Cluster-Based Shrinkage of Correlation Matrices for Portfolio Optimization

Stjepan Begušić and Zvonko Kostanjčar

Laboratory for Financial and Risk Analytics
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb
stjepan.begusic@fer.hr, zvonko.kostanjcar@fer.hr

***Abstract*—The estimation of correlation and covariance matrices from asset return time series is a critical step in financial portfolio optimization. Although sample estimates are reliable when the length of time series is very large compared to the number of assets, in high-dimensional settings estimation issues arise. To reduce estimation errors and mitigate their propagation to out-of-sample performance of portfolios based on noisy estimates, shrinkage methods are applied. In this paper we consider several shrinkage methods for correlation matrix estimation and define a cluster-based shrinkage procedure which introduces information about the structures of communities identified in asset dependence graphs. To test the considered shrinkage methods we apply them in a portfolio optimization scenario using the global minimum variance portfolio, and perform backtests on a large sample of NYSE daily stock return data. We find that shrinkage methods generally improve out-of-sample portfolio performance, and the proposed cluster-based method yields improved results and portfolios which outperform other considered methods.**

***Index Terms*—Correlation, clustering, finance, portfolio optimization, shrinkage**

I. Introduction

A most common statistical tool for measuring pairwise dependencies between variables is the correlation matrix, frequently used in many disciplines, such as physics, image analysis, genomics, engineering, and economics. Specifically, the concept of correlations between asset returns is vital to measuring co-movement in the market and detecting common sources of risk. This is a critical point in many portfolio optimization scenarios, especially in the Markowitz framework where the risk is measured by portfolio variance, to which the cross-correlations of asset contribute largely. Generally, covariance and correlation matrices can reliably be estimated using the standard estimators when the length of asset return time series T is much larger than the number of asset N. However, when the number of asset return time series N is similar or even greater than their length T, severe estimation issues arise [1]. The errors in the covariance matrix estimates further propagate to portfolio optimization problems, leading to poor out-of-sample performance of portfolios optimized using noisy estimates [2], [3]. Since the number of investable assets in a globalized financial system continues to grow, and considering the fact that financial markets are dynamic complex systems which may often change abruptly [4], the considered dimensionalities N are large and the time windows can only be of limited length T, thus making these estimation issues a very much relevant and current topic in finance.

In this paper we study shrinkage techniques which can provide more reliable estimates of the covariance and consider a portfolio optimization scenario for evaluating the various shrinkage approaches. Generally, some shrinkage methods rely on shifting the original empirical estimate towards a shrinkage target, others transform the correlation matrix according to its spectrum [5]. In addition to some state-of-the-art shrinkage methods, we define a cluster-based shrinkage procedures, relying on some previous results demonstrating the emergence of connected components and communities in asset dependence graphs [6], [7]. We test the considered shrinkage methods by employing a risk-based portfolio optimization method to the covariance estimates and performing a backtest on a dataset of $N = 578$ stocks traded at the NYSE from 1999 to 2019. We find that shrinkage methods generally provide estimates which improve the out-of-sample performance of the optimized portfolios in comparison to the empirical noisy estimated. Furthermore, the considered cluster-based procedure is shown to outperform other methods and provide improved risk-adjusted performance of the optimized portfolios. Our results suggest that the clusters representing asset communities may introduce structure in the shrinkage procedure and improve the resulting estimates.

II. Correlation matrix estimation and shrinkage

Most commonly, the empirical correlation matrix of a collection of N time series of length T is calculated as the sample correlation matrix $\mathbf{R} \in \mathbb{R}^{N \times N}$. These estimates are known to be reliable when $T \gg N$. However, in many modern applications, N is rather large and T is commensurate or even smaller than N. In this scenario, empirical estimates are driven by noise and for $N > T$ the inverse is not defined, which may be a problem in many optimization scenarios. In what follows, we consider several popular shrinkage methods for obtaining better correlation estimates in such high-dimensional settings.

978-1-7281-3140-5/19/$31.00 ©2019 IEEE

A. *Basic linear shrinkage*

The most basic and very commonly used shrinkage method is called *basic linear shrinkage*, which is a convex combination of the empirical estimate $\mathbf{R}$ and the $N \times N$ identity matrix $\mathbf{I}_N$:

$$\mathbf{R}^{(lin.)} = \alpha \mathbf{I}_N + (1-\alpha)\mathbf{R}. \tag{1}$$

By employing this method, the off-diagonal correlation estimates are basically pushed towards 0, thus mitigating the effect of spurious large positive or negative pairwise correlations. In practice, although very simple, this method has proven to be quite useful in high-dimensional correlation estimation.

