LARGE COVARIANCE MATRIX ESTIMATION: BRIDGING SHRINKAGE AND TAPERING APPROACHES

Xiaohui Chen[1]*, Z. Jane Wang*[1]*, Martin J. McKeown*[2]

1. Department of Electrical and Computer Engineering, {xiaohuic, zjanew}@ece.ubc.ca
2. Department of Medicine (Neurology), mmckeown@interchange.ubc.ca
University of British Columbia, Canada.

ABSTRACT

In this paper, we propose a shrinkage-to-tapering oracle (STO) estimator for estimation of large covariance matrix when the number of samples is substantially fewer than the number of variables, by combining the strength from both Steinian-type shrinkage and tapering estimators. Our contributions include: (i) Deriving the Frobenius risk and a lower bound for the spectral risk of an MMSE shrinkage estimator; (ii) Deriving a closed-form expression for the optimal coefficient of the proposed STO estimator. Simulations on auto-regression (e.g. a sparse case) and fraction Brownian motion (e.g. a non-sparse case) covariance structures are used to demonstrate the superiority of the proposed estimator.

Index Terms— Covariance matrix, high-dimensionality, shrinkage estimator, tapering estimator.

1. INTRODUCTION

We consider the problem of estimating the covariance structure Σ of n i.i.d.observations $\{\mathbf{x}_i\}_{i=1}^n$ that are realized from a p-dimensional random vector. Covariance estimation problem is of great importance in array signal processing [1], eigen-image analysis [2] and principle component analysis [3]. A natural estimator of Σ is the unstructured sample covariance matrix of $\{\mathbf{x}_i\}$:

$$\hat{S} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T. \tag{1}$$

It is well-known that $\hat{S}$ is a "good" estimator and converges to Σ optimally under the spectral risk when $n \to \infty$ and the number of variables p is fixed. Unfortunately, when the model size p grows as more and more data are collected, e.g. when p is the number of basis functions of a finer expansion in a wavelet domain, random matrix theory predicts that the spectrum of $\hat{S}$ is wider than the spectrum of Σ, even when p and n grow at comparable rates [4].

This work was supported by the Pacific Alzheimer Research Foundation and the Canadian Natural Sciences and Engineering Research Council (NSERC).

To tackle the issues of estimating large covariance matrices, regularization is generally needed. The regularized estimation approaches of large covariance matrices can be broadly classified into two categories: one is based on the shrinkage methods [5, 6] that shrink the covariance matrix to some well-conditioned matrices, and the other is based on operations directly applied to $\hat{S}$ such as banding [7], tapering [8], and thresholding [9]. Both categories aim for a stabilized estimation of eigenvalues as dimensionality grows.

Despite the recent progress on large covariance matrix estimation, there has been relatively little fundamental theoretical study on comparing the shrinkage-type and tapering-type estimators. To fill this gap, we propose in this paper a *shrinkage-to-tapering* estimator for *general* and *high-dimensional* covariance matrices, with the goal that it improves upon both shrinkage and tapering estimators.

The rest of paper is organized as follows. In Section 2, we first discuss the MMSE shrinkage and tapering estimators. We then derive the Frobenius risk and a lower on the spectral risk for the MMSE shrinkage estimator and we show inconsistency of the shrinkage estimator. In Section 3, we propose an alternative solution termed as shrinkage-to-tapering oracle (STO) estimator. Simulations are conducted in Section 4 to demonstrate the improved numeric performances of the proposed estimator. Due to space limit, detail proofs in this paper are omitted, and interested readers are referred to the journal version [10] for details.

2. RISK BOUNDS OF SHRINKAGE AND TAPERING ESTIMATORS

The main purpose of this section is to study and compare the risk bounds of two important regularized covariance matrix estimators. This serves as a motivation of the proposed shrinkage-to-tapering estimator.

