Hautsch, Nikolaus; Kyj, Lada M.; Hautsch, Nikolaus

Working Paper

A blocking and regularization approach to high dimensional realized covariance estimation

CFS Working Paper, No. 2009/20

Provided in Cooperation with:
Center for Financial Studies (CFS), Goethe University Frankfurt

Suggested Citation: Hautsch, Nikolaus; Kyj, Lada M.; Hautsch, Nikolaus (2009) : A blocking and regularization approach to high dimensional realized covariance estimation, CFS Working Paper, No. 2009/20, Goethe University Frankfurt, Center for Financial Studies (CFS), Frankfurt a. M., https://nbn-resolving.de/urn:nbn:de:hebis:30-72694

This Version is available at:
https://hdl.handle.net/10419/43197

No. 2009/20

A Blocking and Regularization Approach to High Dimensional Realized Covariance Estimation

Nikolaus Hautsch, Lada M. Kyj,
and Roel C.A. Oomen

Center for Financial Studies
Goethe-Universität Frankfurt ■ House of Finance
Grüneburgplatz 1 ■ 60323 Frankfurt ■ Deutschland

Telefon: +49 (0)69 798-30050
Fax: +49 (0)69 798-30077
http://www.ifk-cfs.de ■ E-Mail: info@ifk-cfs.de

Center for Financial Studies

The *Center for Financial Studies* is a nonprofit research organization, supported by an association of more than 120 banks, insurance companies, industrial corporations and public institutions. Established in 1968 and closely affiliated with the University of Frankfurt, it provides a strong link between the financial community and academia.

The CFS Working Paper Series presents the result of scientific research on selected topics in the field of money, banking and finance. The authors were either participants in the Center´s Research Fellow Program or members of one of the Center´s Research Projects.

If you would like to know more about the *Center for Financial Studies*, please let us know of your interest.

Prof. Dr. Jan Pieter Krahnen

Center for Financial Studies
Goethe-Universität ■ House of Finance
Grüneburgplatz 1 ■ 60323 Frankfurt am Main ■ Deutschland

Telefon: +49 (0)69 798-30050
Fax: +49 (0)69 798-30077
http://www.ifk-cfs.de ■ E-Mail: info@ifk-cfs.de

CFS Working Paper No. 2009/20

A Blocking and Regularization Approach to High Dimensional Realized Covariance Estimation*

Nikolaus Hautsch[1], Lada M. Kyj[2], and Roel C.A. Oomen[3]

October 2009

Abstract:
We introduce a regularization and blocking estimator for well-conditioned high-dimensional daily covariances using high-frequency data. Using the Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a) kernel estimator, we estimate the covariance matrix block-wise and regularize it. A data-driven grouping of assets of similar trading frequency ensures the reduction of data loss due to refresh time sampling. In an extensive simulation study mimicking the empirical features of the S&P 1500 universe we show that the 'RnB' estimator yields efficiency gains and outperforms competing kernel estimators for varying liquidity settings, noise-to-signal ratios, and dimensions. An empirical application of forecasting daily covariances of the S&P 500 index confirms the simulation results.

JEL-Classifications: C14, C22

Keywords: Covariance Estimation, Blocking, Realized Kernel, Regularization, Microstructure, Asynchronous Trading

* For helpful comments and discussions we thank Torben Andersen, Tim Bollerslev, Ren´e Garcia, Wolfgang H¨ardle, 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 as well as the 2009 European Meeting of the Econometric Society. 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".

1 Hautsch is with Institute for Statistics and Econometrics and Center for Applied Statistics and Economics, Humboldt-Universit¨at zu Berlin as well as Quantitative Products Laboratory, Berlin, and Center for Financial Studies, Frankfurt. Email: nikolaus.hautsch@wiwi.hu-berlin.de.

2 Kyj is with the Quantitative Products Laboratory, Berlin and School of Business and Economics, Humboldt-Universit¨at zu Berlin Email: lada.kyj@wiwi.hu-berlin.de.

3 Oomen is with Deutsche Bank, London and affiliated with Department of Quantitative Economics at the University of Amsterdam, The Netherlands. Email: roel.ca.oomen@gmail.com.

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, Karceski, and Lakonishok, 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 ever increasing horizon is needed to merely ensure positive definiteness of the sample covariance matrix. Since in many applications not only covariances but also the inverses thereof are required, positive definiteness (and well-conditioning) of covariance estimates are necessary properties. Furthermore, Jagannathan and Ma (2003) and Ledoit and Wolf (2003), using daily data, show that conditioning the covariance estimate does also translate into better out-of-sample portfolio risk management for monthly investment horizons. However, today's practitioners often need to manage their risk over much shorter time horizons. Risk measures are required to accurately reflect the risk of trading portfolios over typically a day. The availability of high-frequency asset price data opens up alternative ways of efficiently estimating short-term high-dimensional covariances.

In this paper, a vast-dimensional covariance estimator is constructed using high-frequency data in a manner which addresses market microstructure noise and asynchronous trading effects while preserving consistency, positive definiteness as well as well-conditioning of the covariance matrix. 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. The performance of the resulting regularization and blocking – henceforth "RnB" – estimator is examined within a data-driven simulation setting mimicking the market microstructure and liquidity features of the constituents of the S&P 1500 in 2008. Within this environment, variations in dimension of the covariance matrix, market microstructure effects, and liquidity characteristics are considered. It turns out that the RnB estimator reduces estimation error in every scenario considered. Finally, an empirical application to the forecasting of S&P 500 portfolio volatility illustrates the superior performance of the RnB estimator.

There exists a large body of literature pertaining to realized covariance estimation. The foundations of high frequency covariance estimation are developed in Andersen, Bollerslev, Diebold, and Labys (2001), Andersen, Bollerslev, Diebold,

and Labys (2003) and Barndorff-Nielsen and Shephard (2004). The realized covariance estimator is defined as the cumulative sum of the cross-products of multivariate returns synchronized in calendar time (e.g., every 5 minutes). In addition to the induced efficiency loss due to sparse sampling this estimator becomes ill-conditioned (in the extreme case not positive definite) as the number of cross-sectional dimensions is high relative to the number of intra-day sampling intervals. If on the other hand the sampling frequency is increased, covariance estimates are dominated by market microstructure effects such as bid-ask bounce, price discreteness, and non-synchroneity of price observations. In particular, sampling too frequently results in an over-estimation of the variance elements due to the accumulation of market microstructure noise (see for instance, Zhang, Mykland, and Ait-Sahalia (2005), Hansen and Lunde (2006), and Bandi and Russell (2006)) whereas the covariance elements are under-estimated due to non-synchronous trading effects (Epps, 1979).

