (1 : 64)*					
BLOCK # clusters	HY	1.380	0.755	0.815	1.119	0.579	2.409	0.794	0.751	0.180	0.794	0.180	1.758
	1	2.442	0.189	1.722	2.363	1.665	3.118	1.183	0.184	0.826	1.167	0.815	3.248
	2	2.267	0.160	1.599	2.197	1.549	1.793	0.629	0.098	0.439	0.629	0.439	2.213
	4	2.060	0.144	1.453	2.043	1.441	1.517	0.478	0.076	0.333	0.478	0.333	1.228
	8	1.921	0.135	1.355	1.932	1.362	1.462	0.464	0.072	0.324	0.461	0.322	1.557
	16	1.848	0.129	1.303	1.867	1.316	1.448	0.453	0.069	0.316	0.449	0.313	1.646
	32	1.810	0.124	1.277	1.820	1.283	1.441	0.444	0.066	0.310	0.440	0.307	1.648
		Panel C: Upper half (1 : 128)						*Panel D: Lower half (129 : 256)*					
BLOCK # clusters	HY	0.888	0.752	0.331	0.888	0.331	1.912	1.298	0.757	0.743	1.057	0.514	1.989
	1	1.677	0.185	1.178	1.656	1.163	5.446	1.764	0.192	1.239	1.691	1.188	1.672
	2	0.907	0.101	0.637	0.886	0.622	1.255	1.857	0.202	1.305	1.793	1.259	2.241
	4	0.872	0.094	0.613	0.846	0.594	1.202	1.758	0.180	1.236	1.672	1.175	1.563
	8	0.837	0.089	0.588	0.812	0.571	1.164	1.648	0.168	1.159	1.571	1.104	1.313
	16	0.816	0.085	0.574	0.792	0.556	1.148	1.578	0.160	1.110	1.506	1.058	1.223
	32	0.805	0.082	0.566	0.777	0.546	1.138	1.540	0.154	1.083	1.461	1.027	1.177

Note: This table reports the scaled Frobenius norm of the covariance matrix (FRB) and inverse covariance matrix (INV) estimates, the scaled Euclidean norm of the diagonal elements of the estimate (DIA) and the scaled Euclidean norm of the vectorized off-diagonal elements of the estimate (OFF). The estimates considered are the Hayashi and Yoshida estimator (HY) and the blocking estimator based on varying number of equal-sized clusters (BLOCK) together with regularized versions using eigenvalue cleaning (RnB). Each panel shows the results for the HY estimator and the BLOCK estimator with varying number of asset clusters of equal size for $p = 256$, $\gamma^2 = 0.375$, and the heterogeneous observation arrival set. Panels A, B, C, and D show the corresponding results for various subsets of the matrix.

assets from liquid ones yields improved estimators if the overall liquidity is high (as in panel B or C) but increases estimation errors if the overall liquidity is low (as in panel D). According to our results this effect can only be attributed to the choice of bandwidth.

4.2.3 Simulation 3: Relaxing Equal Asset Cluster Size Structure

While clustering offers a solution to the excessive data reduction problem, an additional question emerges in determining the sizes of clusters. Foreshadowed by the computational burden of the HY estimator, it is of practical need to develop an estimator which can be represented with a parsimonious number of clusters and by extension blocks. The performance of data-driven clustering is examined where size is determined using a simple hierarchical clustering algorithm. The K-means clustering algorithm, according to MacQueen (1967), is a heuristic method that divides the whole set of objects based on attributes into a predefined number (K) of clusters. The classification is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.[4]

The third simulation examines cluster size determination using the K-means algorithm. The simulation environment is the same as in Simulation 2, with $p = 256$, noise level $\gamma^2 = 0.375$, and heterogeneous observation setting. The number of clusters is restricted to four, where the size of these clusters is data driven using K-means. Data-driven clustering results in the illiquid clusters

[4] While the K-means algorithm can allocate the observations amongst the K groups, it lacks systematic guidance for determining the number of clusters. The choice of the number of clusters was determined using finite mixture modeling of Fraley and Raftery (2002). Based on the Bayesian information criterion, it was concluded that there are four or five different clusters, and motivates the choice of four clusters in the simulation study.