2.1. Shrinkage Estimator

Chen et.al [6] define a minimum mean-squared error (MMSE) oracle estimator as the solution of the optimization problem

$$\begin{aligned} \text{minimize}_{\rho\in[0,1]} \quad & E\left\|\hat{\Sigma}(\rho)-\Sigma\right\|_F^2 \\ \text{subject to} \quad & \hat{\Sigma}=(1-\rho)\hat{S}+\rho\hat{F}. \end{aligned} \tag{2}$$

In words, the MMSE oracle estimator seeks the best convex combination between the sample covariance matrix and a scaled identity matrix to approximate the true covariance matrix in terms of MSEs. This estimator is said to be an oracle because the optimal solution depends on Σ which is unknown in practice and is the estimation goal. Under additional Gaussian assumption, the closed-form of ρ_o is given in [6]:

$$\rho_o=\frac{p-2+pt}{p(n+1)-2+(p-n)t},$$

where

$$t=\text{Tr}^2(\Sigma)/\text{Tr}(\Sigma^2).$$

Here t measures the distribution of the off-diagonal entries of Σ. In particular,

$$\text{Tr}(\Sigma^2)\le\text{Tr}^2(\Sigma)\le p\text{Tr}(\Sigma^2),$$

where equalities of the left and right inequalities are attained if and only if $\Sigma=\mathbf{1}\mathbf{1}^T$ and $\Sigma=I$, respectively. So when $t=1$, the matrix entries have the most spread support (dense); while when $t=p$, the energy of Σ concentrates on the diagonal (sparse).

We first give the Frobenius risk of the MMSE oracle estimator (2), assuming that the data are from i.i.d. $N(\mathbf{0},\Sigma)$.

Theorem 2.1. *Suppose $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. Gaussian $N(\mathbf{0},\Sigma)$. The Frobenius risk of the MMSE shrinkage oracle estimator (2) is given by*

$$E\|\hat{\Sigma}_o-\Sigma\|_F^2=\left[(1-\frac{t}{p})\rho_o+\frac{2}{np}\right]\|\Sigma\|_F^2. \tag{3}$$

Next, we also derive a lower bound on the risk under the spectral norm.

Theorem 2.2. *Suppose $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d. Gaussian $N(\mathbf{0},\Sigma)$. The spectral risk of the MMSE shrinkage oracle estimator (2) satisfies*

$$E\|\hat{\Sigma}_o-\Sigma\|^2\ge\rho_o^2(1-\lambda_{\min}(\Sigma))^2. \tag{4}$$

2.2. Tapering Estimation

Tapering estimator is defined through the *covariance matrix taper* (CMT). More specifically, we let $\mathcal{S}$ be the set of $p\times p$ symmetric matrix and $A\circ B$ be the Schur product of two matrices A and B: $A\circ B=(a_{ij}b_{ij})$. Then, we define

Definition 2.1. A *covariance matrix taper* (CMT) A is an element in $\mathcal{S}$ such that $\sum_{j=1}^p\lambda_j(A\circ B)\le\sum_{j=1}^p\lambda_j(B)$ for all $B\in\mathcal{S}$. In other words, multiplication by any CMT decreases the averaged eigenvalue.

Let W be a CMT; a *tapering estimator* of the covariance matrix is defined as

$$\hat{\Sigma}_{\text{taper}}=W\circ\hat{S}. \tag{5}$$

Coupled with tapering estimator, we consider the following class of covariance matrices

$$\begin{aligned}\mathcal{G}(\alpha,C,C_0)=&\{\Sigma:\max_j\sum_{|i-j|>k}|\sigma_{ij}|\le Ck^{-\alpha},\forall k\\ &\text{and }\lambda_{\max}(\Sigma)\le C_0\},\end{aligned}$$

where $C,C_0>0$ are some absolute constants and $\alpha>0$ is a smoothing parameter specifying the rate of decay of σ_{ij} from the main diagonal. The following theorem, proved in [8], shows that a covariance tapering estimator based on data generated from i.i.d $N(\mathbf{0},\Sigma)$ with $\Sigma\in\mathcal{G}(\alpha,C,C_0)$ is minimax.

Theorem 2.3. *(Cai, Zhang, and Zhou [8]) Suppose $\log p=o(n)$ and $p\ge n^\xi$ for some $\xi>0$; then we have the following minimax convergence rate*

1. *under the Frobenius risk/normalized MSE:*

$$p^{-1}\inf_{\hat{\Sigma}}\sup_{\Sigma\in\mathcal{G}(\alpha,C,C_0)}E\|\hat{\Sigma}-\Sigma\|_F^2\asymp n^{-\frac{2\alpha+1}{2(\alpha+1)}};$$

2. *under the spectral risk:*

$$\inf_{\hat{\Sigma}}\sup_{\Sigma\in\mathcal{G}(\alpha,C,C_0)}E\|\hat{\Sigma}-\Sigma\|^2\asymp n^{-\frac{2\alpha}{2\alpha+1}}+\log(p)/n.$$

2.3. Risk Bound Comparison Between Tapering and Shrinakge Estimators

We consider two types of the covariance matrix structures: one is in $\mathcal{G}(\alpha,C,C_0)$ and the other is not.