B. *Constant correlation model*

Somewhat more elaborate, but nevertheless very simple method is the *constant correlation* shrinkage. As opposed to basic linear shrinkage in which the off-diagonal estimates are shifted towards 0, this method assumes a constant correlation between all assets, estimated by the average correlation across all pairs:

$$r_{ij} = r = \frac{1}{N(N-1)} \sum_{i \neq j} R_{ij}, \quad \forall i \neq j. \tag{2}$$

Thus, the resulting shrinkage $\mathbf{R}^{(const.)}$ is a matrix with ones on the diagonal and r on all the off-diagonal elements. Although fairly limited, this version of shrinkage has been shown to produce improved results in financial applications [2].

C. *Eigenvalue clipping*

A more elaborate shrinkage procedure comes from the spectrum of the sample correlation matrix estimate, which is known to carry important structural information about the considered system. The *eigenvalue clipping* estimate aims to keep the information in the K eigenvectors corresponding to the largest K eigenvalues, while reducing the rest:

$$\mathbf{R}^{(clip.)} = \sum_{k=1}^{N} \xi_k^{(clip.)} \mathbf{u}_k \mathbf{u}_k^\intercal, \quad \xi_k^{(clip.)} = \begin{cases} \lambda_k, & \text{if } k \leq K \\ \gamma, & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathbf{u}_k$ and λ_k are eigenvectors and eigenvalues of the sample correlation $\mathbf{R}$, sorted by descending eigenvalue magnitude, and γ is a constant calculated as the mean of the remaining eigenvalues ($k > K$), to preserve the matrix trace. Although it corrects the smaller eigenvalues, this method may yield varying results, depending on the noise present in the eigenvectors associated with the smallest eigenvalues.

III. Cluster-based shrinkage

Some previous research has also focused on asset connectivity graphs, finding that their structures may explain the dynamics of financial markets and help identify common risk factors [8]. Moreover, recent results suggest that connected components emerge and dynamically change within these time-varying financial networks, thus reflecting the underlying risk properties of the system [7]. Therefore, in what follows we hypothesize that communities in asset connectivity graphs form K clusters in the T-dimensional return space of N assets, and that these clusters represent common risk sources. Let $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_K\}$ denote the set of clusters, such that each $\mathcal{C}_p$ is a set of integer indices of assets within cluster p. Correlations within and between clusters are contained in a cluster correlation matrix $\mathbf{S} \in \mathbb{R}^{K \times K}$. The average correlation S_{pp} within cluster p and the average correlation S_{pq} between clusters p and q can be calculated as:

$$S_{pp} = \frac{1}{|\mathcal{C}_p|(|\mathcal{C}_p|-1)} \sum_{i \in \mathcal{C}_p} \sum_{\substack{j \in \mathcal{C}_p \\ j \neq i}} R_{ij}, \tag{4}$$

$$S_{pq} = \frac{1}{2|\mathcal{C}_p||\mathcal{C}_q|} \sum_{i \in \mathcal{C}_p} \sum_{j \in \mathcal{C}_q} R_{ij}, \tag{5}$$

where $|\mathcal{C}_p|$ is the cardinality of (the number of assets in) cluster p. To obtain the shrinkage target matrix $\tilde{\mathbf{R}}$, each pairwise correlation $\tilde{R}_{ij}$ is defined by the intra- or inter-cluster correlations of the corresponding clusters of assets i and j:

$$\tilde{R}_{ij} = S_{pq}, \quad i \in \mathcal{C}_p, j \in \mathcal{C}_q. \tag{6}$$

The final shrinkage can be obtained similar to the basic linear shrinkage principle, using a convex combination of the empirical correlation matrix $\mathbf{R}$ and the cluster-based shrinkage target $\tilde{\mathbf{R}}$:

$$\mathbf{R}^{(clust.)} = \alpha \tilde{\mathbf{R}} + (1-\alpha)\mathbf{R}, \tag{7}$$

with parameter $\alpha \in [0, 1]$. To obtain K clusters from N assets we apply a modified version of the k-means algorithm with an improved randomized seeding technique for centroid initialization [9] and a pairwise correlation distance:

