RESTRICTED TWO-SCALES COVARIANCE AND RISK CORRECTION IN THE CONTEXT OF HIGH-FREQUENCY FINANCIAL TRADING

Cyrille Nzouda, Shunpu Zhang, Kent Eskridge, Richard DeFusco

Abstract
The Two Scales Covariance estimator is a consistent estimator of the true covariance matrix used to evaluate the portfolio risk (Zhang et al. 2005). Used in the context of high frequency financial data, Fan et al (2012) demonstrated that the portfolio risk estimator using the pairwise refresh synchronization method and the portfolio risk estimator using the all-refresh synchronization method converged to the true risk. Moreover, they showed that the portfolio risk estimator using the pairwise refresh synchronization method converged faster. But, their simulation and empirical results showed that as the gross exposure increased, the estimated portfolio risks diverged from the true risk. They argued that the reason could be that the TSCV produced non positive definite matrices. We prove that forcing non positive definite matrices to be positive definite has negative consequences. We suggest a risk based on restricted TSCV - based on positive definite covariance matrices --, however that risk is not unbiased. This article suggests an expression of that bias and proposes a corrected portfolio risk estimator which is unbiased and based on positive definite TSCV. Simulations demonstrate that the estimated risks based on the restricted TSCV are more stable as the gross exposure increases, and we can correct for the bias.

KEYWORDS: Two Scales Covariance (TSCV), risk estimator, Pairwise refresh method, All-refresh method

1. INTRODUCTION

Zhang in 2005 developed a new estimator of the realized covariance matrix capable of handling the bias induced by the Epps effect, the non-synchronized effect and micro-structure effect. The estimator is called Two Scales Realized Volatility (TSRV) in the univariate framework and Two Scales Realized Covariance (TSCV) in the multivariate framework. The TSCV is defined as follows.

Consider a synchronized data set X, a $n \times p$ matrix where n represents the synchronized sample size and p the number of assets. x_{ij} the i^{th} log-price of the j^{th} asset. Since they are not directly observed because of the presence of micro-structures , we denote the actual matrix of log-price by X^o, and the latent price by X such that $X^o = X + \varepsilon$. We assumed that the errors (noise) are i.i.d normal distributed with mean 0 and variance σ^2. The TSCV is formulated as following:

$$< \widehat{X,X} >= [X^o, X^o]_1^{(k)} - \frac{\overline{n_k}}{n}[X^o, X^o]_1^{(1)}$$

where $[X_j^o, X_{j'}^o]_1^{(k)} = \frac{1}{k}\sum_{t=k}^{n}(X_{j_t}^o - X_{j'_{t-k}}^o)^2$ if $j = j'$ and $[X_j^o, X_{j'}^o]_1^{(k)} = \frac{1}{k}\sum_{t=k}^{n}(X_{j_t}^o - X_{j_{t-k}}^o)(X_{j'_t}^o - X_{j'_{t-k}}^o)$ if $j \neq j'$

$\overline{n_k} = (n - k + 1)/k$. k is optimally chosen with order $k = \mathcal{O}(n^{2/3})$.

Theoretically, (Fan, Li, & Yu, 2012) showed that the TSCV was a consistent estimator of the realized covariance matrix. But practically, it did not perform well. For instance, they proved that the TSCV should produce smaller risk when using the pairwise refresh method to synchronized the data. But, simulated data and real data showed that as the gross exposure constraint increased, the risk moved farther away from the true risk, and became even worse than the risk estimated using data synchronized by the all-refresh method.

Since the TSCV is consistent theoretically but not practically, two reasons explain why. Either the TSCV does not estimate true covariance matrices, or the estimation process causes a bias. In this paper, section 2 investigates the nature of the matrices estimated by the TSCV, section 3 describes the suggested Restricted-TSCV and its consequences, and section 4 details the proposed risk correction.

2. RESTRICTED-TSCV

By definition, a covariance matrix must be positive definite. But, the structure of the TSCV does not guaranty that the matrices produced are positive definite.

Theorem 1: TSCV can produce non positive definite matrices.

