# Augment Large Covariance Matrix Estimation With Auxiliary Network Information

Shuyi Ge[*], Shaoran Li[†], Oliver Linton[‡], and Weiguang Liu[§]

Faculty of Economics, University of Cambridge

June 13, 2021

## Abstract

This paper aims to incorporate auxiliary information about the location of significant correlations into the estimation of high-dimensional covariance matrices. With the development of machine learning techniques such as textual analysis, granular linkage information among firms that used to be notoriously hard to get are now becoming available to researchers. Our proposed method provides an avenue for combining those auxiliary network information with traditional economic datasets to improve the estimation of a large covariance matrix. Simulation results show that the proposed adaptive correlation thresholding method generally performs better in the estimation of covariance matrices than previous methods, especially when the true covariance matrix is sparse and the auxiliary network contains genuine information. As a preliminary application, we apply the method to the estimation of the covariance matrix of asset returns. There are several extensions and improvements that we are considering.

[*] Author email:sg751@cam.ac.uk

[†] Author email:sl736@cam.ac.uk

[‡] Author email:obl20@cam.ac.uk

[§] Author email:wl342@cam.ac.uk

# 1 Model and Introduction

Our goal is to estimate $\Sigma = \mathrm{var}(y)$ where $y$ is a $p \times 1$ random vector, say asset returns. We collect the $T$ observations into a matrix $Y : p \times T$. Sample covariance estimate $\hat{\Sigma} = \frac{1}{T}(Y - \bar{Y}\mathbf{1})(Y - \bar{Y}\mathbf{1})'$ is problematic when $p$ is not small relative to $T$. Popular estimation strategies include factor model, shrinkage, thresholding, banding, tapering, etc.

If in addition to the observation of $Y$, we observe a network $G$ among the firms, where $G_{ij}$ either takes value $0, 1$ or a score in $[0, 1]$, with higher $G_{ij}$ implying that it's more "likely" that the returns of firm $i, j$ are correlated. We show that this auxiliary network can be used to improve the estimation the covariance matrix $\Sigma$. Examples of such network include **hoberg2016TextBasedNetwork**, who identifies a product similarity network from financial reports that has been shown to be more accurate than industry block diagonal matrix. As linked firms are potentially subject to similar demand shock, we have reason to believe that $G$ contains valuable information about the comovement among the returns. **israelsen2016does** and **kaustia2020CommonAnalysts** both find that companies covered by the same analysts show similarities in many unobserved dimensions, and this analyst-based network could explain excess co-movement on top of common factors. With the development of machine learning techniques such as textual analysis, we are better at acquiring information from big data. Granular linkage information among firms that used to be notoriously hard to get due to its proprietary properties, now are becoming available to researchers. The question is, how to use those auxiliary network information to better estimate the co-movement between assets?

This paper aims to provide ways to extract the information contained in the auxiliary $G$ matrix to help estimate the covariance $\Sigma$. We consider an *Adaptive Correlation Thresholding* method, where we apply thresholding to the correlation matrix, with the threshold level depending on network information. More specifically, suppose we observe $Y_t$ for $t = 1, \ldots, T$, the procedure is

1. Estimate the sample covariance estimate $\hat{\Sigma}$, and the sample correlation matrix $\hat{R}$.
2. Apply the generalized thresholding function $h(r_{ij}, \tau_{ij})$ to the off-diagonal elements of $\hat{R} = (\hat{r}_{ij})$, as in **rothman2009GeneralizedThresholding**. The novelty

is now we allow the threshold $\tau_{ij}$ to vary across elements and to depend on the network information. Specifications we have considered for the threshold $\tau$ are

- Simple linear model

$$\tau(G_{ij}) = a + bG_{ij}$$

- The probit model

$$\tau_{ij} = \tau\left(G_{ij}\right) = \Phi\left(a + b\left|G_{ij}\right|\right)$$

3. Estimate the unknown parameters in the $\tau$ function by cross validation, as in **bickel2008CovarianceRegularization**, **cai2011AdaptiveThresholding**, where we randomly split the sample $V$ times, for each $v$, compute the new estimator $\hat{\Sigma}_G^{1,v}$ with the first subsample, and sample covariance $\hat{\Sigma}^{2,v}$ and the criterion is

$$L(a, b) = \frac{1}{V}\sum_{v}^{V}\left\|\hat{\Sigma}_G^{1,v} - \hat{\Sigma}^{2,v}\right\|_F^2$$

   we find $a, b$ that minimise this criterion.

4. Then with the estimates of $a, b$, we can estimate $\Sigma$ on the test sample.

There are several advantages of using network guided method:

1. The main advantage is that we are combining economically meaningful network with market-based performance data. Comparing to purely data-driven thresholding or shrinkage methods, the method utilizes valuable information embedded in external network data, which provides more robustness and efficiency. aif our auxiliary network contains the "real" links from the network. The relationship identified will be more stable over time than the relationship identified from return data alone.