DOI: 10.1002/jae

Table V. Results for K-means clustering

Method	BLOCK			RnB		
	FRB	DIA	OFF	FRB	OFF	INV
Panel A: Entire matrix (1 : 256)						
Equal	2.060	0.144	1.453	2.044	1.441	1.517
K-means	1.996	0.152	1.407	2.026	1.428	1.503
Panel B: Upper half (1 : 128)						
Equal	0.872	0.094	0.613	0.846	0.594	1.202
K-means	0.951	0.104	0.668	0.928	0.652	1.327
Panel C: Upper quarter (1 : 64)						
Equal	0.478	0.076	0.334	0.478	0.334	1.229
K-means	0.576	0.091	0.402	0.576	0.402	1.808
Panel D: Lower half (129 : 256)						
Equal	1.758	0.181	1.236	1.673	1.175	1.563
K-means	1.723	0.187	1.211	1.657	1.164	1.498

Note: This table reports the scaled Frobenius norm of the covariance matrix (FRB) and inverse covariance matrix (INV) estimates, the scaled Euclidean norm of the diagonal elements of the estimate (DIA) and the scaled Euclidean norm of the vectorized off-diagonal elements of the estimate (OFF). Results are reported for the BLOCK and RnB estimators for $p = 256$, $\gamma^2 = 0.375$, and the heterogeneous observation arrival set. The number of clusters is fixed to 4.

becoming much smaller, whereas the liquid clusters become larger. In fact, the average cluster sizes from most liquid to least liquid are 109.3, 86.2, 24.3, and 18.2.

Table V reports the results of the K-means clustering for different subsets of the covariance matrix. The restriction to only four clusters allows comparison of the results and benchmarks the results against the naive equal cluster size analysis shown before. As in Simulation 2, the estimation gains are decomposed according to subsets of the entire matrix. At first glance, panel A suggests that clustering with respect to trade durations does not substantially reduce the estimation error compared to the case of equal cluster sizes. Panels B and D, with larger cluster sizes implied by the K-means algorithm, result in greater estimation errors for the liquid subsets. In contrast, the illiquid subset examined in panel C shows estimation error reduction. The error reduction is greater for the off-diagonal elements (OFF) than for the diagonal elements (DIA), further suggesting that the gains are being driven by improved estimates of the covariance elements including illiquid assets. The conclusion from this simulation is that the benefit of adding more clusters is largely driven by dividing the illiquid set into sequentially smaller sets.

5. EMPIRICAL DATA

5.1. Data

The empirical analysis is based on mid-quotes from the NYSE's Trade and Quote (TAQ) database for the constituents of the S&P 500.[5] The S&P 500 includes large-cap, actively traded US equities, and is diverse with respect to variation in liquidity and market microstructure effects. The sample period extends from 1 January 2007 to 1 April 2009, for a total of 562 trading days, and the daily transaction records extend from 9 : 45 until 16 : 00. The first 15 minutes of each day are intentionally omitted to avoid opening effects. The sample period covers the global financial crisis

[5] Qualitatively the S&P 500 has similar liquidity and market microstructure features as the S&P 1500 calibrated simulation and substantiates the study of the RnB estimator in this environment.

DOI: 10.1002/jae

following the collapse of Lehman Brothers Holding Inc. and includes both high- and low-volatility periods. The data are filtered eliminating obvious errors, such as bid prices greater than ask prices, non-positive bid or ask sizes, etc. Moreover, outliers are eliminated when the bid−ask spread is greater that 1% of the current mid-quote and when the mid-quote price does not change. Finally, two additional filters are employed, with both using a centered mean (excluding the observation under consideration) of 50 observations as a baseline. The first is a global filter deleting entries for which the mid-quote price deviates by more than 5 mean absolute deviations for the day. The second is a local filter deleting entries for which the mid-quote deviated by more than 5 mean absolute deviation of 50 observations (excluding the observation under consideration). See Barndorff-Nielsen *et al*. (2008) for a detailed discussion of data filtering and the implications for estimators.

5.2. Summary Statistics

Table VI presents annualized summary statistics for daily open-to-close (9 : 45−16 : 00) log returns of the S&P 500 stocks over the sample period. Summary statistics are computed for each stock and then the minimum, maximum, selected quantiles, and means for the entire index are reported. Panel A considers the entire sample period, Panel B covers only the sample period prior to the Lehman Brothers collapse on 14 September 2008, and Panel C is associated with the post-collapse sample period. The pre-collapse findings are consistent with the large empirical literature on asset returns, for instance, in Andersen *et al*. (2001), Aït-Sahalia and Mancini (2008) and Andersen *et al*. (2010). In all panels, stock returns display excess kurtosis. A greater average kurtosis in the entire sample suggests the occurrence of a structural break between the pre- and post-collapse intervals.