Proof

Consider a non zero $p \times 1$ vector z. It is known that a matrix, here the covariance matrix, is positive definite if and only if $z' < \widehat{X,X} > z > 0$. In fact,

$$\begin{aligned} z' < \widehat{X,X} > z &= z'\left([X^o,X^o]_1^{(k)} - \frac{\overline{n_k}}{n}[X^o,X^o]_1^{(1)}\right)z \\ &= z'\left([X^o,X^o]_1^{(k)}\right)z - z'\left(\frac{\overline{n_k}}{n}[X^o,X^o]_1^{(1)}\right)z \\ &= z'Az - z'Bz \end{aligned}$$

where $A = [X^o,X^o]_1^{(k)}$ with $z'Az > 0$,and $B = \frac{\overline{n_k}}{n}[X^o,X^o]_1^{(1)}$ with $z'Bz > 0$.

From the above result, three scenarios might happen: $z'Az - z'Bz > 0$, $z'Az - z'Bz = 0$, or $z'Az - z'Bz < 0$. The last two possibilities imply that the TSCV can produce non positive definite matrices, therefore the matrices estimated are not covariance matrices.

In order to solve the problem, researchers use the projected matrix onto the space of positive matrix. One approach is to consider a singular value decomposition of the matrix: $< \widehat{X,X} >= \Psi'\Lambda\Psi$, where Ψ is the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues. Since the matrix is not positive definite, some eigenvalues must be zero or negative. To force the matrix to be positive definite, the zero and negative eigenvalues are set to a positive number close to zero. Using the new diagonal matrix and the eigenvectors, the reconstructed matrix is positive definite.

Equivalently, the same positive matrix can be obtained by multiplying the first matrix of the TSCV by the smallest constant a such that z'(aA)z-z'Bz>0. By doing so, we impose a restriction to the TSCV that lead to the Restricted-TSCV. It is defined as following:

$$< \widehat{X,X_r} >= a[X^o,X^o]_1^{(k)} - \frac{\overline{n_k}}{n}[X^o,X^o]_1^{(1)}$$

with $a > 1$

3. CONSEQUENCES OF THE RESTRICTED-TSCV

The Restricted-TSCV is an unbiased estimator of the true covariance matrix.

Proof:

$$\begin{aligned} E\left(< \widehat{X,X_r} \mid a >\right) &= E\left\{\left(a[X^o,X^o]_1^{(k)} - \frac{\overline{n_k}}{n}[X^o,X^o]_1^{(1)}\right) \mid a\right\} \\ &= \frac{1}{k}a\sum(X_t - X_{t-k})^2 - \frac{\overline{n_k}}{n}\sum_{t=k}^{n}(X_t - X_{t-1})^2 + 2\overline{n_k}E\sigma^2(a-1) \\ &= < X,X_r > + \underbrace{2\overline{n_k}\sigma^2(a-1)}_{Bias} \end{aligned}$$

The bias is given by the expression $2\overline{n_k}\sigma^2(a-1)$ where $\overline{n_k} = (n-k+1)/k$ and σ^2 is the covariance matrix of noise. Since the optimal $\}k = \mathcal{O}(n^{2/3})$, we deduce that $\overline{n_k} = \mathcal{O}(n^{1/3})$. Therefore, as the sample size increases, the bias increases slowly.

4. RISK CORRECTION

Recall that the realized volatility is the volatility based on historical data. In other words, data recorded during a period $t - \tau_1$ are used to estimate the volatility matrix. Then this

volatility is used to estimate the weights by minimizing the risk. These weights are then used to evaluate the risk during period $t + \tau_2$. Usually $\tau_1 > \tau_2$. If the volatility matrix estimated using TSCV is not positive definite, then the estimated risk is based on the Restricted-TSCV. Since the Restricted-TSCV is biased, we suggest adjusting the risk evaluated at , $\hat{R} = \widehat{W}' < \widehat{X, X_r} > \widehat{W}$, by subtracting the following expression:

$$2\overline{n_k}\left(\widehat{W}\right)'\widehat{\sigma^2}\left(\widehat{W}\right)(\hat{a} - 1)$$

Then:

$$\hat{R}_{adj} = \hat{R} - 2\overline{n_k}\left(\widehat{W}\right)'\widehat{\sigma^2}\left(\widehat{W}\right)(\hat{a} - 1)$$

where $\widehat{W}$ is the vector of weights that minimizes the risk estimated using pass data. $\hat{a}$ is the estimated value of a. A consistent estimate of σ^2 is available.

Zhang (2005) proved that

$$plim\left(\frac{[X^o, X^o]}{2n} - \sigma^2\right) = 0$$

Then:

$$\widehat{\sigma^2} = \frac{[X^o, X^o]}{2n}$$

The main challenge in using this adjusted risk is how to choose a. The following theorem is useful.