2. This method is very flexible and extensible. Although in our current analysis we only use one of the existing networks as our proxy for $G$, you are free to include many candidate networks in the $\tau$. You may want want to include characteristics-based distances, as it has been documented that companies with similar characteristics exhibit additional co-movement on top of common risk

factors (see **fernandez2011spatial** for example). It also provides a way to discern which set of information is relevant based an estimate of the coefficients $a, b$ in the thresholding level.

3. The networks may provide industry-level comovement that is potentially related to the "weak factors" components, which we intend to investigate.

## 2 Literature Review

There has been extensive research on high-dimensional covariance estimation. Some important lines of thinking include element-wise banding and thresholding method, shrinkage method, factor models, etc. For a book-length review see **pourahmadi2013HighdimensionalCo**

**bickel2008CovarianceRegularization** considers banding or tapering the sample covariance matrix. **bickel2008CovarianceRegularization** considers covariance regularization by hard thresholding. They also compare the results between banding when there is a natural ordering(for example, time series autocorrelation) and thresholding where we need to pay a $\log p$ price in the convergence rate to learn the locations. **cai2011AdaptiveThresholding** considers adptive thresholding where threshold takes the form:

$$\hat{\sigma}_{ij}^{*} = s_{t_{ij}}(\hat{\sigma}_{ij}) \tag{1}$$

where 1. $|s_\lambda(z)| \leq c|y|$ for all $|z - y| \leq \lambda$ 2. $s_\lambda(z) = 0$ for $|z| \leq \lambda$ 3. $|s_\lambda(z) - z| \leq \lambda$. The convergence rate is the same, although here the uniformity class is larger. **fan2015OverviewEstimation** proposes thresholding on the correlation matrix. The choice of thresholding functions can be found **rothman2009GeneralizedThresholding**, **fan2001VariableSelection**, etc.

As an application of thresholding method, **fan2016IncorporatingGlobal** use hard thresholding method in a high-frequency setting based on the sector/industry classifier. $s_{ij}(\sigma_{ij}) = \sigma_{ij}$ if $ij$ are in the same industry. The network they use is a block-diagonal matrix and our results accommodate more general and flexible network information.

**ledoit2004HoneyShrunk** develops an estimation strategy based on linear shrinkage, where the target is identity matrix. This shrinkage guarantees that the estimated covariance matrix is well-conditioned. This approach can be thought of as decreasing variance at the expense of increasing bias a little. There are articles discuss multiple targets, for example, **schafer2005ShrinkageApproach**, **lancewicki2014MultiTargetShrinkage** and **gray2018ShrinkageEstimation**, but their targets are either fixed or data-driven, so different from our guided method where we bring in new information from auxiliary network information. **ledoit2012NonlinearShrinkage** and **ledoit2017NonlinearShrinkage** propose nonlinear shrinkage where the eigenvalues are pulled towards the "correct level" solving a nonrandom limit loss function. The shrinkage method has been shown to have really good performance in estimating large-dimensional covariance matrix, however they are a global method whereas our method is designed to emphasize "economically meaning" links. There is also a vast literature on factor models in high-dimensional models and applications in empirical finance. We refer to **connor2012EfficientSemiparametr** **fan2015OverviewEstimation** and **fan2016ProjectedPrincipal** and literature review therein.

## 3 Simulations

We demonstrate the network guided estimator and examine its small-sample performance using the following simple simulations. First, we consider the case where the true covariance $\Sigma$ comes from an AR(1) model. So for $\{(i,j) : i = 1, \ldots, N, j = 1, \ldots, N\}$, $\sigma_{ij}^2 = \sigma_i \sigma_j \rho_{ij}$ and $\rho_{ij} = \rho^{|i-j|}$. We take $N = 400$.

$$S_{ij} = 3 * \rho^{|i-j|}$$

and assume we observe a matrix $G(l)$ indicating the location of highly correlated pairs $L_{ij}(l) = \mathbf{1}\{\rho_{ij} \geq l\}$. Conditional on $L_{ij} = 1$, we observe $G_{ij} = 1$ with probability $p$ and conditional on $L_{ij} = 0$ , $G_{ij} = 1$ with probability $q$. Hence $p, q$ reflect the probability of missing important locations and including false important locations respectively.