A number of recent paper 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 can be free of any biases. Bandi, Russell, and Zhu (2008), Griffin and Oomen (2009), Martens (2006), Sheppard (2006), and Voev and Lunde (2007) study numerous alternative estimators in a bi-variate setting via optimal sampling or lead-lag estimation to obtain substantial efficiency gains. Most recently, Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008a, hereafter BNHLS) introduce a multivariate realized kernel (RK) estimator which is shown to be consistent in the presence of market microstructure noise and is guaranteed to be positive semi-definite. The RK estimator is a HAC-type estimator composed of a kernel-weighted sum of autocovariance matrices. The choice of kernel ensures the positive semi-definiteness of the resulting estimate. A 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.

Having identified observation synchronization and data preservation as essential components of efficient high-dimensional covariance estimation, the main contribution of this paper is to construct a covariance matrix from a series of sub-sets (blocks) of the matrix. The blocks are composed of asset clusters 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 time scales. Applying the BNHLS RK estimator to individual blocks retains a greater amount of data post-synchronization

and increases the precision of the corresponding estimates compared to an 'all-in-one approach'. However, 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.

To evaluate the performance of the proposed RnB estimator, an extensive simulation study is conducted closely mimicking the empirical features of the S&P 1500 index. In this context market microstructure noise effects as well as observation frequencies are calibrated with respect to the cross-section of the complete S&P1500 universe. The simulation study examines the effects of (i) blocking and regularization, (ii) the number of clusters and (iii) cluster size determination based on different observation distributions and magnitudes of market microstructure noise. It is shown that blocking universally reduces the estimation error with the greatest gain achieved in settings where the cross-sectional variation in observation frequency is large. Moreover, clustering assets into a moderate number of groups isolates illiquid assets from liquid assets and results in improved estimation via blocking. Estimation errors can be further reduced by a data-driven choice of the cluster sizes. Finally, the RnB estimator is applied to estimate daily covariances of the S&P 500 index from January 2007 to April 2009. In a Mincer and Zarnowitz (1969) style forecasting regression the estimator's performance is evaluated with respect to predicting the absolute returns of randomized portfolios. It is shown that the new estimator significantly outperforms competing covariance estimators.

The remainder of the paper is organized as follows: In Section 2, the underlying theoretical setting is presented. Section 3 introduces the used blocking and regularization techniques, whereas Section 4 illustrates the simulation setup. 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 $i - th$ 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 $X^{(i)}(t_j^{(i)})$, 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

$$Y(t) = \int_0^t a(u)du + \int_0^t \sigma(u)dW(u), \tag{1}$$

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:

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

where $U_j^{(i)}$ is covariance stationary and satisfies the following 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. $[Y] = \int_0^1 \Sigma(u)du$ where $\Sigma = \sigma\sigma'$, 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 addresses market microstructure effects and asynchronous price observations while guaranteeing consistency and positive semi-definiteness. RK estimation is a two part process. As in Harris, McInish, Shoesmith, and Wood (1995), the observations are synchronized via refresh time sampling (RTS) as illustrated in Figure 1. 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 time scale. More formally, the first refresh time sampling point is defined as $RFT_1 = \max(t_1^{(1)}, \ldots, t_1^{(p)})$ and $RFT_{j+1} = \text{argmin}(t_{k_i}^{(i)} | t_{k_i}^{(i)} > RFT_j, \forall\, i > 1)$. Refresh time synchronization allows us to define high frequency vector returns as $x_j = X_{RFT_j} - X_{RFT_{j-1}}$, where $j = 1, 2, \ldots, n$, and n is the number of refresh time observations.

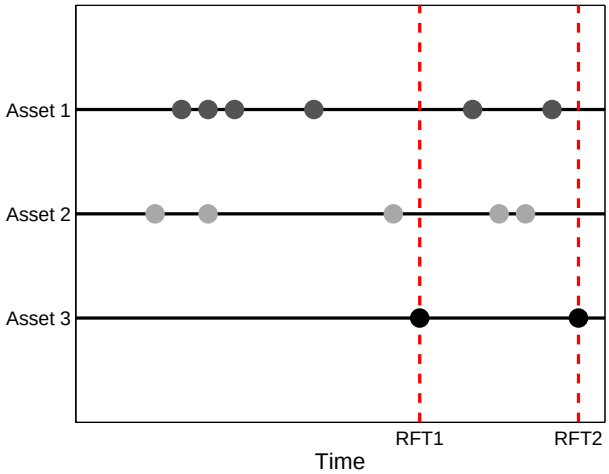
The multivariate realized kernel is defined as

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

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

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

Figure 1: Illustration of the refresh time sampling scheme



Note. This figure illustrates the refresh time sampling scheme when applied to three assets. The solid circles indicate the timing of observations. The dashed vertical lines indicate the refresh time sampling points.

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 the corresponding block. The fact that a global bandwidth may be sub-optimal for a very diverse set of bandwidths is another motivation for grouping similar assets together. For a more detailed discussion of bandwidth selection, see the web appendix of BNHLS.

The RK estimator is related to the heteroskedasticity and autocorrelation consistent (HAC) covariance estimators of Newey and West (1987) and Andrews (1991). Similar to the optimal sampling frequencies derived in Zhang et al. (2005) and Bandi and Russell (2006), the bandwidth parameter is a function of the noise-to-signal ratio. This draws upon the properties of lead-lag estimators, which help filtering out distortions due to market microstructure effects. As noise increases relative to the signal, the bandwidth is increased and more lags of the autocovariance are considered. In the absence of noise there are no autocovariance lags in the estimator and hence it defaults to the realized covariance estimator. A drawback of the kernel structure is that it converges at a rate of $n^{1/5}$, which is slower than the optimal rate of $n^{1/4}$ for realized covariance estimators (see Gloter and Jacod, 2001; Kinnebrock and Podolskij, 2008).