$$d_{ij} = 1 - R_{ij}. \tag{8}$$

Moreover, since the k-means procedure is a greedy algorithm, we use 100 replicates of the set of initial centroids to obtain stable and improved results. To determine the number of clusters in a given matrix of asset returns $\mathbf{X} \in \mathbb{R}^{T \times N}$, we employ tools from random matrix theory (RMT) [1], specifically the Marčenko-Pastur distribution which determines the smallest and largest expected eigenvalues λ_- and λ_+ of the data correlation matrix $\mathbf{R}$ under the assumption of uncorrelated variables, based on the ratio $q = \frac{N}{T}$. Any eigenvalues λ_k of $\mathbf{R}$ larger than λ_+ are considered representative of meaningful components in the data, and so the number of clusters K corresponds to the number of eigenvalues larger than this limit:

$$K = |\{\lambda_i > \lambda_+\}|, \quad \lambda_+ = (1+\sqrt{q})^2 = 1+2\sqrt{\frac{N}{T}}+\frac{N}{T}. \tag{9}$$

The changes in the limit λ_+ depending on the number of assets N and the length of the time window T have an intuitive interpretation in the context of factor models and estimation errors. Firstly, for larger N (more assets) the limit is higher, which is in accordance with the assumption of approximate factor models [10], in which all but the K largest eigenvalues are unbounded with $N \rightarrow \infty$, where the top K eigenvalues grow with N and should exceed the limit λ_+. Moreover, for larger T (longer time window) the limit approaches 1, reflecting the convergence of the correlation estimates toward the true population correlation matrix - in which all the eigenvalues are equal to 1 if the assumption of no correlation holds, and any eigenvalues larger than 1 represent meaningful correlated components.

Rather than using a diagonal target matrix or a constant correlation model, the structure is provided depending on the correlation within and between the identified clusters in the asset space. Moreover, the eigenvalue-based methods may be affected by the noisy eigenvalue estimates, whereas the cluster-based structures provide a more robust and sparse low-rank representation of the data.

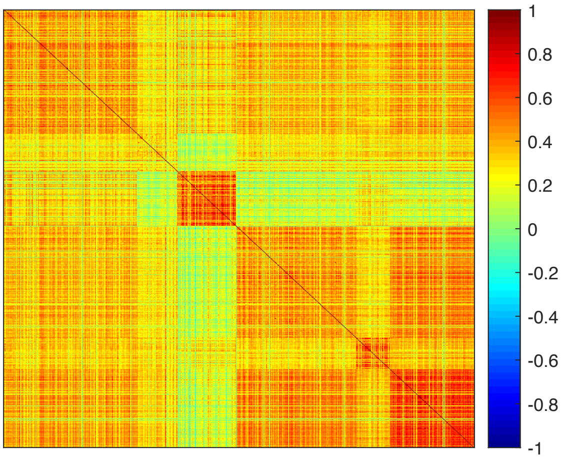

Fig. 1. Empirical correlation matrix estimate for $N = 578$ stock returns from the NYSE, using a time window of $T = 252$ days. The color map on the right indicates correlation magnitude.

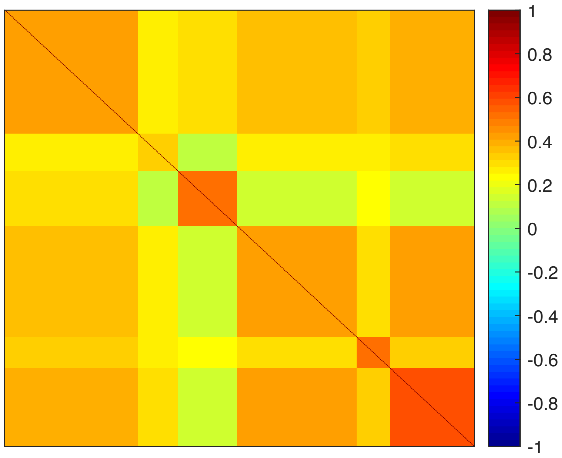

Fig. 2. Cluster-based shrinkage target matrix obtained from the correlation matrix estimate for $N = 578$ stock returns from the NYSE, using a time window of $T = 252$ days, with $\alpha = 1$ and the estimated number of clusters $K = 6$. The color map on the right indicates correlation magnitude.

An example of the original empirical correlation estimate of $N = 578$ stock returns from the NYSE, using a time window of $T = 252$ days, and the cluster-based shrinkage (for the edge case $\alpha = 1$ is given in Figures 1 and 2. The empirical estimate in Figure 1 is suggestive of a block structured correlation, which is extracted by the shrinkage procedure and the cluster-based estimate in Figure 2. Therefore, the cluster-based shrinkage procedure can be thought of as a ”pixelization” or ”blurring” of the original noisy estimate, toward a partitioned block matrix implied by communities in the asset connectivity structures.