We then generate $T = 100$ independent drws of observations $X \sim N(0, \Sigma)$ and estimate $\Sigma$ using 1. Sample covariance ; 2. Linear Shrinkage estimator; 3. Nonlinear Shrinkage estimator; 4. Universal thresholding on the correlation; 5. and Network

Guided thresholding estimator. We now compare their performance. It's worth collecting here the parameters that we will adjust in the experiments:

| Parameter | Description |
|---|---|
| $\rho$ | Determines how strong the correlation is and the sparsity of the covariance matrix $\Sigma$ |
| $l$ | Observation level, determines how we classify a pair $(i, j)$ as important, i.e., $L_{ij} = 1$. |
| $p$ | Conditional on $L_{ij} = 1$, the probability of actually observing $G_{ij} = 1$. |
| $q$ | Conditional on $L_{ij} = 0$, the probability of observing $G_{ij} = 1$ |
| $\tau$ | The Threshold level when we apply generalized thresholding operator on $\sigma_{ij}$ where $G_{ij} = 0$. |

Table 1: Description of varying parameters.

Table 2: The estimation error of various estimators in terms of the Frobenius Norm

| $rho$ | Threshold Level | Sample Cov | Linear Shrinkage | Nonlinear Shrinkage | Universal Threshold | Network Guided |
|---|---|---|---|---|---|---|
| 0.70 | 0.0 | 59.43 | 42.23 | 41.61 | 59.43 | 59.43 |
| | 0.1 | 59.43 | 42.23 | 41.61 | 49.91 | 49.85 |
| | 0.2 | 59.43 | 42.23 | 41.61 | 41.94 | 41.67 |
| | 0.3 | 59.43 | 42.23 | 41.61 | 35.66 | 34.97 |
| | 0.4 | 59.43 | 42.23 | 41.61 | 31.20 | 29.81 |
| | 0.5 | 59.43 | 42.23 | 41.61 | 28.56 | 26.18 |
| | 0.6 | 59.43 | 42.23 | 41.61 | 27.54 | 23.93 |
| | 0.7 | 59.43 | 42.23 | 41.61 | 27.78 | 22.81 |
| | 0.8 | 59.43 | 42.23 | 41.61 | 28.83 | 22.48 |
| | 0.9 | 59.43 | 42.23 | 41.61 | 30.33 | 22.59 |
| | 1.0 | 59.43 | 42.23 | 41.61 | 32.08 | 22.98 |
| 0.80 | 0.0 | 62.54 | 47.59 | 46.59 | 62.54 | 62.54 |

Continued on next page

Table 2: The estimation error of various estimators in terms of the Frobenius Norm

| $rho$ | Threshold Level | Sample Cov | Linear Shrinkage | Nonlinear Shrinkage | Universal Threshold | Network Guided |
|---|---|---|---|---|---|---|
| | 0.1 | 62.54 | 47.59 | 46.59 | 52.60 | 52.80 |
| | 0.2 | 62.54 | 47.59 | 46.59 | 44.30 | 44.52 |
| | 0.3 | 62.54 | 47.59 | 46.59 | 37.89 | 37.85 |
| | 0.4 | 62.54 | 47.59 | 46.59 | 33.55 | 32.85 |
| | 0.5 | 62.54 | 47.59 | 46.59 | 31.32 | 29.48 |
| | 0.6 | 62.54 | 47.59 | 46.59 | 30.95 | 27.58 |
| | 0.7 | 62.54 | 47.59 | 46.59 | 31.93 | 26.78 |
| | 0.8 | 62.54 | 47.59 | 46.59 | 33.75 | 26.71 |
| | 0.9 | 62.54 | 47.59 | 46.59 | 36.02 | 27.05 |
| | 1.0 | 62.54 | 47.59 | 46.59 | 38.49 | 27.59 |
| 0.90 | 0.0 | 63.06 | 53.18 | 52.83 | 63.06 | 63.06 |
| | 0.1 | 63.06 | 53.18 | 52.83 | 53.98 | 54.60 |
| | 0.2 | 63.06 | 53.18 | 52.83 | 46.91 | 47.84 |
| | 0.3 | 63.06 | 53.18 | 52.83 | 42.16 | 42.89 |
| | 0.4 | 63.06 | 53.18 | 52.83 | 39.78 | 39.67 |
| | 0.5 | 63.06 | 53.18 | 52.83 | 39.57 | 37.96 |
| | 0.6 | 63.06 | 53.18 | 52.83 | 41.08 | 37.43 |
| | 0.7 | 63.06 | 53.18 | 52.83 | 43.74 | 37.69 |
| | 0.8 | 63.06 | 53.18 | 52.83 | 47.08 | 38.41 |
| | 0.9 | 63.06 | 53.18 | 52.83 | 50.74 | 39.33 |
| | 1.0 | 63.06 | 53.18 | 52.83 | 54.57 | 40.32 |
| 0.95 | 0.0 | 57.97 | 52.58 | 51.77 | 57.97 | 57.97 |
| | 0.1 | 57.97 | 52.58 | 51.77 | 51.42 | 52.21 |
| | 0.2 | 57.97 | 52.58 | 51.77 | 47.65 | 48.39 |
| | 0.3 | 57.97 | 52.58 | 51.77 | 46.74 | 46.40 |