3 The RnB estimator

3.1 Motivation

As illustrated in Figure 1, RTS may make inefficient use of data. In high dimensional covariance estimation, this 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 $\mathcal{M}(p) = \mathrm{E}[\max(t_1^{(1)}, t_1^{(2)}, ..., t_1^{(p)})]$ as the expected maximum waiting time for all assets to have traded at least once. Then, using the fact that $\Pr[\max(t_1^{(1)}, t_1^{(2)}, ..., t_1^{(p)}) < u] = (1 - e^{-\beta u})^p$, $\mathcal{M}(p)$ can be computed as

$$\mathcal{M}(p) = \frac{1}{\beta} \int_0^\infty p \left(1 - e^{-u}\right)^{p-1} e^{-u} u du, \tag{5}$$

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 Panel A of Figure 2 plots the relationship between $\mathcal{L}(p)$ and p, implying, e.g., 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 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 $\Pr[\max(t_1^{(1)}, t_1^{(2)}, ..., t_1^{(p_1)}, t_1^{(p_1+1)}, ..., t_1^{(p_1+p_2)}) < u] = (1 - e^{-\beta_1 u})^{p_1} (1 - e^{-\beta_2 u})^{p_2}$. The dashed gray line in Figure 2 Panel A represents the data loss for the most active asset in the scenario where $p_1 = p_2$ and $\beta_2 = 5\beta_1$. It is shown that variation in arrival rates further increases the implied data loss.

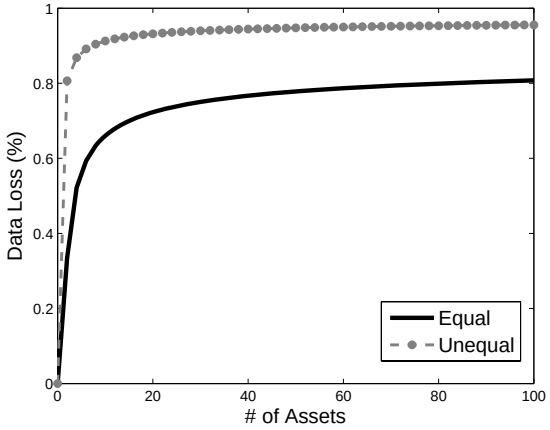
3.2 Blocking strategy

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 right corner. This initial step ensures that subsequent blocks will group together assets with similar arrival rates. Each block is itself a covariance matrix.

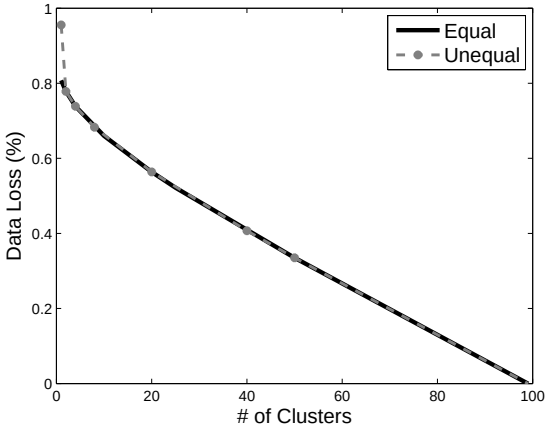
Figure 3 illustrates the construction of the so-called BLOCK estimator with three equal-sized asset clusters. The six resulting covariance blocks, each with a different RTS time scale, combine to form this multi-block estimator. Block 1

Figure 2: Illustration of the data loss implied by the refresh time sampling scheme

Panel A: data loss by number of assets

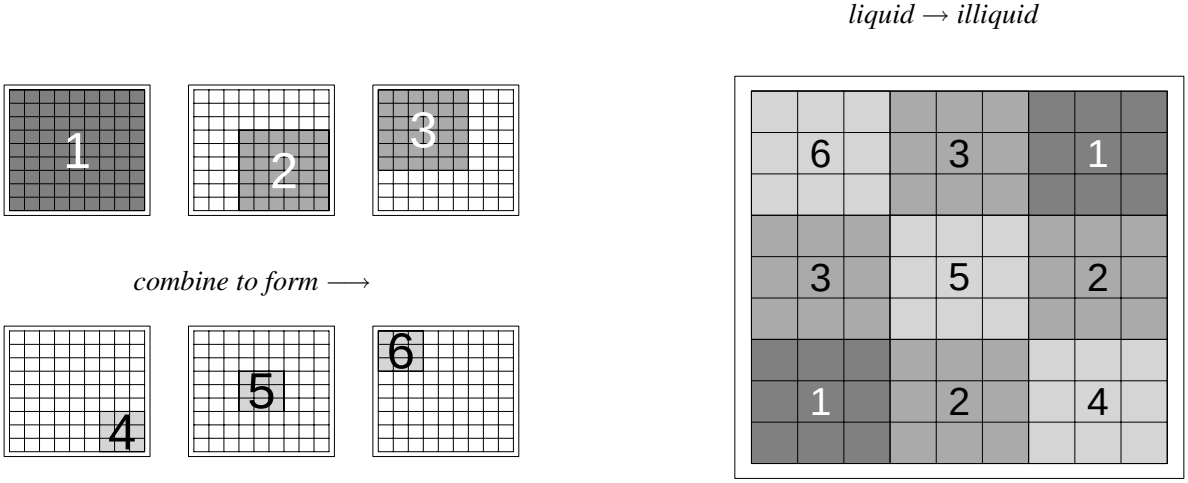


Panel B: data loss by number of clusters



Note. Panel A (B) reports the percentage of data loss as the number of assets (clusters) in the Refresh Time Sampling time scale 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 Panel B, the number of assets is equal to 100.

Figure 3: Visualization of the blocking strategy



Note. 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.

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. Finally, estimates for blocks 4, 5, and 6, composing 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 9-asset realized kernel, the middle off-diagonal blocks (2) and (3) stem from the 6-asset realized kernels, and the diagonal blocks (4), (5), and (6) are from the corresponding 3-asset RKs.