Table VII summarizes the annualized covariance estimates of the S&P 500 stocks using the RK and RnB estimators, employing four equal-sized blocks (henceforth RnB4) for the entire sample. On average, the RnB4 estimates have lower means and standard deviations.[6] All Ljung−Box portmanteau tests are well above the 40.289 critical value at 1% confidence level and strongly reject the null hypothesis of zero autocorrelations up to lag 22, corresponding to about 1 month

Table VI. Summary statistics for daily log-returns of S&P 500 stocks (as a percentage)

	Panel A: Full sample				*Panel B: Pre-collapse*				*Panel C: Post-collapse*			
	Mean	SD	Skew.	Kurt.	Mean	SD	Skew.	Kurt.	Mean	SD	Skew.	Kurt.
Min.	−1.230	2.392	−8.075	5.529	−0.776	1.309	−3.119	4.364	−3.825	3.595	−5.462	3.574
0.10	−0.311	3.252	−0.756	8.088	−0.188	2.294	−0.395	5.450	−0.927	5.161	−0.566	4.298
0.25	−0.179	3.920	−0.335	9.273	−0.086	2.680	−0.149	6.190	−0.507	6.226	−0.256	4.772
0.50	−0.060	4.967	0.055	10.929	−0.005	3.332	0.182	7.245	−0.206	7.830	0.024	5.489
0.75	0.023	6.467	0.372	13.335	0.076	4.190	0.499	9.240	−0.020	10.812	0.299	6.594
0.90	0.103	8.458	0.643	17.423	0.142	5.085	0.837	13.127	0.164	14.219	0.550	8.053
Max.	0.383	17.692	2.291	109.874	0.376	14.393	2.566	49.648	0.795	27.128	2.075	43.084
Mean	−0.094	5.915	−0.033	12.357	−0.016	3.849	0.205	8.494	−0.334	9.861	−0.006	6.030
SD	0.198	5.756	0.699	6.767	0.142	3.872	0.566	4.195	0.558	10.063	0.518	2.531

Note: This table reports summary statistics. The sample period extends from 3 January 2007 to 1 April 2009 for a total of 562 observations. Panel B: Pre-collapse period. The sample period extends from 3 January 2007 to 13 September 2008 for a total of 428 observations. Panel C: Post-collapse period. The sample period extends from 14 September 2008 to 1 April 2009 for a total of 134 observations.

[6] Pre- and post-collapse summary statistics are qualitatively the same and are not reported for the sake of space.

DOI: 10.1002/jae

Table VII. Summary statistics for the annualized covariance distribution as a percentage of S&P 500 stocks

	RK			*RnB*4		
	Mean	SD	Q_{22}	Mean	SD	Q_{22}
Min.	0.013	0.045	259	0.009	0.028	184
0.10	0.023	0.068	1015	0.016	0.043	1087
0.25	0.028	0.081	1261	0.020	0.053	1670
0.50	0.035	0.099	1545	0.025	0.064	2131
0.75	0.043	0.123	1793	0.030	0.080	2401
0.90	0.051	0.152	1967	0.036	0.098	2574
Max.	0.068	0.246	2454	0.047	0.152	2984
Mean	0.036	0.105	1507	0.026	0.068	1982
SD	0.011	0.033	382	0.008	0.021	577

Note: This table reports summary statistics of annualized covariance estimates based on the RK and RnB estimators. The sample period extends from 3 January 2007 to 1 April 2009 for a total of 562 observations. The table reports the Ljung–Box portmanteau test for up to 22nd-order autocorrelation, Q_{22} with 1% critical value of 40.289.

of trading days. Interestingly, Ljung–Box statistics are higher for RnB estimates than for RK estimates, suggesting that the RnB estimator provides estimates with more persistent temporal dependence. In line with recent results by Hansen and Lunde (2010) this finding suggests that RnB estimates are less noisy and thus better reveal the underlying high persistence of the process.