Continued on next page

Table 2: The estimation error of various estimators in terms of the Frobenius Norm

| $rho$ | Threshold Level | Sample Cov | Linear Shrinkage | Nonlinear Shrinkage | Universal Threshold | Network Guided |
|---|---|---|---|---|---|---|
| | 0.4 | 57.97 | 52.58 | 51.77 | 48.35 | 45.96 |
| | 0.5 | 57.97 | 52.58 | 51.77 | 51.82 | 46.66 |
| | 0.6 | 57.97 | 52.58 | 51.77 | 56.46 | 48.04 |
| | 0.7 | 57.97 | 52.58 | 51.77 | 61.73 | 49.73 |
| | 0.8 | 57.97 | 52.58 | 51.77 | 67.33 | 51.51 |
| | 0.9 | 57.97 | 52.58 | 51.77 | 73.05 | 53.24 |
| | 1.0 | 57.97 | 52.58 | 51.77 | 78.77 | 54.84 |
| 0.99 | 0.0 | 104.67 | 115.10 | 106.32 | 104.67 | 104.67 |
| | 0.1 | 104.67 | 115.10 | 106.32 | 114.91 | 106.31 |
| | 0.2 | 104.67 | 115.10 | 106.32 | 125.35 | 108.13 |
| | 0.3 | 104.67 | 115.10 | 106.32 | 135.88 | 110.08 |
| | 0.4 | 104.67 | 115.10 | 106.32 | 146.37 | 112.10 |
| | 0.5 | 104.67 | 115.10 | 106.32 | 156.67 | 114.08 |
| | 0.6 | 104.67 | 115.10 | 106.32 | 166.73 | 115.97 |
| | 0.7 | 104.67 | 115.10 | 106.32 | 176.56 | 117.77 |
| | 0.8 | 104.67 | 115.10 | 106.32 | 186.19 | 119.50 |
| | 0.9 | 104.67 | 115.10 | 106.32 | 195.60 | 121.16 |
| | 1.0 | 104.67 | 115.10 | 106.32 | 204.74 | 122.77 |

Table 3: The estimation error of various estimators in terms of the Matrix-1 Norm

| $rho$ | Threshold Level | Sample Cov | Linear Shrinkage | Nonlinear Shrinkage | Universal Threshold | Network Guided |
|---|---|---|---|---|---|---|
| 0.70 | 0.0 | 63.77 | 33.42 | 33.94 | 63.77 | 63.77 |
| | 0.1 | 63.77 | 33.42 | 33.94 | 49.17 | 49.42 |
| | 0.2 | 63.77 | 33.42 | 33.94 | 37.39 | 38.23 |
| | 0.3 | 63.77 | 33.42 | 33.94 | 28.37 | 29.20 |
| | 0.4 | 63.77 | 33.42 | 33.94 | 22.21 | 22.31 |
| | 0.5 | 63.77 | 33.42 | 33.94 | 18.53 | 17.86 |
| | 0.6 | 63.77 | 33.42 | 33.94 | 16.18 | 14.59 |
| | 0.7 | 63.77 | 33.42 | 33.94 | 14.59 | 12.45 |
| | 0.8 | 63.77 | 33.42 | 33.94 | 13.41 | 11.03 |
| | 0.9 | 63.77 | 33.42 | 33.94 | 12.65 | 10.22 |
| | 1.0 | 63.77 | 33.42 | 33.94 | 12.56 | 9.85 |
| 0.80 | 0.0 | 73.98 | 44.02 | 42.83 | 73.98 | 73.98 |
| | 0.1 | 73.98 | 44.02 | 42.83 | 58.63 | 59.50 |
| | 0.2 | 73.98 | 44.02 | 42.83 | 45.86 | 47.60 |
| | 0.3 | 73.98 | 44.02 | 42.83 | 35.31 | 37.91 |
| | 0.4 | 73.98 | 44.02 | 42.83 | 26.69 | 30.04 |
| | 0.5 | 73.98 | 44.02 | 42.83 | 23.68 | 24.71 |
| | 0.6 | 73.98 | 44.02 | 42.83 | 21.47 | 20.60 |
| | 0.7 | 73.98 | 44.02 | 42.83 | 20.10 | 17.40 |
| | 0.8 | 73.98 | 44.02 | 42.83 | 19.66 | 15.48 |
| | 0.9 | 73.98 | 44.02 | 42.83 | 19.47 | 14.92 |
| | 1.0 | 73.98 | 44.02 | 42.83 | 19.72 | 14.85 |
| 0.90 | 0.0 | 69.65 | 60.41 | 60.57 | 69.65 | 69.65 |
| | 0.1 | 69.65 | 60.41 | 60.57 | 57.72 | 58.98 |
| | 0.2 | 69.65 | 60.41 | 60.57 | 50.38 | 49.66 |