Example 2.1. Consider, for $0<\gamma<1$,

$$\sigma_{ij}=\begin{cases}1, & \text{for } i=j,\\ \gamma^{|i-j|}, & \text{for } i\ne j.\end{cases}$$

The entries of Σ decay exponentially fast when moving away from the main diagonal and thus $\Sigma\in\mathcal{G}(\alpha,C,C_0)$ for every $\alpha>0$. This example corresponds to the covariance structure of auto-regression models with order 1, AR(1). For this Σ, it can be shown that

$$p^{-1}E\|\hat{\Sigma}_o-\Sigma\|_F^2=C(\gamma)+o(1),$$

where $C(\gamma)>0$ is a constant, independent of n. It is clear that the normalized MSE is lower bounded by a positive constant depending on γ and therefore the MMSE shrinkage oracle estimator cannot be a consistency estimator of Σ unless the concentration $p/n\to0$. Fig. 1 plots the finite sample size behavior of the normalized MSE and its limit.

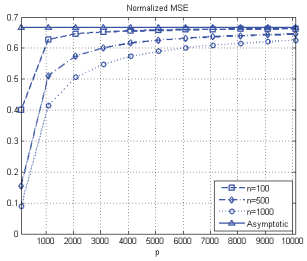


Fig. 1. Normalized MSE curves of the shrinkage MMSE estimator for the large covariances discussed in Example 2.1. The MMSE estimator fails to be consistent when $n/p \to 0$, because the normalized Frobenius risks converge to the asymptotic values that are bounded away from 0.

Example 2.2. In a second example, we consider the covariance structure of a fractional Brownian motion (FBM) with the Hurst parameter $h \in [0.5, 1]$:

$$\sigma_{ij} = 2^{-1}[(|i-j|+1)^{2h} - 2|i-j|^{2h} + (|i-j|-1)^{2h}].$$

The FBM is a model for complex systems that have long-range dependence for h being close to 1, such as modeling the internet traffic [11]. A direct calculation shows that $\|\Sigma\|_1 = p^{2h}$ and from this it easily follows that

$$\max_j \sum_{|i-j|\geq 1} |\sigma_{ij}| \geq p^{2h-1} - 1.$$

Therefore, we see that, when $h > 0.5$, $\Sigma \notin \mathcal{G}(\alpha, C, C_0)$ for any $\alpha > 0$ and the minimax properties stated Theorem 2.3 does not necessarily hold.

3. SHRINKAGE-TO-TAPERING ESTIMATOR

Now, we propose a Steinian shrinkage type estimator. With the important difference from the shrinkage estimator toward a scaled identity matrix, the proposed estimator shrinks the sample covariance matrix to its tapered version.

$$\hat{\Sigma}(\rho) = (1-\rho)\hat{S} + \rho(W \circ \hat{S}),$$

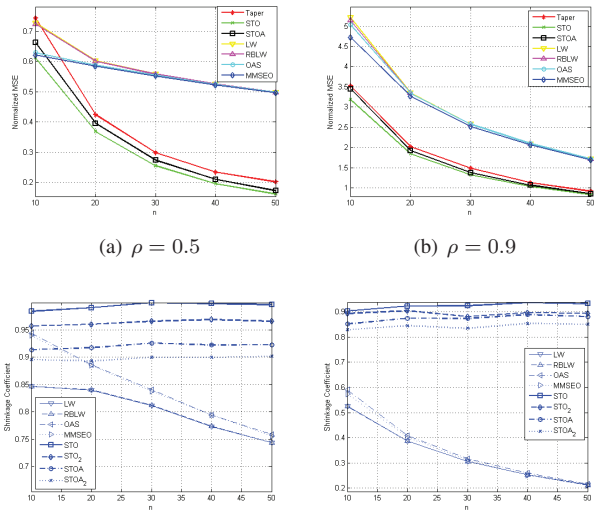
where ρ is determined by the solution to the optimization problem

$$\begin{aligned} \text{minimize}_{\rho\in[0,1]} \quad & E\|\hat{\Sigma}(\rho) - \Sigma\|_F^2 \\ \text{subject to} \quad & \hat{\Sigma}(\rho) = (1-\rho)\hat{S} + \rho(W \circ \hat{S}). \end{aligned}$$

