SHRINKAGE ESTIMATION OF HIGH DIMENSIONAL COVARIANCE MATRICES

Yilun Chen, Ami Wiesel, Alfred O. Hero III

Department of EECS, University of Michigan, Ann Arbor, MI 48109, USA
{yilun,amiw,hero}@umich.edu

ABSTRACT

We address covariance estimation under mean-squared loss in the Gaussian setting. Specifically, we consider shrinkage methods which are suitable for high dimensional problems with small number of samples (large p small n). First, we improve on the Ledoit-Wolf (LW) method by conditioning on a sufficient statistic via the Rao-Blackwell theorem, obtaining a new estimator RBLW whose mean-squared error dominates the LW under Gaussian model. Second, to further reduce the estimation error, we propose an iterative approach which approximates the clairvoyant shrinkage estimator. Convergence of this iterative method is proved and a closed form expression for the limit is determined, which is called the OAS estimator. Both of the proposed estimators have simple expressions and are easy to compute. Although the two methods are developed from different approaches, their structure is identical up to specific constants. The RBLW estimator provably dominates the LW method; and numerical simulations demonstrate that the OAS estimator performs even better, especially when n is much less than p.

Index Terms— Shrinkage, covariance estimation, Rao-Blackwell, mean-squared loss

1. INTRODUCTION

Covariance matrix estimation is a fundamental problem in signal processing and related fields. Different application varying from array processing [6] to functional genomics [7] rely on accurately estimated covariance matrices. In recent years, estimation of high dimensional $p \times p$ covariance matrices under small sample size n has attracted considerable interest. Examples include gene expression arrays, financial forecasting, spectroscopic imaging, fMRI data and many others. Classical estimation methods perform poorly in these settings and this is the main motivation for this work.

The sample covariance is most commonly used as an estimate for the unknown covariance matrix. When it is invertible, the sample covariance coincides with the classical maximum likelihood estimate. However, while it is an unbiased estimator, it does not minimize the mean-squared error. Indeed, Stein demonstrated that superior performance may be obtained by shrinking the sample covariance towards a structured estimate [1]. Since then, many shrinkage estimators have been proposed under different performance measures, *e.g.*, [2, 3, 4]. The majority of these works addressed the case of invertible sample covariance when $n > p$. Recently, Ledoit and Wolf (LW) proposed a shrinkage estimator for the case $n < p$ which asymptotically minimizes the mean-squared error in the covariance [5]. The estimator is well conditioned under small sample sizes and can be applied to high dimensional problems. In contrast to the previous works, the performance advantages are not restricted to the Gaussian assumption and are distribution free.

In this paper, we show that the LW estimator can be significantly improved when the sample is Gaussian. We begin by providing a closed form expression for the optimal clairvoyant shrinkage estimator under mean-squared loss criteria. This estimator is an explicit function of the unknown covariance matrix that can be used as an oracle performance bound. Our first estimator is obtained by applying the classical Rao-Blackwell theorem [9] to the LW method, and is therefore denoted by RBLW. After tedious integral computations, we can obtain a simple closed form estimator which provably dominates the LW method in terms of mean-squared loss. We then introduce an iterative shrinkage estimator which tries to approximate the oracle. Beginning with an initial rough estimate, each iteration is defined as the oracle solution where the unknown covariance is replaced by its estimate obtained in the previous iteration. Remarkably, a closed form expression can be determined for the limit of these iterations, called the oracle approximating shrinkage (OAS) estimator.

The OAS and RBLW estimators share similar structure. In fact, we show that this special structure is related to the locally most powerful invariant test for covariance sphericity [10]. Both methods are simple, easy to compute and perform well with finite sample size. The RBLW estimator provably dominates the LW and our numerical results demonstrate that for small sample sizes, the OAS estimator is superior to both the RBLW and the LW techniques.

The paper is organized as follows. Section 2 provides a formulation of the problem. We then develop the oracle estimator, the RBLW estimator and the OAS estimator in Section 3. Section 4 includes numerical simulation results and we conclude the paper in Section 5.

Notation: In the following of the paper, $(\cdot)^T$ denotes the transpose operator, $\mathrm{tr}\,(\cdot)$ denotes the trace operator, $E\,[\cdot]$ and $E\,[\cdot\,|\cdot\,]$ denote the expectation and conditional expectation respectively, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, and $|\cdot|$ denotes the determinant of a matrix or the absolute value of a scalar.