Continued on next page

Table 3: The estimation error of various estimators in terms of the Matrix-1 Norm

| $rho$ | Threshold Level | Sample Cov | Linear Shrinkage | Nonlinear Shrinkage | Universal Threshold | Network Guided |
|---|---|---|---|---|---|---|
| | 0.3 | 69.65 | 60.41 | 60.57 | 46.30 | 42.43 |
| | 0.4 | 69.65 | 60.41 | 60.57 | 43.35 | 38.31 |
| | 0.5 | 69.65 | 60.41 | 60.57 | 41.31 | 35.10 |
| | 0.6 | 69.65 | 60.41 | 60.57 | 39.91 | 32.66 |
| | 0.7 | 69.65 | 60.41 | 60.57 | 39.22 | 31.48 |
| | 0.8 | 69.65 | 60.41 | 60.57 | 39.73 | 30.93 |
| | 0.9 | 69.65 | 60.41 | 60.57 | 40.38 | 30.56 |
| | 1.0 | 69.65 | 60.41 | 60.57 | 41.17 | 30.41 |
| 0.95 | 0.0 | 95.18 | 94.47 | 92.88 | 95.18 | 95.18 |
| | 0.1 | 95.18 | 94.47 | 92.88 | 88.39 | 85.83 |
| | 0.2 | 95.18 | 94.47 | 92.88 | 81.86 | 76.60 |
| | 0.3 | 95.18 | 94.47 | 92.88 | 76.00 | 68.02 |
| | 0.4 | 95.18 | 94.47 | 92.88 | 71.17 | 60.34 |
| | 0.5 | 95.18 | 94.47 | 92.88 | 67.30 | 57.41 |
| | 0.6 | 95.18 | 94.47 | 92.88 | 69.93 | 56.36 |
| | 0.7 | 95.18 | 94.47 | 92.88 | 73.19 | 56.04 |
| | 0.8 | 95.18 | 94.47 | 92.88 | 76.37 | 55.70 |
| | 0.9 | 95.18 | 94.47 | 92.88 | 79.61 | 55.75 |
| | 1.0 | 95.18 | 94.47 | 92.88 | 82.59 | 56.93 |
| 0.99 | 0.0 | 49.73 | 43.91 | 50.83 | 49.73 | 49.73 |
| | 0.1 | 49.73 | 43.91 | 50.83 | 46.98 | 48.49 |
| | 0.2 | 49.73 | 43.91 | 50.83 | 58.00 | 47.60 |
| | 0.3 | 49.73 | 43.91 | 50.83 | 70.41 | 50.18 |
| | 0.4 | 49.73 | 43.91 | 50.83 | 82.87 | 54.76 |
| | 0.5 | 49.73 | 43.91 | 50.83 | 94.81 | 58.76 |

Continued on next page

Table 3: The estimation error of various estimators in terms of the Matrix-1 Norm

| $rho$ | Threshold Level | Sample Cov | Linear Shrinkage | Nonlinear Shrinkage | Universal Threshold | Network Guided |
|---|---|---|---|---|---|---|
| | 0.6 | 49.73 | 43.91 | 50.83 | 105.96 | 61.89 |
| | 0.7 | 49.73 | 43.91 | 50.83 | 115.92 | 63.92 |
| | 0.8 | 49.73 | 43.91 | 50.83 | 128.73 | 66.98 |
| | 0.9 | 49.73 | 43.91 | 50.83 | 141.60 | 69.59 |
| | 1.0 | 49.73 | 43.91 | 50.83 | 154.46 | 71.30 |

In Table 2, we show the general performance of these estimators when we simulate using different $\rho$ and thresholding level $\tau$. Here we have taken the thresholding operator to be soft thresholding. It can be seen that generally speaking, when the covariance matrix becomes denser, linear, nonlinear shrinakge estimators and the sample covariance estimator become superiro to

Then we consider simulations with varying observation levels $l$. In Figure 1 when we set observation level equal to 0, the network guided estimator will be the same as the sample covariance estimator, on the other extreme, when observation level is set to 1, the network guided estimator is equivalent to universal thresholding. In between these cases, when we have information about the locations of the important pairs, we have a range where the estimation error is lowered.