5.3. Estimating Portfolio Volatility

Following the procedure outlined in Briner and Connor (2008), the quality of competing estimators is evaluated according to their ability to predict the realized volatility of a (random) portfolio. Random portfolio weights are drawn from a uniform distribution $U(-0.5, 1.5)$ and scaled such that $\sum w = 1$. As in Bollerslev *et al*. (2008), the realized portfolio variance $\text{RCov}^m_{P_w,t}$ is computed from intra-day portfolio returns $r_{P_w,j,t} := \sum_{i=1}^{p} w_i r_{i,j,t},\ j = 1, \ldots, m$, sampled every 17.5 minutes ($m = 22$). As most investment decisions invariably involve the volatility over multiple days, overnight returns are incorporated into the covariance estimates using the scaling method presented in Hansen and Lunde (2005). Close-to-close returns (r_{CC}) are divided into two segments: overnight close-to-open returns (r_{CO}), and intra-day open-to-close returns (r_{OC}). Specifically, full-day covariances are constructed using

$$\hat{\Sigma}_{\text{CC}}(\omega) = \overline{\omega_1} r_{\text{CO}} r'_{\text{CO}} + \overline{\omega_2}\hat{\Sigma}_{\text{OC}}, \text{ where } \omega = (\overline{\omega_1} + \overline{\omega_2})' = \left(\sum_{j=1}^{p} \omega_{1,j} + \sum_{j=1}^{p} \omega_{2,j}\right)'$$

In a univariate setting, Hansen and Lunde (2005) propose choosing the weights by minimizing the mean squared error with respect to the underlying ('true') quadratic variation and show that the scaling factors can be computed as $\omega_{1,j} = (1-\phi)\frac{\mu_0}{\mu_1}$ and $\omega_{2,j} = \phi\frac{\mu_0}{\mu_2}$, with $\mu_0 = E(IV_t)$, $\mu_1 = E(r^2_{CO,t})$, $\mu_2 = E(\text{RV}_{\text{OC},t})$ with ϕ denoting the relative importance factor. As shown by Hansen and Lunde, the quantities μ_0, μ_1, μ_2 and ϕ can be estimated using sample averages of estimated quadratic variations and close-to-open returns. To apply this procedure in a multivariate setting and to ensure positive definiteness of the resulting matrix, we use the multivariate adjustment suggested in Chiriac and Voev (2010), where $\overline{\omega_1}$ and $\overline{\omega_2}$ are the cross-sectional average scaling factor of the variances.

DOI: 10.1002/jae

The estimates are evaluated within a Mincer and Zarnowitz (1969) framework. The regression is specified as

$$\sqrt{\frac{\pi}{2}} \times \widehat{\sigma}_{t,w} = \alpha_0 + \alpha_1 \sqrt{(w' \widehat{\Sigma}_{(1),t} w)} + \alpha_2 \sqrt{(w' \widehat{\Sigma}_{(2),t} w)}$$

where the realized portfolio volatility is $\widehat{\sigma}_{t,w} = (\mathrm{RCov}^{m}_{P_w,t} + w'(r_{\mathrm{CO},t} r'_{\mathrm{CO},t})w)^{1/2}$ and $\widehat{\Sigma}_{(j),t},\ j = \{1, 2\}$ are competing covariance estimates. The scaling suggested by Patton and Sheppard (2008) imposes an assumption of conditionally normally distributed portfolio returns, but allows for time-varying variances.[7]

Table VIII shows the Mincer–Zarnowitz forecast evaluation regression results. Newey and West (1987) robust standard errors are reported, where the bandwidths are determined following the procedure outlined in Newey and West (1994). The coefficient α_0 is a measure of regression bias, whereas α_1 and α_2 are measures of regression efficiency.