Grouping assets according to their trading frequency directly addresses the data reduction problem. Hence, 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-time-scale design substantially increases the effective number of observations used without imposing any additional structure on the covariance estimate.

As a result, each resulting block has an individual RTS time scale, 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

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

Figure 2 Panel 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 first derivative 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., 4 or 5. Finally, the impact of blocking is even 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, Caporin, and Ranaldo (2008) and Disatnik (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 approach, both of these methods by no means guarantee efficient use of data.

3.3 Regularization

While our proposed BLOCK estimator improves estimation precision it is done at the expense of positive semi-definiteness. This necessitates the consideration of regularization techniques which yield positive definite and well-conditioned covariance estimates. Regularization procedures help form meaningful approximate solutions of ill-conditioned or singular covariance matrices, see Neumaier (1998) for a discussion of regularization. Ill-conditioned matrices are characterized by eigenvalues vanishing to zero, behave similar to numerically singular matrices and result in unstable matrix inversions. Hence, our regularization objective is two-fold. First, covariance matrices must be non-singular, and, second, ensure that these are numerically stable. There are many regularization techniques that can be applied to covariance estimates (see Ledoit and Wolf (2004), Qi and Sun (2006) or Bickel and Levina (2008)). In this study we employ "Eigenvalue Cleaning", a regularization technique introduced by Laloux, Cizeau, Bouchaud, and Potters (1999), and further developed in Tola, Lillo, Gallegati, and Mantegna (2008). Applications of Random Matrix Theory have emerged as a common regularization technique for high dimensional realized covariance matrices (see also Onatski (2009), Wang and Zou (2009), 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 focus is on the correlation matrix R with spectral decomposition $R = Q\Lambda Q'$, where Q is the matrix of eigenvectors and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of eigenvalues. Under the null hypothesis of independent assets, the correlation matrix R is the identity matrix. Under this hypothesis, the distribution of eigenvalues is given by the Marchenko Pastur distribution with maximum eigenvalue given by $\lambda_{max} = \sigma^2(1 + \frac{1}{q} + 2\sqrt{\frac{1}{q}})$, where σ^2 is the variance of the entire portfolio equal to one in case of a correlation matrix.

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. Suppose 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 λ_{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) λ_{max} are identified as further "signals". On the other hand, eigenvalues smaller than this threshold are identified as eigenvalues to be driven by noise 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,

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

Hence, δ is determined as the ratio of the trace of the positive semi-definite projection of the correlation matrix R_+ minus the sum of the estimated eigenvalues which exceed the Marchenko Pastur threshold over the number of dimensions that fail to exceed the threshold. This results in a matrix $\hat{R} = Q\tilde{\Lambda}Q'$, where $\tilde{\Lambda} = diag(\tilde{\lambda}_i)$. Finally, the RnB estimator is defined as the corresponding covariance constructed from $\hat{R}$.

The Marchenko Pastur density gives the limit of the largest eigenvalue, whereas the Tracy-Widom distribution, as discussed in Johnstone (2001), gives the distribution of the largest eigenvalue as a function of number of dimensions and observations. The simpler threshold value obtained from the Marchenko Pastur distribution will overestimate the number of signals but is used for the following two reasons. First, recall that the objective of eigenvalue cleaning is to regularize vanishing eigenvalues, and not to fit a principal component model. Hence the focus is on addressing the bottom tail of the eigenvalue distribution. Second, a feature of the BLOCK estimator is that it has a different set of RTS observations for each block. The estimation error associated with approximating N may result in large modifications to the centering and scaling constants and render the formal hypothesis testing using the Tracy-Widom distributions problematic.

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 as the condition number of the matrix, $\kappa(A) = \left|\frac{\lambda_{max}}{\lambda_{min}}\right|$, is greater than 10*$p$.

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.,

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

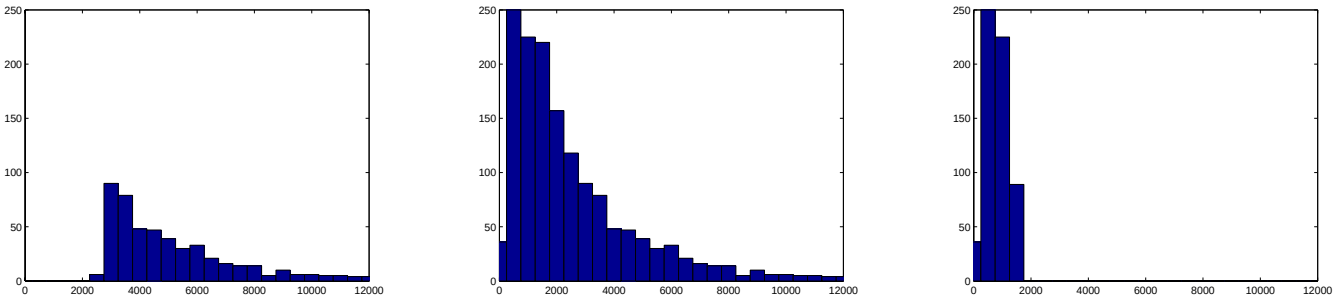
where $\Theta^{'}$ is the Cholesky factorization of the covariance matrix such that $\Theta\Theta^{'} = \Sigma$, and Z is a $(p \times 1)$ vector of independent standard Brownian motions. To simulate the process, we use an Euler discretization scheme with step size $\Delta = 1/23400$. The covariance structure is generated from an ad hoc statistical three-factor model that closely mimics the cross-sectional distribution of correlations for the S&P1500 universe. 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&P6 00 indices. Then, arrival times are modeled by uniformly sampling $M^{(i)}$ observations for asset i from $[0, 1]$. Figure 4 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.

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

Figure 4: Liquidity classification by observation frequency

Panel A: liquid Panel B: heterogeneous Panel C: illiquid



Note. Panels A, B, and C show the distribution of number of observations from the top 500 assets, the entire sample, and the bottom 600 assets of the S&P 1500 universe when ordered by number of observations.

Table 1: 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 January 2, 2008 through December 31, 2008. The index constituent lists are from January 2009. Assets are grouped according to liquidity characteristics into Illiquid 600, Middle 400, and Liquid 500.