Theorem 2: Consider $\Phi = A - B$ a $n \times n$ matrix written as difference of two covariance matrices. There exists a $n \times n$ non singular matrix L such that $A = LL'$ and $B = LDL'$ for some $n \times n$ diagonal matrix D with diagonal elements $\{d_i\}$. There exists a non zero vector L_m^{-1}, a column of the matrix $(L^{-1})'$ such that $(L_m^{-1})B(L_m^{-1})' = d_m(L_m^{-1})A(L_m^{-1})'$ where d_m is the positive maximum non zero value of the diagonal matrix D.

Lemma 1: Covariance matrices A and B are simultaneously diagonalizable.

Proof of lemma 1:

Consider two $n \times n$ covariance matrices A and B. Let's show that A and B are simultaneously diagonalizable. Let's assume a $n \times n$ nonsingular matrix L such that $L^{-1}AL = D_A$ and $L^{-1}BL = D_B$ where D_A and D_B are some diagonal matrices. Then:

$$\begin{aligned} L^{-1}ABL &= L^{-1}ALL^{-1}BL \\ &= D_A D_B \\ &= D_B D_A \\ &= L^{-1}BLL^{-1}AL \\ &= L^{-1}BAL \end{aligned}$$

Therefore

$$\begin{aligned} AB &= LL^{-1}ABLL^{-1} \\ &= LL^{-1}BALL^{-1} \\ &= BA \end{aligned}$$

Since A and B are both symmetric, condition is necessary and sufficient for matrices A and B to be simultaneously diagonalizable.

Proof of Theorem 2:

Since covariance matrices A and B are simultaneously diagonalizable, let's apply the Cholesky decomposition to A. There exist a singular matrix L such that $A = LL'$. Second, let's apply the Single Value Decomposition to B. Therefore, there exist the same singular matrix L with a diagonal matrix D such that $B = LDL'$.

$$\begin{aligned} A = LL' and B = LDL' \Leftrightarrow (L^{-1})A(L^{-1})' = I \quad and \quad (L^{-1})B(L^{-1})' = D' \\ (L^{-1})B(L^{-1})' \quad = \quad D(L^{-1})A(L^{-1})' \\ (L_i^{-1})B(L_i^{-1})' \quad = \quad d_i(L_i^{-1})A(L_i^{-1})' \end{aligned}$$

where L_i^{-1} is the i^{th} vector of the singular matrix L^{-1}, and d_i is its corresponding diagonal element in the diagonal matrix D. To complete the proof, choose the L_i^{-1} that corresponds to the positive and minimum $\{d_i\}$. Then $d_m = min\{d_i > 0\}$ and L_m^{-1} is its corresponding vector.

In the context of $A - B$ being negative definite, $(L_i^{-1})A(L_i^{-1})'$ is less than $(L_i^{-1})B(L_i^{-1})'$. Therefore, to establish the equality between them d_m most be greater than 1. Moreover, if we choose $\hat{a}$ to be equal to d_m, there could be a possibility for $d_m z'Az - z'Bz = 0$. To avoid it, we have to add to d_m a number that is positive and close to 0. Then, $\hat{a} = d_m + \eta$ and $\hat{a} > 1$.

We suggest to choose η between 0 and 0.0001.

5. Simulations and Results

5.1. Model

We use the model Fan et al (2012) used in their paper, however with a small change on the noise parameters.

Let the latent log-price X_t^i follows

$$dX_t^i = \mu^i dt + \rho^i \sigma_t^i dB_t^i + \sqrt{1 - (\rho^i)^2}\sigma_t^i dW_t + v^i dZ_t i = 1 \cdots p$$

Where B, W, and Z represent the standard Brownian motions, therefore, they are all $i.i.d\ N(0,1)$. The instantaneous volatility σ_t^i is modeled by the independent Ornstein-Uhlenbeck processes

$$d\eta_t^i = \alpha^i(\beta_0^i - \eta_t^i)dt + \beta_1^i dU_t^i$$

where $\eta_t^i = log(\sigma_t^i)$ and U_t^i is an independent Brownian motion.