Table 4: The estimation error of the Network Guided Estimator with varying probabilities $p$, $q$ that determine how $G$ is generated.

| q<br>p | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 47.09 | 48.44 | 49.91 | 51.19 | 52.43 | 53.68 | 54.89 | 56.14 | 57.37 | 58.50 | 59.61 |
| 0.1 | 46.69 | 48.00 | 49.51 | 50.80 | 52.11 | 53.36 | 54.62 | 55.92 | 57.09 | 58.09 | 59.32 |
| 0.2 | 46.35 | 47.81 | 49.19 | 50.38 | 51.73 | 53.10 | 54.43 | 55.47 | 56.76 | 57.80 | 58.99 |

Continued on next page

Table 4: The estimation error of the Network Guided Estimator with varying probabilities $p$, $q$ that determine how $G$ is generated.

| q | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| p | | | | | | | | | | | |
| 0.3 | 45.91 | 47.40 | 48.82 | 50.12 | 51.50 | 52.80 | 53.88 | 55.10 | 56.39 | 57.47 | 58.79 |
| 0.4 | 45.53 | 47.08 | 48.40 | 49.61 | 51.01 | 52.37 | 53.61 | 54.72 | 56.00 | 57.14 | 58.45 |
| 0.5 | 45.15 | 46.74 | 48.00 | 49.46 | 50.71 | 52.16 | 53.33 | 54.60 | 55.77 | 57.00 | 57.99 |
| 0.6 | 44.78 | 46.14 | 47.79 | 48.98 | 50.38 | 51.47 | 52.93 | 54.10 | 55.32 | 56.59 | 57.85 |
| 0.7 | 44.37 | 45.94 | 47.10 | 48.96 | 50.10 | 51.29 | 52.52 | 53.96 | 55.03 | 56.23 | 57.44 |
| 0.8 | 44.06 | 45.58 | 46.85 | 48.36 | 49.66 | 50.87 | 52.15 | 53.62 | 54.81 | 55.90 | 57.22 |
| 0.9 | 43.51 | 45.06 | 46.28 | 48.02 | 49.33 | 50.56 | 51.99 | 53.27 | 54.46 | 55.69 | 56.82 |
| 1.0 | 43.14 | 44.71 | 46.08 | 47.52 | 48.87 | 50.15 | 51.69 | 52.80 | 54.13 | 55.29 | 56.53 |

In Table 4, we have when $p = q = 0$ the estimation error of the universal thresholding estimator, and $p = q = 1$ the sample covaraince estimation error. As we can see, as long as $q$ is not large, the estimation error will be smaller when we have a higher probability $p$ of observing the true large elements. It should be noted that $q$ in fact cannot be very large, given that the whole matrix is sparse.

## 4 Empirical Study

In this section, we apply the adaptive correlation thresholding method to a portfolio construction problem. First we describe the procedure and then present some of the results we have.

Assume we observe the excess return $Y_{it}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$ follows

$$Y_{it} = B_i' F_t + u_{it}; \quad \Sigma_u = E(u_t u_t')$$

where $F_t$ are factor excess returns. Here we have considered Fama-French 3 and the Carhart's momentum factor. The goal is to estimate $\Sigma_Y = E(YY')$ and use that estimate to construct portfolio following **ledoit2017NonlinearShrinkage**. The auxialiary network we have include the **hoberg2016TextBasedNetwork**'s network(henceforth