IV. Application in portfolio optimization

A. Global minimum variance portfolio

To inspect the validity of the considered shrinkage methods, we apply them in a risk-based portfolio optimization scenario. Specifically, a very popular portfolio optimization method, stemming from the Markowitz mean-variance framework is the *global minimum variance* (GMV) portfolio [11], in which the optimal portfolio is the one which minimizes the variance:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^\intercal \mathbf{Q}\mathbf{w} \\ \text{s. t.} \quad & \mathbf{w}^\intercal \mathbf{1} = 1 \\ & \mathbf{w} \geq \mathbf{0}. \end{aligned} \tag{10}$$

The portfolio weight vector $\mathbf{w} \in \mathbb{R}^N$ is constrained so that all weights are positive (no short-selling) and they sum to 1 (investment constraint) - these are common constraints in practical applications. The only parameter estimated from the data here is the asset covariance matrix $\mathbf{Q}^{\in}\mathbb{R}^{N\times N}$. The optimization problem can be solved using sequential quadratic programming (SQP) - a nonlinear optimization algorithm.

As opposed to the general mean-variance framework, this approach is agnostic toward expected returns and thus avoids errors which plague the expected return estimation. However, the errors associated with the estimation of $\mathbf{Q}$ may still affect the out-of-sample portfolio performance, especially when the number of asset time series N is large compared to their length T. This is why this setting is very suitable for the application and evaluation of shrinkage methods. Since the considered methods are primarily designed and mostly used for correlation matrices, we apply them to the correlation estimates $\mathbf{R}$ to obtain the shrunk estimate $\mathbf{R}^{(s)}$, from which the covariance can be reconstructed as:

$$\mathbf{Q}^{(s)} = \mathbf{R}^{(s)} \circ (\hat{\sigma}\hat{\sigma}^\intercal), \tag{11}$$

where $\hat{\sigma} \in \mathbb{R}^N$ is a vector of individual asset volatilities (standard deviation estimates), and $\circ$ denotes the Hadamard element-wise product. By using the in-sample estimates of the individual asset volatilities we avoid any look-ahead bias, and make the empirical estimate (without shrinkage) equal to the in-sample covariance matrix.

B. Data and backtests

To test the considered shrinkage methods, we use a dataset containing daily returns of $N = 578$ stocks traded at the NYSE from 1999 to 2019. The dataset contains stocks from various sectors and provides some diversification potential for building diversified portfolios.

TABLE I
PERFORMANCE STATISTICS OF THE PORTFOLIOS BUILT USING THE CONSIDERED SHRINKAGE METHODS, COMPARED TO THE STANDARD 1/N PORTFOLIO AND THE GMV PORTFOLIO WITHOUT SHRINKAGE. THE BEST ENTRIES IN EACH COLUMN ARE EMPHASIZED IN BOLDFACE.

	r_{ann}	σ_{ann}	Sharpe	Sortino	Max. DD
1/N	5.20%	19.84%	0.15	0.19	68.12%
GMV (Empirical estimate)	7.00%	7.68%	0.63	0.7	37.71%
GMV (Lin. shrinkage, $\alpha = 0.5$)	7.59%	7.62%	0.71	0.79	36.49%
GMV (Constant corr.)	**8.21%**	7.92%	0.76	0.88	28.12%
GMV (Eigenvalue clip.)	6.83%	7.65%	0.61	0.68	34.71 %
GMV (Clust., $\alpha = 0.5$)	7.48%	**7.46%**	0.71	0.81	30.15%
GMV (Clust, $\alpha = 0.75$)	7.73%	7.63%	0.73	0.84	**28.05%**
GMV (Clust., $\alpha = 1$)	8.18%	7.83%	**0.77**	**0.89**	28.49%

To test the methods, we perform backtests on the dataset, using a lookback window of $T = 252$ trading days, which correspond to approximately 1 year. Each quarter (once every 3 months), we use the past T daily returns to estimate the original and shrunk versions of the covariance matrix, and based on these matrices find the corresponding GMV portfolios. We hold these portfolios until the next rebalance time, in 3 months. We also build an equal-weighted portfolio, which is used as a general benchmark for what the investor could have earned without any optimization.

