Signal Processing 148 (2018) 223–233

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Linear shrinkage estimation of covariance matrices using low-complexity cross-validation

Jun Tong[a,*], Rui Hu[a], Jiangtao Xi[a], Zhitao Xiao[b], Qinghua Guo[a], Yanguang Yu[a]

[a] School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia
[b] School of Electronic and Information Engineering, Tianjin Polytechnic University, China

ARTICLE INFO

Article history:
Received 25 August 2017
Revised 20 February 2018
Accepted 21 February 2018

Keywords:
Covariance matrix
Cross-validation
Linear shrinkage
Ordinary least squares
Sample covariance matrix

ABSTRACT

Shrinkage can effectively improve the condition number and accuracy of covariance matrix estimation, especially for low-sample-support applications with the number of training samples smaller than the dimensionality. This paper investigates parameter choice for linear shrinkage estimators. We propose data-driven, leave-one-out cross-validation (LOOCV) methods for automatically choosing the shrinkage coefficients, aiming to minimize the Frobenius norm of the estimation error. A quadratic loss is used as the prediction error for LOOCV. The resulting solutions can be found analytically or by solving optimization problems of small sizes and thus have low complexities. Our proposed methods are compared with various existing techniques. We show that the LOOCV method achieves near-oracle performance for shrinkage designs using sample covariance matrix (SCM) and several typical shrinkage targets. Furthermore, the LOOCV method provides low-complexity solutions for estimators that use general shrinkage targets, multiple targets, and/or ordinary least squares (OLS)-based covariance matrix estimation. We also show applications of our proposed techniques to several different problems in array signal processing.

1. Introduction

In statistical signal processing, one critical problem is to estimate the covariance matrix, which has extensive applications in correlation analysis, portfolio optimization, and various signal processing tasks in radar and communication systems [1–5]. One key challenge is that when the dimensionality is large but the sample support is relatively low, the estimated covariance matrix $\mathbf{R}$, which may be obtained using a general method such as sample covariance matrix (SCM) or ordinary least squares (OLS), becomes ill-conditioned or even singular, and suffers from significant errors relative to the true covariance matrix $\mathbf{\Sigma}$. Consequently, signal processing tasks that rely on covariance matrix estimation may perform poorly or fail to apply. Regularization techniques have attracted tremendous attention recently for covariance matrix estimation. By imposing structural assumptions of the true covariance matrix $\mathbf{\Sigma}$, techniques such as banding [6], thresholding [7], and shrinkage [8–18] have demonstrated great potential for improving the performance of covariance matrix estimation. See [19–21] for recent surveys.

This paper is concerned with the linear shrinkage estimation of covariance matrices. Given an estimate $\mathbf{R}$ of the covariance matrix, a linear shrinkage estimate is constructed as

$$\widehat{\mathbf{\Sigma}}_{\rho,\tau} = \rho\mathbf{R} + \tau\mathbf{T}_0, \quad (1)$$

where $\mathbf{T}_0$ is the shrinkage target and ρ and τ are nonnegative shrinkage coefficients. In general, the shrinkage target $\mathbf{T}_0$ is better-conditioned, more parsimonious or more structured, with lower variance but higher bias compared to the original estimate $\mathbf{R}$ [11]. The coefficients ρ and τ are chosen to provide a good tradeoff between bias and variance, such that an estimate outperforming both $\mathbf{R}$ and $\mathbf{T}_0$ is achieved and a better approximation to the true covariance matrix $\mathbf{\Sigma}$ can be obtained. Compared to other regularized estimators such as banding and thresholding, linear shrinkage estimators can be easily designed to guarantee positive-definiteness. Such shrinkage designs have been employed in various applications which utilize covariance matrices and have demonstrated significant performance improvements. The linear shrinkage approach has also been generalized to nonlinear shrinkage estimation of covariance matrices [22,23], and is closely related to several unitarily invariant covariance matrix estimators that shrink the eigenvalues of the SCM, such as those imposing condition number constraints on the estimate [24,25]. There are also a body of studies on shrinkage estimation of precision matrix (the inverse of covariance matrix) [26–30] and on application-

* Corresponding author.
E-mail address: jtong@uow.edu.au (J. Tong).

https://doi.org/10.1016/j.sigpro.2018.02.026
0165-1684/

oriented design of shrinkage estimators. See [31–36] for example applications in array signal processing.