2. PROBLEM FORMULATION

Let $\{X_i\}_{i=1}^n$ be a sample of independent identical distributed p-dimensional Gaussian vectors with zero mean and covariance Σ. Note that we do not assume $n \geq p$. Given these realizations, our goal is to find an estimator $\hat{\Sigma}\left(\{X_i\}_{i=1}^n\right)$ which minimizes the mean-squared error:

$$E\left[\left\|\hat{\Sigma}\left(\{X_i\}_{i=1}^n\right)-\Sigma\right\|_F^2\right]. \tag{1}$$

It is impractical to minimize this loss without additional constraints and therefore we restrict ourselves to a specific class of estimators that employ shrinkage [1, 8]. The classical estimator is the

This research was partially supported by the NSF grant CCF 0830490. The work of A. Wiesel was supported by a Marie Curie Outgoing International Fellowship within the 7th European Community Framework Programme.

sample covariance $\hat{S}$ defined as

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T. \tag{2}$$

On the other hand, if we assume that the elements of X_i are uncorrelated and of equal variance, an intuitive estimate for Σ is

$$\hat{F} = \hat{\nu} I, \tag{3}$$

where $\hat{\nu} = \mathrm{tr}(\hat{S})/p$. This structured estimate will result in reduced variance but will increase the bias when the diagonal assumption is incorrect. A reasonable tradeoff achieved by shrinkage of $\hat{S}$ towards $\hat{F}$ results in the following class of estimators

$$\hat{\Sigma} = \hat{\rho}\hat{F} + (1-\hat{\rho})\hat{S}, \tag{4}$$

parameterized by the shrinkage coefficient $\hat{\rho}$. $\hat{F}$ is also referred to as the shrinkage target.

Altogether, our goal is to find a shrinkage intensity $\hat{\rho}$ as a function of the observations $\{X_i\}_{i=1}^n$ in order to minimize the squared loss in (1).

3. GAUSSIAN SHRINKAGE ESTIMATORS

3.1. The Oracle estimator

The oracle estimator $\hat{\Sigma}_O$ is given by (4) with ρ being the solution to

$$\begin{array}{ll} \min\limits_{\rho} & E\left[\left\|\hat{\Sigma}_O - \Sigma\right\|_F^2\right] \\ s.t. & \hat{\Sigma}_O = \rho\hat{F} + (1-\rho)\,\hat{S} \end{array}. \tag{5}$$

The optimal ρ is provided in the following theorem.

Theorem 1. *Let $\{X_i\}_{i=1}^n$ be independent p-dimensional Gaussian vectors with zero mean and covariance Σ, the optimal solution to (5) is*

$$\rho = \frac{E\left[\mathrm{tr}\left(\left(\Sigma - \hat{S}\right)\left(\hat{F} - \hat{S}\right)\right)\right]}{E\left[\left\|\hat{S} - \hat{F}\right\|_F^2\right]} \tag{6}$$

$$= \frac{(1-2/p)\,\mathrm{tr}\left(\Sigma^2\right) + \mathrm{tr}^2\left(\Sigma\right)}{(n+1-2/p)\,\mathrm{tr}(\Sigma^2) + (1-n/p)\mathrm{tr}^2(\Sigma)}. \tag{7}$$

Equation (6) was proved in [5] and its mean square optimality does not depend on the distribution of signals. Under the additional Gaussian assumption, (7) can be obtained from straightforward evaluation of the expectations.

3.2. The Rao-Blackwell Ledoit-Wolf (RBLW) estimator

The starting point for our derivation of the RBLW estimator is the LW method [5]. Ledoit and Wolf proposed to approximate the oracle (5) using the following consistent estimate of (6):

$$\hat{\rho}_{LW} = \min\left(\frac{\sum\limits_{i=1}^{n}\left\|X_i X_i^T - \hat{S}\right\|_F^2}{n^2\left[\mathrm{tr}\left(\hat{S}^2\right) - \mathrm{tr}^2\left(\hat{S}\right)/p\right]}, 1\right). \tag{8}$$

The LW estimator $\hat{\Sigma}_{LW}$ is then defined by plugging $\hat{\rho}_{LW}$ to (4).