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. Table1 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 fact, 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.[1] Note that portfolios

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

of this high dimension size have rarely been studied in the realized covariance literature. A notable exception is Wang and Zou (2009) 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.

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

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

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.,

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

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 2 and 3 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 (PSD). The estimates considered are the multivariate realized kernel (RK), and the blocking estimator based on 4 clusters of equal size (BLOCK) together with regularized versions thereof using eigenvalue cleaning (henceforth RRK and RnB, respectively). All criteria are evaluated under varying noise levels, observation arrival structures, and dimension sizes.

size equality across clusters.

Tables 2 and 3 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 3 shows that for higher dimensions the PSD statistic is virtually zero for all RK estimates in the heterogeneous and illiquid settings. The illiquid setting has only 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 results suggest that at this dimension RTS results in fewer than p observations. On the other hand, the BLOCK estimator is never positive definite at this dimension, whereas RnB is always positive definite by construction. The additional reduction between the unregularized and regularized statistics suggests that by imposing structure via regularization estimation error in high dimensional systems can be mitigated. 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 results 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.[2]

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

[2]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.

Table 2: Performance of realized kernel and blocking estimators for $p = 64$ and 4 asset clusters

	unregularized				regularized			
	RK		BLOCK		RRK		RnB	
	FRB	PSD	FRB	PSD	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 (PSD). The estimates considered are the multivariate realized kernel (RK), blocking estimator based on 4 clusters of equal size (BLOCK) together with regularized versions using eigenvalue cleaning (RRK and RnB).

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 time scale 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

Table 3: Performance of realized kernel and blocking estimators for $p = 256$ and 4 asset clusters

	unregularized				regularized			
	RK		BLOCK		RRK		RnB	
	FRB	PSD	FRB	PSD	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 (PSD). The estimates considered are the multivariate realized kernel (RK), blocking estimator based on 4 clusters of equal size (BLOCK) together with regularized versions using eigenvalue cleaning (RRK and RnB).

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 distinguish from errors in the off-diagonal elements as well as in those of the entire matrix. Accordingly, Table 4 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 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 presents the corresponding results for the least liquid half of assets. A comparison against 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 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 the bandwidth.

4.2.3 Simulation 3: asset cluster size determination

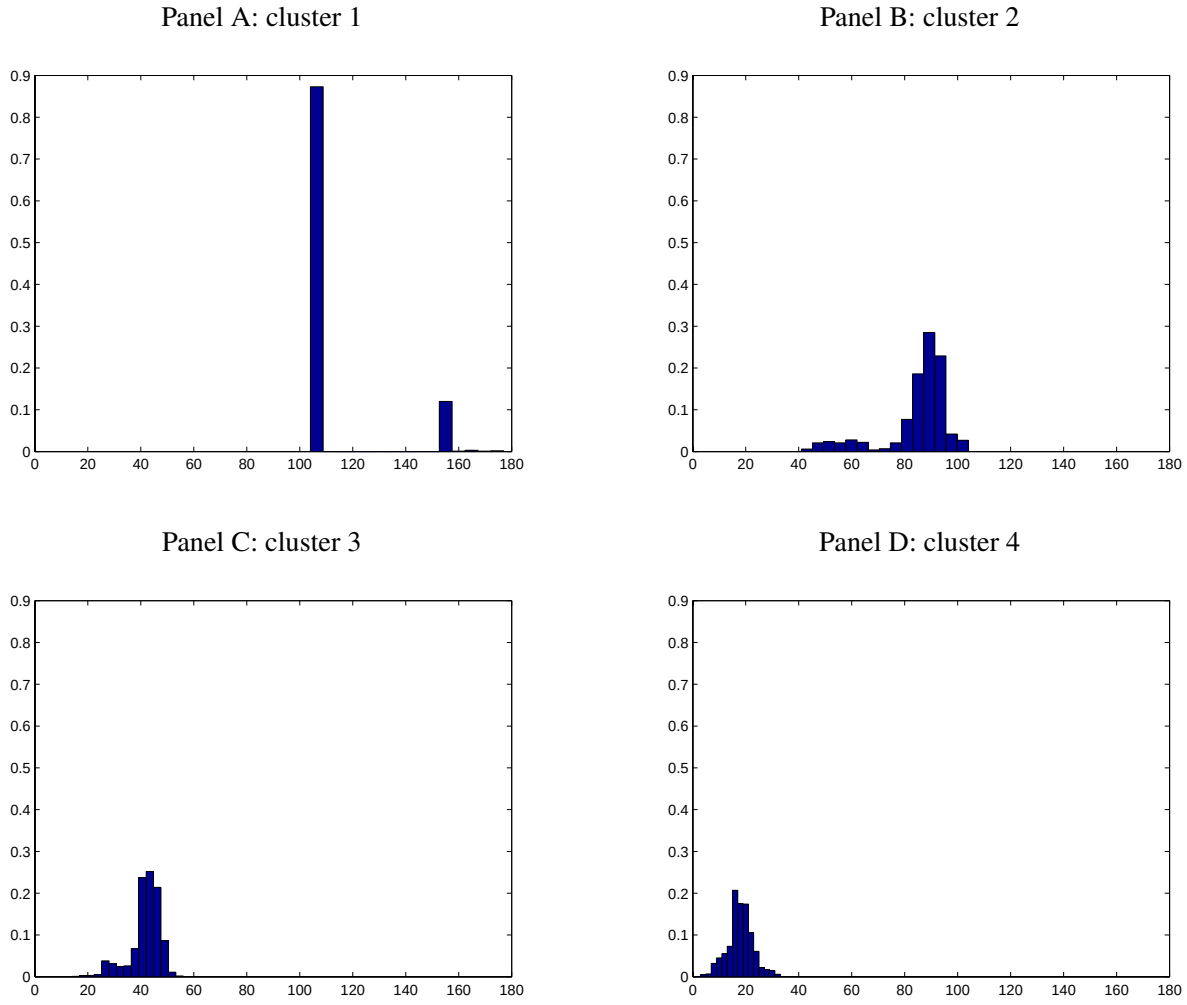
While clustering offers a solution to the excessive data reduction problem, an additional question emerges in determining the sizes of clusters. Foreshadowed by the computational burden of the HY estimator, it is of practical need to develop an estimator which can be represented with a parsimonious number of clusters and by extension blocks. The performance of a data-driven cluster is examined where size is determined using a simple hierarchical clustering algorithm to identify groups of observations that are self-similar, but distinct from other groups.