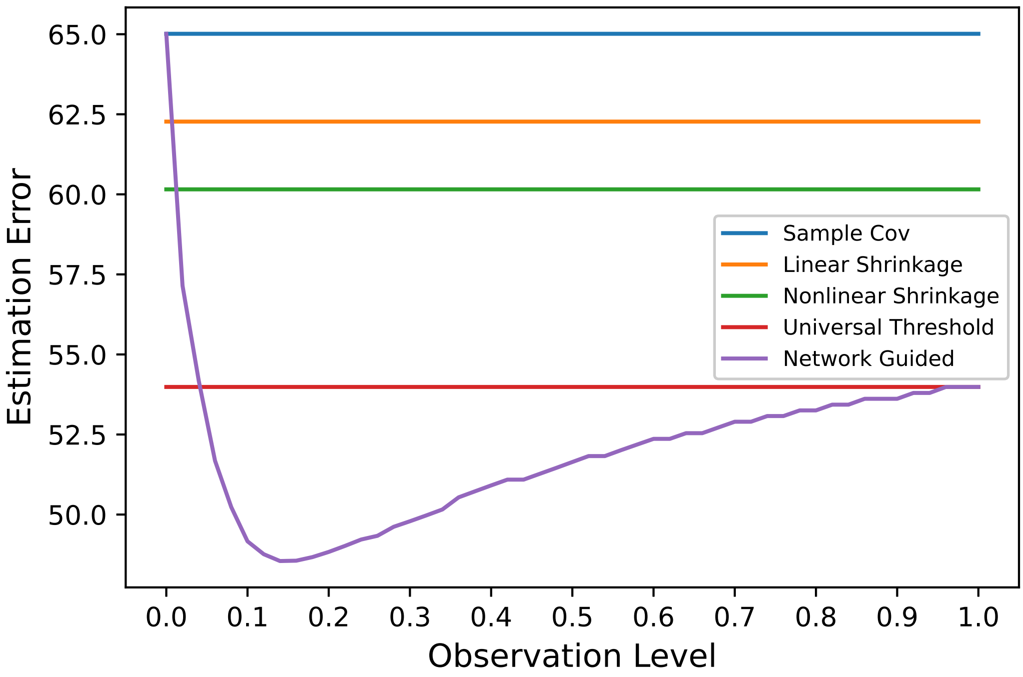

Figure 1: The estimation error against the observation level

Hoberg's Network) and IBES analysts cocoverage network. Here we present the results for SP500 returns using Hoberg's Network.

The procedure we take is as follows.

1. We run time series linear regressions of $Y_{it}$ on $F_{k,t}$, obtain the beta estimates $\hat{B}_i$ and the residual $\hat{u}_{it}$.

2. Compute the covaraince matrix $S_{\hat{u}} = \frac{1}{T}\hat{u}\hat{u}'$ and $S_F = \frac{1}{T}\sum_t (F_t - \bar{F})(F_t - \bar{F})'$ and appply *adaptive correlation thresholding* on $S_{\hat{u}}$, denote the estimate as $\hat{S}_{\hat{u}}$. where the second step adaptive correlation thresholding is achieved in the following way. Let $R_u$ be the correlation matrix calculated from $S_u$. We use soft thresholding $h(r_{ij}, \tau_{ij}) = \text{sign}\left(r_{ij}\right)\left(r_{ij} - \tau_{ij}\right)_+$ on the off-diagonal elemetns $r_{ij}$ of $R_u$, where

$$\tau_{ij} = \delta_{ij}\sqrt{\frac{\log N}{T}}$$

   and

$$\delta_{ij} = a + bG_{ij}$$

   Let the threshold estimate be $\hat{R}_{\hat{u}}(a, b)$, given $a, b$, our estimate will be

$$\hat{S}_{\hat{u}} = \hat{S}_{\hat{u}}(a, b) = \text{diag}(S_{\hat{u}})^{\frac{1}{2}}\hat{R}_{\hat{u}}\,\text{diag}(S_{\hat{u}})^{\frac{1}{2}}$$

   In order to guarantee positive definiteness, I follow the suggestion in **fan2015OverviewEstimation** and **fan2013LargeCovariance**, by first finding the minimum $\underline{\delta}$ such that the $\hat{S}(\delta, 0)$ has its smallest eigenvalue larger than 0 if we choose $\tau_{ij} = \underline{\delta}\sqrt{\frac{\log N}{T}}$.

   Then $a, b$ are estimated using cross-validation following **bickel2008CovarianceRegularization** by randomly spliting the sample $V$ times, for each $v = 1, \ldots, V$, compute the estimate $\hat{S}_u^{1,v}$ with the first subsample, and sample covariance estimate $\hat{\Sigma}_u^{2,v}$ with the second subsample and let the criterion function be

$$L(a, b) = \frac{1}{V}\sum_v^V \left\|\hat{S}_u^{1,v} - \hat{\Sigma}_u^{2,v}\right\|_F^2$$

we find $\hat{a}, \hat{b}$ that minimise this criterion subject to the constraints:

$$0 \leq a\sqrt{\frac{\log N}{T}} \leq 1 \tag{2}$$

$$b\sqrt{\frac{\log N}{T}} \leq 0 \tag{3}$$

$$\underline{\delta} \leq a + b \tag{4}$$

3. Construct an estimate of $\Sigma_Y$ by $\hat{\Sigma}_Y = \hat{B}S_F\hat{B}' + \hat{S}_{\hat{u}}$

We have estimated the covariance matrices of SP500 stocks from 1996 to the end of 2017; using stock return data and Fama-French 3 factor returns $F_{kt}, k = 1, 2, 3$. We incorporate Hoberg's network $G_t$ that are updated yearly into our estimation procedure.

The Hoberg's Network is a yearly updated $N \times N$ network with elements in $[0, 1]$ with higher score $G_{ij}$ reflecting potentially higher correlation between the $i$-th and $j$-th firms.