The motivation for the RBLW originates from the fact that under the Gaussian assumption, a sufficient statistic for estimating Σ is the sample covariance $\hat{S}$ in (2). Intuitively, the LW estimator is a function of ancillary and unnecessary statistics and therefore can be improved. Specifically, the Rao-Blackwell theorem [9] states that if $g(X)$ is an estimator of a parameter θ, then the conditional expectation of $g(X)$ given $T(X)$, where T is a sufficient statistic, is typically a better estimator of θ, and is at least never worse under any convex loss criteria. Applying this classical theorem to the LW estimator yields the following theorem.

Theorem 2. *Let $\{X_i\}_{i=1}^n$ be independent p-dimensional Gaussian vectors with zero mean and covariance Σ, then the conditioned expectation of the LW covariance estimator is*

$$\hat{\Sigma}_{RBLW} = E\left[\hat{\Sigma}_{LW}\,\middle|\,\hat{S}\right] \tag{9}$$

$$= \hat{\rho}_{RBLW}\hat{F} + (1-\hat{\rho}_{RBLW})\hat{S}, \tag{10}$$

where

$$\hat{\rho}_{RBLW} = \min\left(\frac{(n-2)/n\cdot\mathrm{tr}\left(\hat{S}^2\right) + \mathrm{tr}^2\left(\hat{S}\right)}{(n+2)\left[\mathrm{tr}\left(\hat{S}^2\right) - \mathrm{tr}^2\left(\hat{S}\right)/p\right]}, 1\right). \tag{11}$$

Due to the Rao-Blackwell theorem, this estimator satisfies

$$E\left[\left\|\hat{\Sigma}_{RBLW} - \Sigma\right\|_F^2\right] \le E\left[\left\|\hat{\Sigma}_{LW} - \Sigma\right\|_F^2\right]. \tag{12}$$

The proof of Theorem 2 is quite involved and is omitted for lack of space.[1] It involves the calculation of expectations conditioned on a Wishart matrix ($n \ge p$) and a singular Wishart matrix ($n < p$) via Haar integrals. Hereby we list some necessary lemmas.

Lemma 1. *If X_i is a $p\times 1$ vector, M is a $p\times p$ positive definite matrix, i.e., $M \succ 0$, then for any integer $m > -2$,*

$$\int_{X_i X_i^T \prec M} \|X_i\|^4 \frac{\left|M - X_i X_i^T\right|^{\frac{1}{2}m}}{|M|^{\frac{1}{2}(m+1)}} dX_i = \frac{\pi^{\frac{p}{2}}}{4}\frac{\Gamma\{m/2+1\}}{\Gamma\{(m+p)/2+3\}}\left[2\mathrm{tr}(M^2) + \mathrm{tr}^2(M)\right], \tag{13}$$

where $\Gamma\{\}$ is the Gamma function defined by $\Gamma\{z\} = \int_o^\infty t^{z-1}e^{-t}dt$.

Lemma 2. *If X_i is a $p\times 1$ Gaussian vector with zero mean and covariance Σ, then*

$$E\left[\left\|X_i\right\|_2^4\middle|\,\hat{S}\right] = \frac{n}{n+2}\left[2\mathrm{tr}\left(\hat{S}^2\right) + \mathrm{tr}^2\left(\hat{S}\right)\right], \tag{14}$$

which holds for both $n \ge p$ and $n < p$.

3.3. The Oracle Approximating Shrinkage (OAS) estimator

The OAS estimator is an iterative approximation for the unimplementable oracle method.[2] We start from any other estimator as an initial guess of Σ and iteratively refine it. The initial guess $\hat{\Sigma}_0$ could be the sample covariance, the RBLW estimate or others. We replace Σ in the oracle estimator by $\hat{\Sigma}_0$ yielding $\hat{\Sigma}_1$ which in turn generates

[1] The reader is referred to [12] for the complete proof.

[2] Note that a similar iteration scheme is also employed in [8] in the context of linear regression.