To measure the performance of the various portfolios based on the considered shrinkage methods, we employ several statistics. Firstly, the annual expected returns r_{ann} and volatilities σ_{ann} reflect what the investor could expect as a level of annual return and the corresponding risk. Furthermore, we calculate the risk-adjusted performance measures, namely the ratios of expected return and volatility: Sharpe ratio $\mathrm{E}[r_p - r_{rf}]/\sigma_p$, and the Sortino ratio $\mathrm{E}[r_p - r_{rf}]/\sigma_d$, where r_{rf} is the risk-free rate, σ_p is the portfolio volatility (through the backtest period) and σ_d is the downside volatility. As opposed to the Sharpe ratio, the Sortino ratio only uses the downside volatility as a more accurate measure of risk (since it is the negative returns which are really adverse to the portfolio performance). Finally, we also calculate the maximum drawdown - the largest loss an investor could incur throughout the backtested period.

V. RESULTS

Using the NYSE stock returns dataset, with a $T = 252$ day lookback window and a 3 month rebalance interval, we backtest the considered methods on a total time frame of 20 years, from 1999 to 2019. For the cluster-based shrinkage we test three different values of the parameter α: $0.25, 0.5$, and 0.75. Furthermore, the number of clusters K are estimated at each rebalance point from the lookback window data using the aforementioned RMT-based method. The same number of components is also used for the eigenvalue clipping method. The estimated number of components K at each rebalance point is given in Figure 3. The dynamics of K in Figure 3 suggests that around the bubble, crash, and rebound period of 2006-2010, most assets in the market were most likely very correlated and were explainable by a lower number of factors (with K reaching as low as 5), while during periods of lower volatility, the estimated K seems to fluctuate around 8 to 10. These results are in line with previous findings suggesting that the amount of variance explained by a first few components in the cross section of asset returns is indicative of systemic risk in the financial market [8].

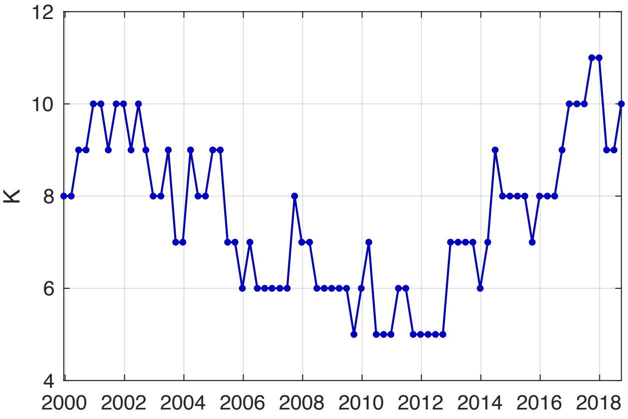

Fig. 3. The estimated number of clusters K through the backtest period.

The performance statistics for the portfolios built using the considered shrinkage methods are given in Table I. The results indicate that the cluster-based shrinkage methods yields reliable estimates of the covariance which can be used to build portfolios that outperform traditional ones relying on noisy empirical estimates. Moreover, in terms of risk-adjusted performance, as measured by the Sharpe and Sortino ratios, the portfolios using cluster-based shrinkage are considerably outperforming other methods, only matched by the very efficient constant correlation model. Nevertheless, the cluster-based shrinkage outperforms the constant correlation model for the $\alpha = 1$ case. Generally, all shrinkage methods outperform the no-shrinkage noisy empirical estimates, except the eigenvalue clipping method which seems to perform very similarly. Finally, the results suggest that overall the cluster-based methods outperform the other methods, as indicated by the best or second-best results in all considered statistics.

In addition to the performance statistics, to visualize the performance of the GMV portfolio, and especially the benefits of the proposed cluster-based shrinkage, we display the portfolio performance of the 1/N portfolio, the GMV portfolio based on empirical covariance estimates, and the GMV portfolio based on the cluster-based shrinkage in Figure 4.