In Figure 1, we present the distribution, we present the distribution (blue) of sample covariance estimates $S_{\hat{u},ij}$ of residuals after regressing the the stocks returns on the Fama-French 3 factor for the stocks that haIn Figure 1, we present the distribution (blue) of sample covariance estimates $S_{\hat{u},ij}$ of residuals after regressing the the stocks returns on the Fama-French 3 factor for the stocks tsiduals after regressing the the stocks returns on the Fama-French 3 factor for the stocks that have no missing data in the dataset; alongwith the distribution of sample covariance $S_{\hat{u},ij}$ for those $ij$ with $G_{ij} > 0$ in the Hoberg's network. It's clear that the distribution is shifted to the right, implying that Hoberg's network can pick up information that are not explained by the factors.

Then we use a rolling-window estimation, with 252-day estimation period and then move the window forward by 21 days. In the estimation periods in window $m = 1, \ldots, M$ we construct estimate $\hat{\Sigma}_{Y,m} = \hat{B}S_F\hat{B}' + \hat{S}_{\hat{u}}\left(\hat{a}_m, \hat{b}_m\right)$. The estimated parameters $\hat{a}_m, \hat{b}_m$ have mean $(1.177, -0.252)$, where the $\hat{b}$ measures the effect of knowing the auxiliary network on the thresholding level.

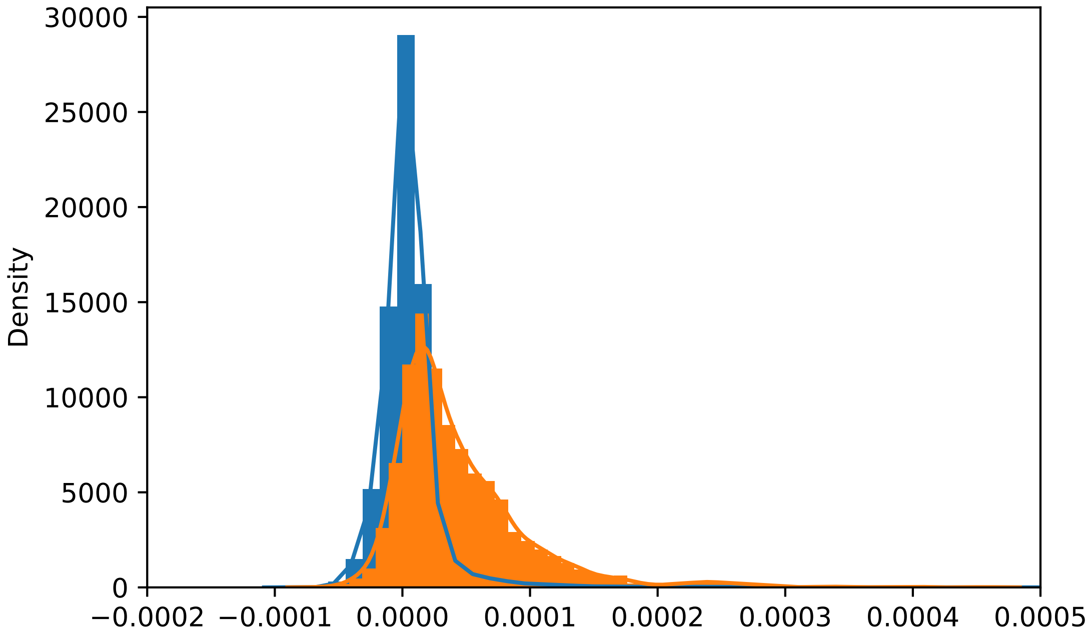

Figure 2: Distribution of $S_{\hat{u},ij}$, $i, j = 1, \ldots, N$ and $S_{\hat{u},ij}$ for $i, j$ such that $G_{ij} > 0$

# 5 Conclusion and Further Works

This paper considers the problem of incorporating ever-increasing auxiliary data from machine learning techniques such as textual analysis into the estimation of large covariance matrices. This current version is preliminary with ongoing research on the following applications.

Firstly, we are applying the covaraince estimation technique on portfolio construction, following the problem considered in **ledoit2004HoneyShrunk** and **ledoit2017NonlinearShrinkage**, where the estimation of the sparse covaraince matrices are vital for constructing the minimum-variance portfolio.

Secondly, the method can be applied to study spatial-APT under large $N$ case. **kou2018asset** finds that common risk factors are insufficient to capture all the significant inter-dependencies in asset returns, and local interactions are also important. Spatial-APT and spatial CAPM type of models have not been popular in large N case since the measure of contiguity is challenging. Our method can uncover contemporaneously correlated entities by combining market-based information and auxiliary network information, thus providing a natural contiguity measure. Relying solely on either statistical methods or external network information is not as desirable as the links identified by the former are hard to interpret and the external network may miss some important links.

Thirdly, we are expanding the set of auxialiary networks beyond the Hoberg's network as well as applying the technique on larger datasets. We have collected IBES analysts cocovarage network and are constructing new network based on firms' characteristics. The flexibility of the methods allows us many potential improvements.