$\hat{\Sigma}_2$ through our proposed iteration. The iteration process is continued until convergence and the limit defines the OAS estimator, denoted as $\hat{\Sigma}_{OAS}$. Specifically, the proposed iteration is as follows:

$$\hat{\Sigma}_j = \hat{\rho}_j \hat{F} + (1-\hat{\rho}_j)\hat{S}, \tag{15}$$

$$\hat{\rho}_{j+1} = \frac{(1-2/p)\mathrm{tr}\left(\hat{\Sigma}_j\hat{S}\right) + \mathrm{tr}^2\left(\hat{\Sigma}_j\right)}{(n+1-2/p)\mathrm{tr}\left(\hat{\Sigma}_j\hat{S}\right) + (1-n/p)\mathrm{tr}^2\left(\hat{\Sigma}_j\right)}. \tag{16}$$

By comparison between (16) and (11), notice that in (16) $\mathrm{tr}\left(\Sigma\right)$ and $\mathrm{tr}\left(\Sigma^2\right)$ are replaced by $\mathrm{tr}\left(\hat{\Sigma}_j\right)$ and $\mathrm{tr}\left(\hat{\Sigma}_j\hat{S}\right)$, respectively. We use $\mathrm{tr}\left(\hat{\Sigma}_j\hat{S}\right)$ instead of $\mathrm{tr}\left(\hat{\Sigma}_j^2\right)$ since the latter would always forces $\hat{\rho}_j$ to converge to 1 while the former leads to a more meaningful limiting value.

Theorem 3. *The iterative process in (15) $\sim$ (16) converges to the expressions:*

$$\hat{\Sigma}_{OAS} = \hat{\rho}_{OAS}\hat{F} + (1-\hat{\rho}_{OAS})\hat{S}, \tag{17}$$

$$\hat{\rho}_{OAS} = \min\left(\frac{(1-2/p)\mathrm{tr}\left(\hat{S}^2\right) + \mathrm{tr}^2\left(\hat{S}\right)}{(n+1-2/p)\left[\mathrm{tr}\left(\hat{S}^2\right) - \mathrm{tr}^2\left(\hat{S}\right)/p\right]}, 1\right), \tag{18}$$

as long as the initial $\hat{\rho}_0$ is between 0 and 1.

Proof. Substitute (15) into (16). After simplifications we obtain

$$\hat{\rho}_{j+1} = \frac{1-(1-2/p)\,\hat{\phi}\cdot\hat{\rho}_j}{1+n\hat{\phi}-(n+1-2/p)\,\hat{\phi}\cdot\hat{\rho}_j}, \tag{19}$$

where

$$\hat{\phi} = \frac{\mathrm{tr}\left(\hat{S}^2\right) - \mathrm{tr}^2\left(\hat{S}\right)/p}{\mathrm{tr}\left(\hat{S}^2\right) + \mathrm{tr}^2\left(\hat{S}\right)} \in [0,1). \tag{20}$$

Define $\hat{b}_j = \left[1-(n+1-2/p)\,\hat{\phi}\cdot\hat{\rho}_j\right]^{-1}$. Equation (19) is equivalent to

$$\hat{b}_{j+1} = \frac{n\hat{\phi}}{1-(1-2/p)\hat{\phi}}\cdot\hat{b}_j + \frac{1}{1-(1-2/p)\hat{\phi}}, \tag{21}$$

and it is easy to see that

$$\lim_{j\to\infty}\hat{b}_j = \begin{cases} \infty & \frac{n\hat{\phi}}{1-(1-2/p)\hat{\phi}} \geq 1 \\ \frac{1}{1-(n+1-2/p)\hat{\phi}} & \frac{n\hat{\phi}}{1-(1-2/p)\hat{\phi}} < 1 \end{cases}, \tag{22}$$

therefore $\hat{\rho}_j$ also converges as $j\to\infty$ and $\hat{\rho}_I$ is given by

$$\hat{\rho}_{OAS} = \lim_{j\to\infty}\hat{\rho}_j = \begin{cases} \frac{1}{(n+1-2/p)\,\hat{\phi}} & \hat{\phi} \geq \frac{1}{n+1-2/p} \\ 1 & \hat{\phi} < \frac{1}{n+1-2/p} \end{cases}. \tag{23}$$

Equation (18) is obtained by substituting (20) into (23). Therefore, (15) and (16) converge to (17) and (18) as $j\to\infty$. □