The performance of different versions of RnB estimators are compared against the original RK estimator as well as a regularized version of the RK estimator (RRK). The latter is included to assess the impact of regularization solely (without blocking). Hence the set of estimators considered includes RRK (equivalent to RnB1), RnB2, RnB4, and RnB8, where the digits indicate the number of underlying equally sized clusters. Panel A of Table VIII gives results of single estimates evaluated on the basis of the null hypothesis, $\alpha_0 = 0$ and $\alpha_1 = 1$. It is shown that in terms of R^2 the RnB estimates are more accurate than the RK estimates. The regression coefficient α_0 is statistically insignificant. The regression coefficients α_1 are closer to 1 for RnB estimates than for RK estimates, demonstrating that RnB estimates are also more efficient. Panel B gives results of encompassing regressions requiring $\alpha_0 = 0$ and $\alpha_1 + \alpha_2 = 1$ if the two competing forecasts are jointly unbiased and efficient. A forecast is encompassed when its coefficient is not statistically different from zero. It is shown that the RK estimates are encompassed by RnB estimates as the

Table VIII. Mincer–Zarnowitz regression evaluations

$\widehat{\Sigma}_{(1)}$	$\widehat{\Sigma}_{(2)}$	α_0	α_1	α_2	R^2
Panel A: Single regressions					
RK		0.000	0.742		0.84
		(0.000)	(0.013)		
RRK		0.000	0.743		0.84
		(0.000)	(0.013)		
RnB2		0.000	0.818		0.86
		(0.000)	(0.013)		
RnB4		0.000	0.853		0.87
		(0.000)	(0.015)		
RnB8		0.000	0.859		0.87
		(0.000)	(0.015)		
Panel B: Encompassing regressions					
RnB2	RK	0.000	0.960	−0.130	0.86
		(0.000)	(0.326)	(0.298)	
RnB2	RRK	0.000	0.917	−0.090	0.86
		(0.000)	(0.248)	(0.227)	

Note: This table reports Mincer–Zarnowitz regression. Random portfolios are generated from all available constituents of the S&P 500 from 3 January 2007 to 1 April 2009. Newey–West robust standard errors are reported in parentheses below.

[7] The relative performance of the estimator is robust to various return volatility scalings as well as the log transformation.

DOI: 10.1002/jae

coefficient α_2 is never statistically significant. This is consistent with panel A and confirms that at this dimension RK is inferior to the RnB estimates.

Studying the impact of the number of underlying blocks, we observe performance improvements with an increasing number of clusters. The α_1 estimate is closer to one as the number of clusters increases. While we observe clear efficiency gains when increasing the number of clusters from one to four, there are only marginal differences between RnB4 and RnB8. This result is also confirmed by the corresponding R^2 statistics. These findings are in line with the simulation results presented above and demonstrate that most of the efficiency gains essentially stem from separating between liquid and illiquid assets. This suggests using a low number of categories, which keeps the estimator parsimonious and does not require a strong regularization in the second step.

6. CONCLUSIONS

This paper introduces a regularization and blocking (RnB) estimator for vast-dimensional covariance estimation. The estimator limits data loss due to asynchroneity by grouping assets together according to liquidity and estimating a series of covariance blocks using the realized kernel (RK) estimator introduced by Barndorff-Nielsen *et al.* (2010). These blocks are combined to form a complete covariance matrix, which is then regularized using eigenvalue cleaning, as introduced by Laloux *et al.* (1999), guaranteeing positive definite and well-conditioned covariance matrix estimates.

The performance of the RnB estimator is analyzed within an extensive simulation study designed to mimic the empirical features of the S&P 1500 universe. The RnB estimator shows significant gains compared to the realized kernel estimator, especially in settings with high dimensionality and heterogeneous observation frequencies of individual assets. Moreover, most of the efficiency gains can be captured with a parsimonious number of clusters. Finally, switching from equal-sized clusters to cluster sizes arising from a K-means algorithm based on trade-to-trade durations yields further reductions of estimation errors. Applying the RnB estimator to predict the volatility of portfolios composed of S&P 500 constituents shows significant performance gains over the realized kernel estimator and (a regularized version) of the realized covariance estimator.

The empirical results show that the new estimator is useful whenever high-dimensional covariances over short time horizons have to be estimated with preferably high precision. Furthermore, the comparison of the performance of RnB estimators with varying asset clusters suggests that most of the efficiency gains stem from separating between liquid and less liquid assets.

Overall, this paper shows that the idea of combining high-frequency-based covariance estimators with blocking and regularization techniques is promising and opens up new lines of research to estimate and forecast vast-dimensional covariances. In a recent study, Hautsch and Kyj (2010) show that the idea of blocking covariances does not only lead to more efficient estimates but does also allow for better factor extraction, which can be exploited to improve out-of-sample covariance forecasts.