Table 4: Results for different number of asset clusters for $p = 256$

	BLOCK			RnB			BLOCK			RnB		
# clusters	FRB	DIA	OFF	FRB	OFF	INV	FRB	DIA	OFF	FRB	OFF	INV
	Panel A: entire matrix (1:256)						*Panel B: upper quarter (1:64)*					
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
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
	Panel C: upper half (1:128)						*Panel D: lower half (129:256)*					
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
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

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.

Figure 5: Clustering based on trade durations



Note. K-means clustering based on trade durations for $p = 256$. Clusters are presented from most liquid to least liquid groups.

The K-means clustering algorithm according to MacQueen (1967) is one of the simplest and widely used methods of hierarchical agglomerative clustering. K-means clustering 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. It is a two-step algorithm that alternates between assigning each observation to the cluster with the closest mean and updating the new means of the observations in the cluster. The algorithm is deemed to have converged when the assignments no longer change.

The third simulation examines cluster size determination using the K-means algorithm. We use the same simulation

environment 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. Figure 5 shows the distribution of cluster sizes for K-means algorithm based on trade-to-trade durations. The clusters are ordered from most liquid to least liquid groups. Recalling that there were 64 assets in each equal-sized cluster, data-driven clustering results in the illiquid clusters becoming much smaller whereas the liquid clusters become much larger. In fact, the K-means algorithm classifies approximately more than half of the observations as being in a liquid group, and then further divides the remaining illiquid half into three sequentially smaller sets.

Table 5: Results for K-means clustering

	BLOCK			RnB		
Method	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
Kmeans	1.9963	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
Kmeans	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
Kmeans	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
Kmeans	1.723	0.187	1.211	1.657	1.164	1.498

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.

Table 5 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 benchmark 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. It is clear that clustering with respect to trade durations further reduces the estimation error compared to the case of equal cluster sizes.[3] Consistent with the results shown in Simulation 2, Panels B and D, 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. Finally, it is observed that the error reduction is greater for the off-diagonal elements (OFF) than for the diagonal elements (DIA), suggesting that the gains are being driven by improved estimates of the covariance elements including illiquid assets.

5 Empirical analysis

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.[4] 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 January 1, 2007 to April 1, 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 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 midquote and when the midquote 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

[3] Note that we also analyzed clustering with respect to observation frequencies. It is shown that in this case K-means clustering does not reduce estimation error. Hence, observation asynchroneity is better understood in terms of waiting times.

[4] The S&P 500 has a number of illiquid assets and has qualitatively similar market microstructure features as the S&P 1500 calibrated simulation and substantiates the study of the RnB estimator in this environment.

Table 6: Summary Statistics for the daily log-return in percentage of S&P 500 stocks

	Panel A: full sample				*Panel B: pre-collapse*				*Panel C: post-collapse*			
	Mean	Std.	Skew.	Kurt.	Mean	Std.	Skew.	Kurt.	Mean	Std.	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
Std.	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 January 3, 2007 to April 1, 2009 for a total of 562 observations. Panel B: Pre-collapse period. The sample period extends from January 3, 2007 to September 13, 2008 for a total of 428 observations. Panel C: Post-collapse period. The sample period extends from September 14, 2008 to April 1, 2009 for a total of 134 observations.

consideration). See Barndorff-Nielsen et al. (2008b) for a detailed discussion of data filtering and the implications for estimators.

5.2 Summary statistics

Table 6 presents annualized summary statistics for daily 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 September 14, 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, Bollerslev, Diebold, and Labys (2001), Ait-Sahalia and Mancini (2008) and Andersen, Bollerslev, Frederiksen, and Nielsen (2009). 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 7 summaries the annualized covariance estimates of the S&P 500 stocks using the RK and RnB estimators for the entire sample. The RnB estimators is restricted to four equal-sized clusters. On average, the RnB estimators have

Table 7: Summary statistics for the annualized covariance distribution in percentage of S&P 500 stocks

	RK			RnB		
	Mean	Std.	Q_{22}	Mean	Std.	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
Std.	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 January 3, 2007 to April 1, 2009 for a total of 562 observations. The table reports the Ljung-Box Portmanteau test for up to 22nd order autocorrelation, Q_{22}, the 1% critical value is 40.289.

lower means and standard deviations.[5] 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 one month 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.

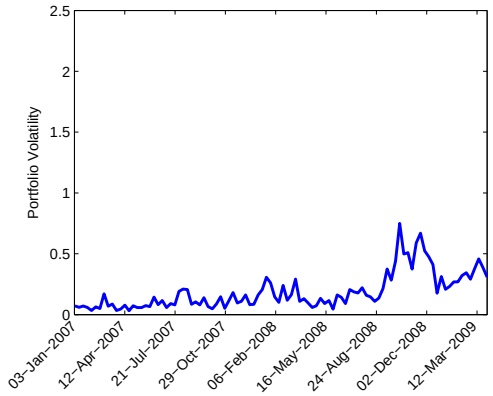
5.3 Forecasting portfolio volatility

Following the procedure outlined in Briner and Connor (2008), the forecast quality of estimates is evaluated according to the predictability of the future volatility of a (random) portfolio. Random portfolio w weights are drawn from a uniform distribution $U(-1,1)$ and scaled such that $\sum w = 1$. The estimated portfolio volatility is $\widehat{\sigma}_w = (w'\widehat{\Sigma}w)^{1/2}$, where $\widehat{\Sigma}$ is a covariance estimate. The realized portfolio volatility is computed from the daily absolute returns as $\widehat{\sigma}_w = |w'r_t|$. Figure 6 reports the portfolio volatility of an equally weighted portfolio in Panel A, and the corresponding median (50%) and the 25% and 75% quantiles in Panel B. The randomized portfolios capture the salient feature of the summary statistics presented in Table 6, namely higher market volatility following the collapse of Lehman Brothers.