3.4. Comparison

It is clear that the $\hat{\rho}_{OAS}$ shares the same structure as $\hat{\rho}_{RBLW}$. In fact, they can both be expressed as

$$\hat{\rho}_{OAS} = \min\left(\alpha_{OAS} + \frac{\beta_{OAS}}{\hat{U}}, 1\right) \tag{24}$$

and

$$\hat{\rho}_{RBLW} = \min\left(\alpha_{RBLW} + \frac{\beta_{RBLW}}{\hat{U}}, 1\right) \tag{25}$$

with $\hat{U}$ defined as

$$\hat{U} = \frac{1}{p-1}\left(\frac{p\cdot\mathrm{tr}\left(\hat{S}^2\right)}{\mathrm{tr}^2\left(\hat{S}\right)} - 1\right), \tag{26}$$

where

$$\alpha_{OAS} = \frac{1}{n+1-2/p}, \quad \beta_{OAS} = \frac{p+1}{(n+1-2/p)(p-1)}, \tag{27}$$

and

$$\alpha_{RBLW} = \frac{n-2}{n(n+2)}, \quad \beta_{RBLW} = \frac{(p+1)n-2}{n(n+2)(p-1)}. \tag{28}$$

Thus the only difference between $\hat{\rho}_{OAS}$ and $\hat{\rho}_{RBLW}$ is the shrinkage coefficients. Interestingly, the statistic $\hat{U}$ is also adopted to test the sphericity of Σ, *i.e.*, testing whether $\Sigma = \nu I$. In particular, $\hat{U}$ is the locally most powerful invariant test statistic for sphericity [10]. The smaller $\hat{U}$ is, the more likely Σ is proportional to an identity matrix I, and the more shrinkage occurs in $\hat{\Sigma}_{OAS}$ and $\hat{\Sigma}_{RBLW}$.

4. NUMERICAL SIMULATIONS

In this section, we compare the RBLW and the OAS with the LW method by numerical simulation. The oracle estimator (5) is also included using the true Σ as a benchmark lower bound of MSE for comparison. For all simulations, we set $p = 100$ and let n range from 5 to 120. Each simulation is repeated 100 times and the averaged MSE and the shrinkage coefficients are plotted as a function of n.

In the first example, we let Σ be the covariance matrix of a Gaussian AR(1) process,

$$\Sigma_{ij} = r^{|i-j|}, \tag{29}$$

where Σ_{ij} denotes the entry of Σ in row i and column j. We take $r = 0.5$ for purposed illustration. Fig. 1 and Fig. 2 show the estimated MSE and shrinkage coefficient respectively. One sees that the OAS performs very closely to the ideal oracle estimator. When n is small compared with p, the OAS significantly outperforms both of the RBLW and the LW. The RBLW improves the LW slightly but this is not easily seen at the scale used for plots in Fig. 1 and Fig. 2. As expected, all the estimators converge towards each other as n increases.

In the second example, we let Σ be the covariance matrix of the increment process of fractional Brownian motion (FBM) which exhibits long-range dependence. Such processes are often used to model Internet traffic [11]. The covariance matrix is given by

$$\Sigma_{ij} = \frac{1}{2}\left[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\right],$$

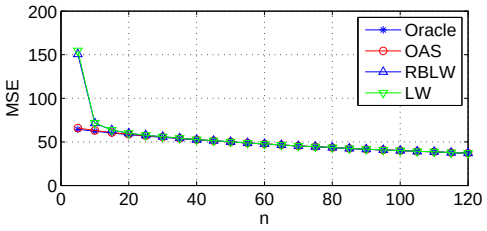

Fig. 1. AR(1) process: Comparison of MSE with different n when $p = 100$, $r = 0.5$.

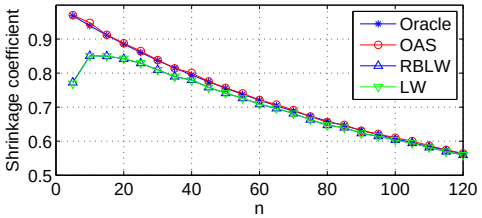

Fig. 2. AR(1) process: Comparison of shrinkage intensity with different n when $p = 100$, $r = 0.5$.

where $H \in [0.5, 1]$ is the Hurst parameter. The typical value of H is below 0.9 in practical applications and we set $H = 0.75$. From Fig. 3 and Fig. 4 we obtain similar performances of the shrinkage estimators.