ACKNOWLEDGEMENTS

For helpful comments and discussions we thank Torben Andersen, Tim Bollerslev, René Garcia, Peter Reinhard Hansen, Wolfgang Härdle, Hartmuth Henkel, Christian Hesse, Asger Lunde, Nour Meddahi, Markus Reiss, Jeffrey Russell, Neil Shephard and the participants of the 2009 Humboldt–Copenhagen Conference on Financial Econometrics, the 2009 CREATES Conference on Financial Econometrics and Statistics, the 2009 SoFiE conference, the 2009 EC^2 Conference,

DOI: 10.1002/jae

the 2009 European Meeting of the Econometric Society, the 2010 Conference on Volatility and Systemic Risk at NYU Stern School of Business, as well as of seminars at Université Libre de Bruxelles, University of Rotterdam, Singapore Management University, University of Technology, Sydney, and University of Melbourne. This research is supported by the Deutsche Bank AG via the Quantitative Products Laboratory and the Deutsche Forschungsgemeinschaft via the Collaborative Research Center 649 'Economic Risk'. This work was done while Kyj was at Quantitative Products Laboratory and the views expressed are strictly those of Kyj and not necessarily of Barclays Capital.

REFERENCES

Aït-Sahalia Y, Mancini L. 2008. Out of sample forecasts of quadratic variation. *Journal of Econometrics* **147**: 17–33.

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

Andersen T, Bollerslev T, Frederiksen P, Nielsen M. 2010. Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns. *Journal of Applied Econometrics* **25**: 233–261.

Bandi F, Russell J. 2006. Separating microstructure noise from volatility. *Journal of Financial Economics* **79**: 655–692.

Bandi F, Russell J, Zhu Y. 2008. Using high-frequency data in dynamic portfolio choice. *Econometric Reviews* **27**: 163–198.

Barndorff-Nielsen O, Shephard N. 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* **64**: 253–280.

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

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

Barndorff-Nielsen O, Shephard N. 2005. Measuring the impact of jumps in multivariate price processes using bipower variation. Working paper, University of Oxford.

Barndorff-Nielsen O, Hansen P, Lunde A, Shephard N. 2008. Realised kernels in practice: trades and quotes. *Econometrics Journal* **4**: 1–32.

Barndorff-Nielsen O, Hansen P, Lunde A, Shephard N. 2010. Multivariate realized kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. Journal of Econometrics (forthcoming). (Web appendix available at http://www.hha.dk/∼alunde/BNHLS/PDF/WEBAPPENDIX_26_6_08.PDF.

Bickel PJ, Levina E. 2008. Regularized estimation of large covariance matrices. *Annals of Statistics* **36**: 199–227.

Bollerslev T, Law T, Tauchen G. 2008. Risks, jumps and diversification. *Journal of Econometrics* **144**: 234–256.

Bonato M, Caporin M, Ranaldo A. 2009. Forecasting realized (co)variances with a block structure Wishart autoregressive model. Working paper no. 2009-3, Swiss National Bank.

Briner B, Connor G. 2008. How much structure is best? A comparison of market model, factor model, and unstructured equity covariance matrices. *Journal of Risk* **10**: 3–30.

Chan L, Karceski J, Lakonishok J. 1999. On portfolio optimization: forecasting covariances and choosing the risk model. *Review of Financial Studies* **12**: 937–974.

Chiriac R, Voev V. 2010. Modelling and forecasting multivariate realized volatility. *Journal of Applied Econometrics* (forthcoming).

Duffie D, Pan J. 1997. An overview of value at risk. *Journal of Derivatives* **4**: 7–49.

Epps T. 1979. Comovement in stock prices in the very short run. *Journal of the American Statistical Association* **74**: 291–298.

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**: 611–631.

Griffin J, Oomen R. 2010. Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Journal of Econometrics* (forthcoming).

Hansen PR, Lunde A. 2005. A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics* **3**: 525–554.