For $\Sigma \in \mathcal{G}(\alpha, C, C_0)$, we can see from Theorem 2.3 that the proposed shrinkage-to-tapering oracle (STO) estimator reduces to the tapering estimator for large n and p. While for $\Sigma \notin \mathcal{G}(\alpha, C, C_0)$, the proposed estimator reduces to an analogy of the MMSE shrinkage oracle estimator. Therefore, we expect that, for an arbitrary large covariance matrix Σ, the proposed estimator could improve upon both tapering and MMSE shrinkage oracle estimators.

The optimal coefficient of the MMSE STO estimator can be given in a closed-form.

Fig. 2. Model 1: The normalized MSE curves as a function of n, averaged over 100 replications. The tapering [8], LW [5], RBLW [6], MMSE shrinkage oracle (MMSEO) [6], and OAS [6] are compared with the proposed STO and STOA estimators.



(a) $\rho = 0.5$ (b) $\rho = 0.9$

(c) $\rho = 0.5$ (d) $\rho = 0.9$

Theorem 3.1. *The coefficient of the proposed STO estimator under the minimum Frobenius risk is*

$$\hat{\rho}^{STO} = \frac{E(\|\hat{S}\|_F^2 - \|V \circ \hat{S}\|_F^2) - (\|\Sigma\|_F^2 - \|V \circ \Sigma\|_F^2)}{E\|\hat{S}\|_F^2 + E\|W \circ \hat{S}\|_F^2 - 2E\|V \circ \hat{S}\|_F^2}. \tag{6}$$

Under further Gaussian assumption, we can write (6) in a closed-form given by

$$\begin{aligned} \hat{\rho}^{STO} =& [\|\Sigma\|_F^2 + Tr^2(\Sigma) - \|V \circ \Sigma\|_F^2 - Tr(DV^2D)] \\ & \Big/ [(n+1)(\|\Sigma\|_F^2 + \|W \circ \Sigma\|_F^2 - 2\|V \circ \Sigma\|_F^2) \\ & + Tr^2(\Sigma) + Tr(DW^2D) - 2Tr(DV^2D)]. \end{aligned} \tag{7}$$

Since the STO estimator depends on the unknown Σ, we also present an iterative STO approximation (STOA) algorithm. For details, please see [10].

4. SIMULATIONS

Simulations based on the two examples discussed in Section 2 are performed to study the finite sample size numeric performances of the proposed estimators. We fix $p = 100$ for all models and consider different values of n with $n = \{10, 20, 30, 40, 50\}$. The STOA algorithm is initialized at $\hat{\Sigma}_0 = \hat{S}$ and $\rho = 0.5$. The maximum number of iterations in the STOA algorithm is set to be 10. We compare the proposed STO estimator and its variant STOA with the tapering [8] and several shrinkage estimators including

the LW [5], Rao-Blackwellized LW (RBLW) [6], MMSE shrinkage oracle (MMSEO) estimator and its variant oracle approximating shrinkage (OAS) [6] estimator.

4.1. Model 1 - AR(1) model

We chose $\rho = \{0.5, 0.9\}$. Due to space limit, we only plot the estimated normalized MSEs, i.e. the Frobenius risk in Fig. 2 for different estimators. For the spectral risk, we refer interested readers to the journal version [10].