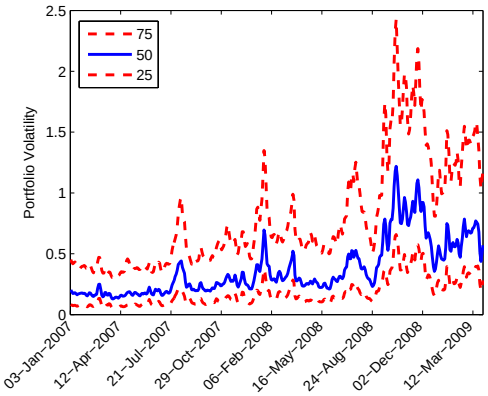
[5] Pre and post-collapse summary statistics are qualitatively the same and are not reported for the sake of space.

Figure 6: Portfolio volatility

Panel A: equally weighted



Panel B: randomly weighted



Note. Portfolio volatilities, $\widehat{\sigma}_w = |w'r_t|$, are realized absolute daily returns. Panel A shows the equally weighted portfolio volatility annualized. Panel B shows the median and quartiles of the randomly weighted portfolio volatility annualized.

Employing the Mincer and Zarnowitz (1969) framework, the forecast regression is specified as

$$\sqrt{\pi/2} * |w'r_t| = \alpha_0 + \alpha_1 \sqrt{w'\widehat{\Sigma}_{(1),t-1}w} + \alpha_2 \sqrt{w'\widehat{\Sigma}_{(2),t-1}w} + \varepsilon_t.$$

This regresses a proxy for ex post volatility of a randomly weighted portfolio on an intercept and competing portfolio volatility forecasts. Table 8 shows the out-of-sample Mincer Zarnowitz forecast evaluation regression results. Newey and West (1987) robust t-statistics are reported, where the bandwidths are determined following the procedure outlined in Newey and West (1994). The coefficient α_0 is a measure of forecast bias, whereas α_1 and α_2 are measures of forecast efficiency.

The performance of different versions of RnB are compared against the original RK estimator and a regularized version of the RK estimator (RRK) is included to assess the impact of regularization (without blocking). Panel A gives results of single forecasts with the null hypothesis, $\alpha_0 = 0$ and $\alpha_1 = 1$. It is shown that in terms of R^2, the RnB forecasts are more accurate than the RK forecasts. The regression coefficients α_0 is statistically significant only for the RK estimators, indicating that RK forecasts are biased whereas RnB forecasts are unbiased. The regression coefficients β_1 are closer to 1 for RnB forecasts than for RK forecasts, demonstrating that RnB forecasts are also more efficient. Panel B gives results of encompassing forecasts with the null hypothesis, $\alpha_0 = 0$ and $\alpha_1 + \alpha_2 = 1$. A forecast is encompassed when its coefficient is not statistically different than zero. It is shown that the RK forecasts are encompassed by the

alternative as the 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 forecasts.

Table 8: Mincer Zarnowitz forecast evaluation

$\widehat{\Sigma}_{(1)}$	$\widehat{\Sigma}_{(2)}$	α_0	α_1	α_2	R^2
Panel A: Single Forecasts					
RK		0.035 (3.133)	0.677 (7.821)		0.220
RRK		0.048 (4.411)	0.604 (6.807)		0.185
RnB		0.010 (0.574)	0.762 (9.202)		0.280
RnB-K		0.012 (0.657)	0.768 (9.193)		0.283
Panel B: Encompassing Forecasts					
RnB	RK	0.009 (0.351)	1.023 (4.776)	−0.272 (−1.230)	0.288
RnB-K	RK	0.011 (0.571)	0.930 (4.806)	−0.171 (−0.858)	0.288

Note. This table reports Mincer Zarnowitz regression

$$\sqrt{\pi/2}|w'r_t| = \alpha_0 + \alpha_1\sqrt{w'\widehat{\Sigma}_{(1),t-1}w} + \alpha_2\sqrt{w'\widehat{\Sigma}_{(2),t-1}w} + \varepsilon_t.$$

Random portfolios are generated from all available constituents of the S&P 500 from January 3, 2007 to April 1, 2009. Newey-West robust t-statistics are reported in parenthesis below.

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, Hansen, Lunde, and Shephard (2008a). These blocks are combined to form a complete covariance matrix, which is then regularized using a procedure called eigenvalue cleaning, as introduced by Laloux, Cizeau, Bouchaud, and Potters (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 em-

pirical 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 to cluster sizes arising from K-means algorithm based on trade-to-trade durations yields further reduction of estimation errors. Applying the RnB estimator to (out-of-sample) forecasts of 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 quite useful whenever high-dimensional covariances over short time horizons have to be estimated and forecast with preferably high precision. Given its computational tractability it might serve as a valuable tool in portfolio risk management applications.

References

Ait-Sahalia, Y. and Mancini, L. (2008), "Out of Sample Forecasts of Quadratic Variation," *Journal of Econometrics*, 147, 17–33.

Andersen, T., Bollerslev, T., Diebold, F., and Labys, P. (2001), "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55.

— (2003), "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Andersen, T., Bollerslev, T., Frederiksen, P., and Nielsen, M. (2009), "Continuous-time Models, Realized Volatilities, and Testable Distributional Implications for Daily Stock Returns." *Journal of Applied Econometrics*, forthcoming.

Andrews, D. (1991), "Heteroscedasticity and Autocorrelation Consistent Covaraince Matrix Estimation," *Econometrica*, 59, 817–858.

Bandi, F. and Russell, J. (2006), "Separating Microstructure Noise from Volatility," *Journal of Financial Economics*, 79, 655–692.

Bandi, F., Russell, J., and Zhu, Y. (2008), "Using High-Frequency Data in Dynamic Portfolio Choice," *Econometric Reviews*, 27, 163–198.

Barndorff-Nielsen, O., Hansen, P., Lunde, A., and Shephard, N. (2008a), "Multivariate Realized Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-synchronous Trading," Working Paper, University of Oxford, (web appendix available at `http://www.hha.dk/~alunde/`).

— (2008b), "Realised Kernels in Practice: Trades and Quotes," *Econometrics Journal*, 4, 1–32.