DOI: 10.1002/jae

Hansen PR, Lunde A. 2006. Realized variance and market microstructure noise. *Journal of Business and Economic Statistics* **24**: 127–161.
Hansen PR, Lunde A. 2010. Estimating the persistence and the autocorrelation function of a time series that is measured with error. Working paper, Stanford University, CA.
Harris F, McInish T, Shoesmith G, Wood R. 1995. Cointegration, error correction and price discovery on informationally-linked security markets. *Journal of Financial and Quantitative Analysis* **30**: 563–581.
Hautsch N, Kyj LM. 2010. Forecasting vast dimensional covariances using a dynamic multi-scale realized spectral components model. Working paper, Humboldt-Universität, Berlin.
Hayashi T, Yoshida N. 2005. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* **11**: 359–379.
Jagannathan R, Ma T. 2003. Risk reduction in large portfolios: why imposing the wrong constraints helps. *Journal of Finance* **58**: 1651–1683.
Laloux L, Cizeau P, Bouchaud J-P, Potters M. 1999. Noise dressing of financial correlation matrices. *Physical Review Letters* **83**: 1467–1470.
Ledoit O, Wolf M. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10**: 603–621.
Ledoit O, Wolf M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**: 365–411.
MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Cam L, Neyman J (eds). University of California Press: Berkeley, CA; 281–297.
Martens M. 2006. Estimating unbiased and precise realized covariances. Working paper, Erasmus University, Rotterdam.
Michaud RO. 1989. The Markowitz optimization enigma: is optimized optimal? *Financial Analysts Journal* **45**: 31–42.
Mincer J, Zarnowitz V. 1969. The evaluation of economic forecasts. In *Economic Forecasts and Expectations*, Mincer J (ed.). Columbia University Press: New York; 3–46.
Newey W, West K. 1987. A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**: 703–708.
Newey W, West K. 1994. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies* **61**: 631–653.
Onatski A. 2009. Testing hypotheses about the number of factors in large factor models. *Econometrica* **77**(5): 1447–1479.
Oomen R. 2006. Comment to realized variance and market microstructure noise. *Journal of Business and Economic Statistics* **24**: 195–202.
Oomen R. 2009. A universal scaling law of noise in financial markets. Working paper, University of Amsterdam.
Patton A, Sheppard K. 2008. Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series*, Andersen TG, Davis RA, Kreiss JP, Mikosch T. Springer: Berlin; 801–828.
Qi H, Sun D. 2006. A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM Journal of Matrix Analysis and Applications* **28**: 360–385.
Sheppard K. 2006. Realized covariance and scrambling. Working paper, Oxford University.
Stein C. 1977. Lectures on the theory of estimation of many parameters. In *Studies in the Statistical Theory of Estimation*, Part 1, Ibragimov I, Nikulin M (eds). *Proceedings of Scientific Seminars of the Steklov Institute*, no. 74; 4–65.
Tola V, Lillo F, Gallegati M, Mantegna R. 2008. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control* **32**: 235–258.
Voev V, Lunde A. 2007. Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics* **5**: 68–104.
Wang Y, Zou J. 2010. Vast volatility matrix estimation for high-frequency financial data. *Annals of Statistics* **2**: 943–978.
Zhang L, Mykland P, Ait-Sahalia Y. 2005. A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* **100**: 1394–1411.
Zumbach G. 2009. The empirical properties of large covariance matrices. *RiskMetrics Journal* **9**: 31–54.

DOI: 10.1002/jae

APPENDIX

The simulated return process follows a three-factor model:

$$x_i = F\beta_i + \varepsilon, \text{ where } \varepsilon \sim N(0, \Psi)$$

where F is a $(p \times 3)$ matrix of factors, x_i is a $(p \times 1)$ vector of returns at time i, and β is a (3×1) vector of factor loadings, which are themselves random variables with the following distribution:

$$\beta \sim N\left(\begin{bmatrix} 0.5 \\ 0 \\ -0.1 \end{bmatrix}, \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix}\right)$$

The covariance matrix of such a process can be written as

$$\Phi = \beta\beta' + \Psi$$

where it is assumed that all the correlation between processes is captured by the factors and as a result the off-diagonal elements of Ψ are all set to 0.

The goodness-of-fit of the model, R^2, follows a Beta(1, 1) distribution.[8] Solve for the error to add to the factor model:

$$\Psi = \frac{\text{diag}(\beta\beta')}{R^2} - \text{diag}(\beta\beta')$$

Finally, take the correlation matrix of Φ such that $\Sigma = \text{Corr}[\Phi]$ to ensure that the variances are equal and control for perceived errors in covariance estimation due to unequal variances.

[8] Beta(1, 1) distribution was chosen as it reflects the goodness-of-fit of a three-factor model of the S&P 1500 in 2008.

DOI: 10.1002/jae