The parameters are defined as: $(\mu^i, \beta_0^i, \beta_1^i, \alpha^i, \rho^i) = (0.03x_1^i, -x_2^i, 0.75x_3^i, -1/40x_4^i, -0.7)$ and $v^i = exp(\beta_0^i), x_j^i$ follows an independent uniform distribution on [0.7,1.3]. Notice that the parameters are kept fixed in the simulations.

In order to simulate the latent price at the order of one second (Benchmark data set), the Euler method was used. Equations and (4) can be written as followed:

$$X_{t+1} = X_t + \mu\Delta t + \rho\sigma_t^i\sqrt{\Delta t}B_t + \sqrt{1-(\rho^i)^2}\sigma_t^i\sqrt{\Delta t}W_t + v\sqrt{\Delta t}Z_t$$

$$\eta_{t+1} = \alpha^i(\beta_0^i - \eta_t^i)\Delta t + \beta_1\sqrt{\Delta t}U_t$$

Where $\Delta t = 1/N$, N is the number of observations per asset.

The micro-structure noise is introduced by perturbing the latent log-prices with $\varepsilon^i \overset{i.i.d}{\sim} N(0,(0.5)^2)$. The Poisson processes with parameters $(\lambda_1 \cdots \lambda_p)$ such that $\lambda_i = 0.02i \times 23400$ is used to simulate times to be used to non-synchronize the data according to each synchronization method.

5.2. Benchmark Portfolio Risk Evaluation

The Benchmark portfolio risk is used to assess the performance of other portfolio risks. In order to construct the benchmark portfolio, we simulate trading prices per second for 200 trading days. The number of data point per asset is 23400. In other words, trading prices are simulated secondly and there are 23400 seconds during a trading day. We call this data set the Benchmark-data set because no synchronization is applied. We start investing 1 unit of capital to p assets at day 101. In order to determine the portfolio allocation weights, we use the previous 100 trading days to estimate the covariance for day 101. Once we estimate the covariance, we use the optimization process described by equation

$$\begin{aligned} & min_{(w)} w^T \Sigma w \\ s.t. \quad & w^T 1 = 1 \\ & ||\, w \,||_1 \leq C \end{aligned}$$

to estimate the weights that minimize the risk for each gross exposure C. C varies from 1 to 3 by 0.1. After the first day, we restart the process for day 102, 103 until 200. In order to evaluate the performance of our portfolio, we use the estimated weights to calculate the risk.

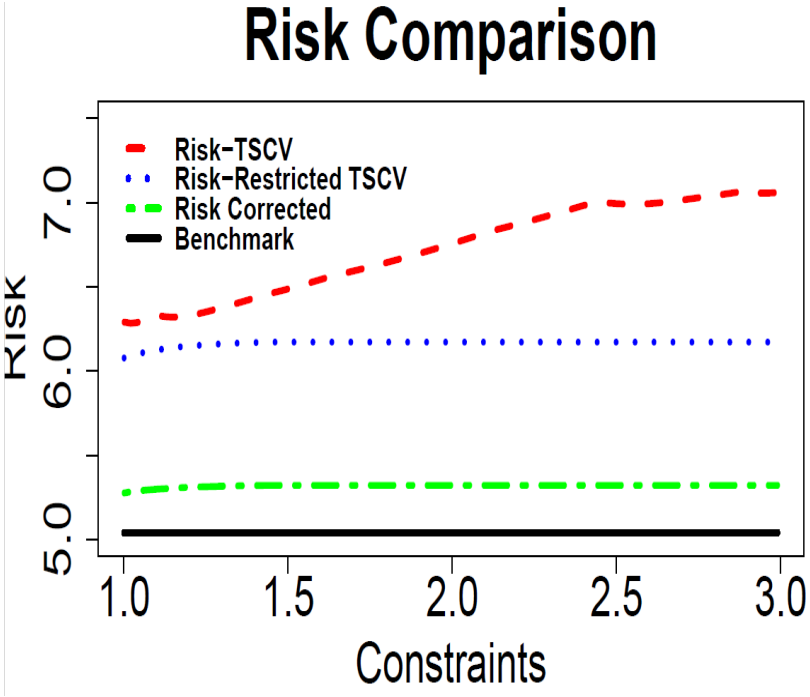
5.3. Risks Evaluation

The trading time scaled per second is not realistic because trades do not happen secondly all the time. In order to approximate the real world trading process, we use the Poisson process described above to simulate trading times and consider the trading prices that correspond to those trading times to construct a non-synchronized data set. Then, we use the the Pairwise refresh method described by Fan et al (2012) to synchronize the data set. Once the data are synchronized, we estimate the risk based on the TSCV, the risk based on the Restricted TSCV, and the corrected risk. We use the same strategy used by the benchmark portfolio described above. Note here that the estimated risk based on the TSCV is based on non-positive definite matrices. Therefore, we force the matrices to be positive definite by replacing the negatives eigenvalues by a small but positive value (0.00001). The risk based on the restricted TSCV is based on positive definite matrices, but it is biased. We estimate the bias and subtract it from the the estimated risk to obtain the corrected risk.