In both of the examples, the oracle shrinkage coefficient ρ decreases in the sample number n, which makes sense since $(1 - \rho)$ can be regarded as "confidence" assigned to $\hat{S}$. Intuitively, as more and more observations are available, one has higher confidence in the sample covariance $\hat{S}$ and therefore ρ decreases. This characteristic is shown by $\hat{\rho}_{OAS}$ but not by $\hat{\rho}_{RBLW}$ and $\hat{\rho}_{LW}$. This may partly explain why the OAS estimator outperforms the RBLW and the LW estimators with small sample sizes.

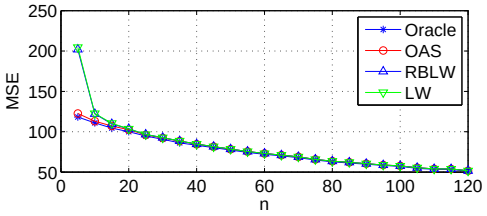

Fig. 3. Incremental FBM process: Comparison of MSE with different n when $p = 100$, Hurst parameter $H = 0.75$.

5. CONCLUSION

In this paper, we have introduced two new shrinkage estimators of covariance matrices. The RBLW estimator is proposed to improve the LW method via the Rao-Blackwell theorem. The OAS estimator is developed by iterating on the optimal oracle estimate, where the converged limit is determined analytically. The RBLW provably dominates the LW, and the OAS outperforms both the RBLW and the LW numerically. The proposed estimators have simple explicit expressions and are easy to implement. Furthermore, they share the same structure and differ from each other only in the shrinkage coefficient.

In this paper we set the shrinkage target $\hat{F}$ as the identity matrix. The theory behind the proposed estimators can be extended to other possible shrinkage targets. An interesting question for future research is how to choose appropriate targets to further reduce the estimation error.

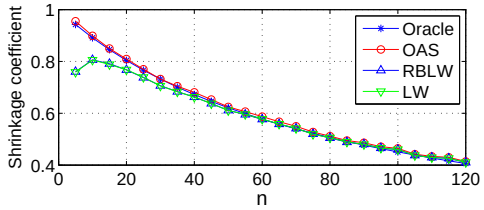

Fig. 4. Incremental FBM process: Comparison of shrinkage intensity with different n when $p = 100$, Hurst parameter $H = 0.75$.

6. REFERENCES

[1] C. Stein, Estimation of a covariance matrix. Rietz Lecture, *39th Annual Meeting IMS*, Atlanta, GA, 1975.

[2] L. R. Haff, Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix, *Annals of Statistics*, Volume 8, Number 3, Page 586-597, 1980.

[3] D. K. Dey and C. Srinivasan, Estimation of a covariance matrix under Stein's loss. *Annals of Statistics*, Volume 13, Page 1581 - 1591, 1985.

[4] R. Yang, J. O. Berger, Estimation of a covariance matrix using the reference prior, *Annals of Statistics*, Volume 22, Page 1195 - 1211, 1994.

[5] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis archive*, Volume 88, Issue 2, Pages 365 - 411, February 2004.

[6] R. Abrahamsson, Y. Selén and P. Stoica, Enhanced covariance matrix estimators in adaptive beamforming, *IEEE Proc. of ICASSP*, Pages 969 - 972, 2007

[7] J. Schäfer and K. Strimmer, A Shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology*, Volume 4, Issue 1, Article 32, 2005.

[8] Y.C., Eldar and J.S. Chernoi, A pre-test like estimator dominating the least-squares method, *J. Statist. Plann. Inference*, 2008, doi: 10.1016/j.jspi.2007.12.002.

[9] H.L. Van Trees, Detection, Estimation, and Modulation Theory, Part I. New York, NY: John Wiley & Sons, Inc., 1971.

[10] S. Johh, Some optimal multivariate tests, *Biometrika*, Volume 58, Page 123 - 127, 1971.

[11] W.E. Leland, M.S. Taqqu, W. Willinger and D.V.,Wilson, On the self-similar nature of Ethernet traffic, *IEEE Trans. on Networking*, Volume 2, Issue 1, Page 1-15, 1994.

[12] Y. Chen, A. Wiesel and A.O. Hero, Gaussian shrinkage estimation of covariance matrices under mean-squared loss, technical report, `http://sitemaker.umich.edu/yilun/files/report_cov_est.pdf`