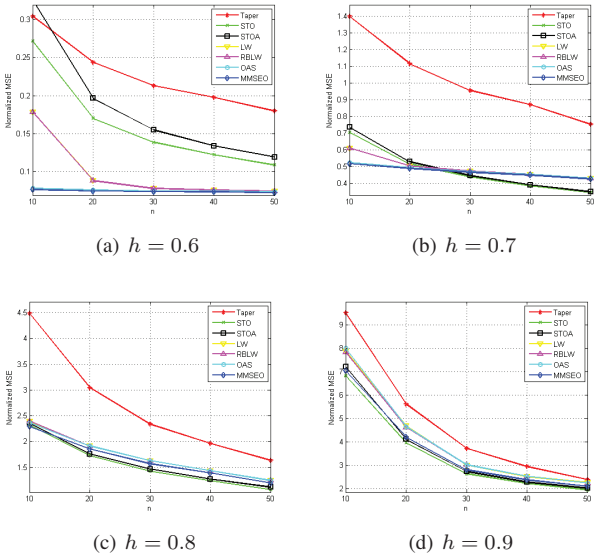
Several interesting observations can be made from Fig. 2. First, in terms of estimation risks, the STO, STOA, and tapering estimators uniformly improve upon the previous shrinkage-type estimators including LW, RBLW, OAS, and the MMSEO. This validates our Theorem 2.1 on finite sample size data. The improvement is visually appreciable even when n is not so large as considered in the asymptotic setup. Second, the proposed STO and STOA also outperform the tapering estimator, although the improvement is smaller than those from the previous shrinkage-type estimators. Third, it is clear from these two figures that STOA can well approximate the STO estimator.

4.2. Model 2 - fractional Brownian motion

We simulate observations from the FBM covariance structure with the Hurst parameter h selected from $h = \{0.6, 0.7, 0.8, 0.9\}$.

From Fig. 3, we can see that the normalized MSEs of the MMSE shrinkage estimators are smaller than that of the tapering estimator. This is not surprising because: (i) the assumption $\Sigma \in \mathcal{G}(\alpha, C, C_0)$ is violated and therefore no optimality under the Frobenius risk can be expected in the tapering estimator; (ii) the MMSE estimators are designed to minimize the Frobenius risk. It is also observed that the STO and STOA estimators uniformly outperform other shrinkage estimators when $h = 0.8$ and $h = 0.9$. In the case of $h = 0.6$, they are outperformed by LW, RBLW, OAS, and MMSEO estimators but still yield smaller MSEs than the tapering estimator. The case of $h = 0.7$ appears to be non-uniform; however the curve trends shown in Fig 3(b) suggest that STO and STOA may eventually yield a smaller Frobenius risk as n gets larger.

Fig. 3. Model 2 (FBM): The normalized MSE curves as a function of n, averaged over 100 replications.



(a) $h = 0.6$ (b) $h = 0.7$

(c) $h = 0.8$ (d) $h = 0.9$

5. CONCLUSION

In this paper, we first show that the MMSE shrinkage oracle estimator is inconsistent under both Frobenius and spectral risks for some typical covariance matrices in $\mathcal{G}(\alpha, C, C_0)$. We then propose a STO estimator that combines the advantages from both the MMSE shrinkage and tapering estimators.

6. REFERENCES

[1] Richard Abrahamsson, Yngve Selen, and Petre Stoica, "Enhanced Covariance Matrix Estimators in Adaptive Beamforming," *2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 969–972, 2007.

[2] Guangzhi Cao, Leonardo R. Bachega, and Charles A. Bouman, "The Sparse Matrix Transform for Covariance Estimation and Analysis of High Dimensional Signals," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 625–640, 2011.

[3] Iain M. Johnstone and Arthur Yu Lu, "Sparse Principal Components Analysis," *ArXiv:0901.4392v1*, 2004.

[4] Debashis Paul, "Asymptotics of Sample Eigenstructures for a Large Dimensional Spiked Covariance Model," *Statistica Sinica*, vol. 17, pp. 1617–1642, 2007.

[5] Olivier Ledoit and Michael Wolf, "A Well-Conditioned Estimator for Large Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[6] Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero, "Shrinkage Algorithms for MMSE Covariance Estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5016–5029, 2010.

[7] Peter J. Bickel and Elizaveta Levina, "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[8] Tony Cai, Cun-Hui Zhang, and Harrison Zhou, "Optimal Rates of Convergence for Covariance Matrix Estimation," *The Annals of Statistics*, vol. 38, no. 4, pp. 2118–2144, 2010.

[9] Peter J. Bickel and Elizaveta Levina, "Covariance Regularization by Thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[10] Xiaohui Chen, Z. Jane Wang, and Martin J. McKeown, "Shrinkage-to-Tapering Estimation of Large Covariance Matrices," *Submitted to IEEE Transactions on Signal Processing*, 2011.

[11] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.