5.4. **Results**

Figure 1 displays the behavior of the Risk based on TSCV, the risk based on restricted TSCV, the corrected risk and the benchmark risk as the gross exposure increases. On one hand, it appears that the risk based on the TSCV departs from the benchmark risk as the gross exposure increases. This result is the same found by Fan et al (2012). The reason is that the covariance matrices are non-positive definite. We force them to be positive definite by replacing the negative eigenvalues by small and positive value (0.0001). On the other hand, the risk based on the restricted TSCV -- positive definite covariance matrices -- is more stable as the gross exposure increases. This implies that the technique used to force covariance matrices to be positive definite has different impacts on the estimated risk. Moreover, the restricted TSCV allows us to estimate the bias, the price we pay for having positive definite matrices. Subtracting that bias from the estimated risk based on restricted TSCV gives the risk corrected.

Figure 1: Comparison between Risk beased on TSCV, Risk based on restricted TSCV, Corrected risk, and benchmark risk



6. CONCLUSION

In this article, we proved that there was a bias when forcing a non-positive Two Scales Covariance matrix to be positive definite. We suggested an expression of that bias and proposed a corrected portfolio risk estimator which was unbiased and based on positive definite TSCV. Simulation results showed that forcing non positive covariance matrices by replacing negative eigenvalues by small but positive values (0.0001) has negative consequences on the estimated portfolio risk. We recommend to use the risk based on the restricted TSCV because the estimated risks produced are more stable as the gross exposure increases, and we can correct for the bias.

References

Ait-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial studies, 18*(2), 351-416.

Bachelier, L. (1900). *Th{\'e}orie de la sp{\'e}culation.* Gauthier-Villars.

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

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics, 162*(2), 149-169.

Bickel, P. J., & Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 199-227.

Burdzy, K. (n.d.). Brownian motion and its applications to mathematical analysis.

Campbell, J. Y., Lo, A. W.-C., MacKinlay, A. C., & others. (1997). *The econometrics of financial markets* (Vol. 2). princeton University press Princeton, NJ.

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

Cochrane, J. H. (2009). *Asset Pricing:(Revised Edition).* Princeton university press.

Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association, 74*(366a), 291-298.

Fan, J., & Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association, 102*(480), 1349-1362.

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

Fan, J., Zhang, J., & Yu, K. (2008). Asset allocation and risk assessment with gross exposure constraints for vast portfolios. *arXiv preprint arXiv:0812.2604*.

Halmos, P. R. (1950). *Measure Theory. vol. 2.*. (v. Nostrand, Ed.) New York.

Hautsch, N., Kyj, L. M., & Oomen, R. C. (2012). A blocking and regularization approach to high-dimensional realized covariance estimation. *Journal of Applied Econometrics, 27*(4), 625-645.

Hayashi, T., Yoshida, N., & others. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli, 11*(2), 359-379.

Iacus, S. M. (2009). *Simulation and inference for stochastic differential equations: with R examples* (Vol. 1). Springer Science \& Business Media.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance, 7*(1), 77-91.

Markowitz, H. M. (1968). *Portfolio selection: efficient diversification of investments* (Vol. 16). Yale university press.

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association, 104*(485), 177-186.

Wang, Y., & Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics*, 943-978.

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics, 159*(1), 235-250.

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

Zhang, L., & others. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli, 12*(6), 1019-1043.

Zhang, L., Mykland, P. A., & A{\"\i}t-Sahalia, Y. (2005). A tale of two time scales. *Journal of the American Statistical Association, 100*(472).

Zheng, X., Li, Y., & others. (2011). On the estimation of integrated covariance matrices of high dimensional diffusion processes. *The Annals of Statistics, 39*(6), 3121-3151.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *J. Bus. Econ. Statist*.