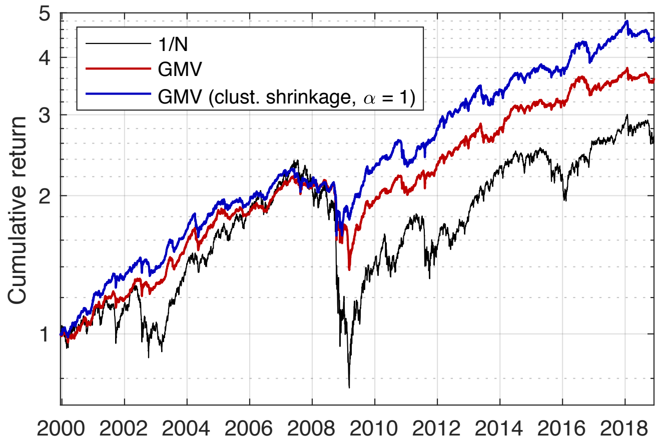

Fig. 4. The performance of the 1/N portfolio, the GMV portfolio based on empirical covariance estimates, and the GMV portfolio based on the cluster-based shrinkage through the backtest period.

Firstly, the results in Figure 4 demonstrate how dramatic the difference between the optimized GMV portfolios and the naive 1/N portfolio really is. Moreover it is evident that the cluster-based shrinkage method results in a portfolio which outperforms the one based on noisy empirical estimates - not only through a significantly lower drawdown in the midst of the 2008-2009 crisis (also noticeable in Table I), but as well as through the entire period.

VI. Conclusion

In this paper we consider shrinkage methods for obtaining more reliable correlation matrix estimates from noisy financial return data. In addition to several standard methods, such as basic linear shrinkage, eigenvalue clipping and the constant correlation model, we also define a cluster-based shrinkage approach, using clusters of original assets to represent the common risk factors which define the pairwise asset correlations. The number of clusters is determined by using a method from RMT based on the spectrum of the correlation matrix estimates. We test the considered shrinkage methods in a portfolio optimization scenario, using the global minimum variance portfolios, which only include the covariance estimates as an input. The backtest results indicate that shrinkage methods generally improve the out-of-sample portfolio performance. Moreover, portfolios based on the considered cluster-based shrinkage are shown to outperform other methods. The results suggest that the cluster-based structures identified in financial return data can augment shrinkage and yield reliable estimates of the asset correlation matrices.

References

[1] J. Bun, J. P. Bouchaud, and M. Potters, “Cleaning large correlation matrices: Tools from Random Matrix Theory,” *Physics Reports*, vol. 666, pp. 1–109, 2017. doi: 10.1016/j.physrep.2016.10.005

[2] O. Ledoit and M. Wolf, “Honey, I Shrunk the Sample Covariance Matrix,” *The Journal of Portfolio Management*, vol. 30, no. 4, pp. 110–119, jul 2004. doi: 10.3905/jpm.2004.110

[3] E. Pantaleo, M. Tumminello, F. Lillo, and R. N. Mantegna, “When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators,” *Quantitative Finance*, vol. 11, no. 7, pp. 1067–1080, jul 2011. doi: 10.1080/14697688.2010.534813

[4] Z. Kostanjčar, S. Begušić, H. E. Stanley, and B. Podobnik, “Estimating Tipping Points in Feedback-Driven Financial Networks,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1040–1052, sep 2016. doi: 10.1109/JSTSP.2016.2593099

[5] J. Bun, J.-P. Bouchaud, and M. Potters, “Cleaning correlation matrices,” *Risk Magazine*, vol. 2015, no. April, 2015.

[6] M. MacMahon and D. Garlaschelli, “Community Detection for Correlation Matrices,” *Physical Review X*, vol. 5, no. 2, p. 021006, apr 2015. doi: 10.1103/PhysRevX.5.021006

[7] S. Begušić, Z. Kostanjčar, D. Kovač, H. E. Stanley, and B. Podobnik, “Information Feedback in Temporal Networks as a Predictor of Market Crashes,” *Complexity*, vol. 2018, pp. 1–13, sep 2018. doi: 10.1155/2018/2834680

[8] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, “Econometric measures of connectedness and systemic risk in the finance and insurance sectors,” *Journal of Financial Economics*, vol. 104, no. 3, pp. 535–559, 2012. doi: 10.1016/j.jfineco.2011.12.010

[9] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007. doi: 10.1145/1283383.1283494. ISBN 9780898716245. ISSN 978-0-898716-24-5 pp. 1027–1025.

[10] J. Bai and S. Ng, “Rank regularized estimation of approximate factor models,” *Journal of Econometrics*, apr 2019. doi: 10.1016/j.jeconom.2019.04.021

[11] R. G. Clarke, H. de Silva, and S. Thorley, “Minimum-Variance Portfolios in the U.S. Equity Market,” *The Journal of Portfolio Management*, vol. 33, no. 1, pp. 10–24, oct 2009. doi: 10.3905/jpm.2006.661366