Barndorff-Nielsen, O. and Shephard, N. (2002), "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society, Ser. B.*, 64, 253–280.

— (2004), "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics," *Econometrica*, 72, 885–925.

Bickel, P. J. and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227.

Bonato, M., Caporin, M., and Ranaldo, A. (2008), "Forecasting Realized (Co)variances with a Block Structure Wishart Autoregressive Model," Working Paper Swiss Banking Institute of the University of Zrich.

Briner, B. and 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., and Lakonishok, J. (1999), "On Portfolio Optimization: Forecasting Covaraince and Choosing the Risk Model," *The Review of Financial Studies*, 12, 937–974.

Disatnik, D. (2009), "Portfolio Optimization Using a Block Structure for the Covaraince Matrix," Working Paper University of Michigan.

Duffie, D. and 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*, 291–298.

Gloter, A. and Jacod, J. (2001), "Diffusion with Measurement Errors. II. Optimal Estimators." *ESAIM Probability and Statistics*, 5, 243–260.

Griffin, J. and Oomen, R. (2009), "Covariance Measurement in the Presence of Non-synchronous Trading and Market Microstructure Noise," *Journal of Econometrics*, forthcoming.

Hansen, P. R. and Lunde, A. (2006), "Realized Variance and Market Microstructure Noise," *Journal of Business and Economic Statistics*, 24, 127–161.

Harris, F., McInish, T., Shoesmith, G., and Wood, R. (1995), "Cointegration, Error Correction and Price Discovery on Informationally-linked Security Markets," *Journal of Financial and Quantitative Analysis*, 30, 563–581.

Hayashi, T. and Yoshida, N. (2005), "On Covariance Estimation of Non-synchronously Observed Diffusion Processes," *Bernoulli*, 11, 359–379.

Jagannathan, R. and Ma, T. (2003), "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps," *Journal of Finance*, 58, 1651–1683.

Johnstone, I. (2001), "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *Annals of Statistics*, 29, 295–327.

Kinnebrock, S. and Podolskij, M. (2008), "An Econometric Analysis of Modulated Realised Covariance, Regression and Correlation in Noisy Diffusion Models," OFRC Working Papers Series 2008fe25, Oxford Financial Research Centre.

Laloux, L., Cizeau, P., Bouchaud, J.-P., and Potters, M. (1999), "Noise Dressing of Financial Correlation Matrices," *Physical Review Letters*, 83, 1467–1470.

Ledoit, O. and 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.

— (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*, eds. Cam, L. and Neyman, J., University of California Press, vol. 1, pp. 281–297.

Martens, M. (2006), "Estimating Unbiased and Precise Realized Covariances," Manuscript Erasmus University Rotterdam.

Michaud, R. O. (1989), "The Markowitz Optimization Enigma: Is Optimized Optimal," *Financial Analysts Journal*, 45, 31–42.

Mincer, J. and Zarnowitz, V. (1969), "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations*, ed. Mincer, J., Columbia University Press, pp. 3–46.

Neumaier, A. (1998), "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization," *SIAM Review*, 40, 636–667.

Newey, W. and West, K. (1987), "A Simple, Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covaraince Matrix," *Econometrica*, 55, 703–708.

— (1994), "Automatic Lag Selection in Covariance Matrix Estimation," *The Review of Economic Studies*, 61, 631–653.

Onatski, A. (2009), "Testing Hypotheses About the Number of Factors in Large Factor Models," *Econometrica*, forthcoming.

Oomen, R.C.A. (2006), Comment on "Realized Variance and Market Microstructure Noise," by P.R. Hansen and A. Lunde *Journal of Business and Economic Statistics*, 24, 195–202.

Oomen, R.C.A. (2009), "A Universal Scaling Law of Noise in Financial Markets", Manuscript University of Amsterdam.

Qi, H. and 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 of Oxford University.

Tola, V., Lillo, F., Gallegati, M., and Mantegna, R. (2008), "Cluster Analysis for Portfolio Optimization," *Journal of Economic Dynamics and Control*, 32, 235–258.

Voev, V. and Lunde, A. (2007), "Integrated Covariance Estimation Using High-Frequency Data in the Presence of Noise," *Journal of Financial Econometrics*, 5, 68–104.

Wang, Y. and Zou, J. (2009), "Vast Volatility Matrix Estimation for High-Frequency Financial Data," *Annals of Statistics*, forthcoming.

Zhang, L., Mykland, P., and Ait-Sahalia, Y. (2005), "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 1394–1411.

Zumbach, G. (2009), "The Empirical Properties of Large Covariance Matrices," Tech. rep., RiskMetrics Group.

CFS Working Paper Series:

No.	Author(s)	Title
2009/19	**Guenter W. Beck** **Volker Wieland**	Money in Monetary Policy Design: Monetary Cross-Checking in the New-Keynesian Model
2009/18	**Wolfgang Karl Härdle** **Nikolaus Hautsch** **Andrija Mihoci**	Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics
2009/17	**John F. Cogan** **Tobias Cwik** **John B. Taylor** **Volker Wieland**	New Keynesian versus Old Keynesian Government Spending Multipliers
2009/16	**Christopher D. Carroll**	Precautionary Saving and the Marginal Propensity to Consume Out of Permanent Income
2009/15	**Christopher D. Carroll** **Olivier Jeanne**	A Tractable Model of Precautionary Reserves, Net Foreign Assets, or Sovereign Wealth Funds
2009/14	**Christopher D. Carroll** **Patrick Toche**	A Tractable Model of Buffer Stock Saving
2009/13	**Jan Pieter Krahnen** **Günter Franke**	Instabile Finanzmärkte
2009/12	**Christopher D. Caroll** **Jiri Slacalek**	The American Consumer: Reforming, Or Just Resting?
2009/11	**Jan Pieter Krahnen** **Christian Wilde**	CDOs and Systematic Risk: Why bond ratings are inadequate
2009/10	**Peter Gomber** **Markus Gsell**	Algorithmic Trading Engines Versus Human Traders – Do They Behave Different in Securities Markets?

Copies of working papers can be downloaded at http://www.ifk-cfs.de