Shrinkage has a Bayes interpretation [2,9]. The true covariance matrix $\mathbf{\Sigma}$ can be assumed to be within the neighborhoods of the shrinkage target $\mathbf{T}_0$. There can be various different approaches for constructing $\mathbf{R}$ and $\mathbf{T}_0$. For example, when a generative model about the observation exists, one may first estimate the model parameters and then construct $\mathbf{R}$ [20]. A typical example of this is linear models seen in communication systems. Furthermore, different types of shrinkage targets, not necessarily limited to identity or diagonal targets, can be used to better utilize prior knowledge. For example, knowledge-aided space-time signal processing (KA-STAP) may set $\mathbf{T}_0$ using knowledge about the environment [3] or past covariance matrix estimates [37]. Even multiple shrinkage targets can be applied when distinct guesses about the true covariance matrix are available [17].

The choice of shrinkage coefficients significantly influences the performance of linear shrinkage estimators. Various criteria and methods have been studied. Under the mean squared error (MSE) criterion, Ledoit and Wolf (LW) [2] derived closed-form solutions based on asymptotic estimates of the statistics needed for finding the optimal shrinkage coefficients, where $\mathbf{R}$ and $\mathbf{T}_0$ are assumed as the SCM and identity matrix, respectively. Later the LW solution was extended for more general shrinkage targets [3,17]. Chen et al. [4] assumed Gaussian distribution and proposed an oracle approximating shrinkage (OAS) estimator, which achieves near-optimal parameter choice for Gaussian data even with very low sample supports. The shrinkage coefficients determination can also be cast as a model selection problem and thus generic model selection techniques such as cross-validation (CV) [38–40] can be applied. In general, CV splits the training samples for multiple times into disjoint subsets and then fits and assesses the models under different splits based on a properly chosen prediction loss. This has been explored, e.g., in [10,13], where the Gaussian likelihood is used as the prediction loss.

All these data-driven techniques achieve near-optimal parameter choice when the underlying assumptions hold. However, there are also limitations to their applications: almost all existing analytical solutions to shrinkage coefficients [2–4], Lancewicki and Aladjem [17] were derived under the assumption of SCM and certain special forms of shrinkage targets. They need to be re-designed when applied to other cases, which is generally nontrivial. The asymptotic analysis-based methods [2,3] may not perform well when the sample support is very low. Although the existing CV approaches [10,13] have broader applications, they assume Gaussian distribution and employ grid search to determine the shrinkage coefficients. The likelihood cost of [10,13] must be computed for multiple data splits and multiple candidates of shrinkage coefficients, which can be time-consuming.

In this paper, we further investigate data-driven techniques that automatically tune the linear shrinkage coefficients using leave-one-out cross-validation (LOOCV). We choose a simple quadratic loss as the prediction loss for LOOCV, and derive analytical and computationally efficient solutions. The solutions do not need to specify the distribution of the data. Furthermore, the LOOCV treatment is applicable to different covariance matrix estimators including the SCM- and ordinary least squares (OLS)-based schemes. It can be used together with general shrinkage targets and can also be easily extended to incorporate multiple shrinkage targets. The numerical examples show that the proposed method can achieve oracle-approximating performance for covariance matrix estimation and can improve the performance of several array signal processing schemes.

The remainder of the paper is organized as follows. In Section 2, we present computationally efficient LOOCV methods for choosing the linear shrinkage coefficients for both SCM- and OLS-based covariance matrix estimators and also compare the proposed LOOCV methods with several existing methods which have attracted considerable attentions recently. In Section 3, we extend our results for multi-target shrinkage. Section 4 reports numerical examples, and finally Section 5 gives conclusions.

2. LOOCV choice of linear shrinkage coefficients

This paper deals with the estimation of covariance matrices of zero-mean signals whose fourth-order moments exist. We study the LOOCV choice of the shrinkage coefficients for the linear shrinkage covariance matrix estimator (1), i.e., $\widehat{\mathbf{\Sigma}}_{\rho,\tau} = \rho\mathbf{R} + \tau\mathbf{T}_0$. The following assumptions are made:

1. The true covariance matrix $\mathbf{\Sigma}$, the estimated covariance matrix $\mathbf{R}$, and the shrinkage target $\mathbf{T}_0$ are all Hermitian and positive-semidefinite (PSD).
2. T independent, identically distributed (i.i.d.) samples $\{\mathbf{y}_t\}$ of the signal are available.
3. The shrinkage coefficients are nonnegative, i.e.,

$$\rho \geq 0, \quad \tau \geq 0. \tag{2}$$

Assumption 3 follows the treatments in [2–4] and is sufficient but not necessary to guarantee that the shrinkage estimate $\widehat{\mathbf{\Sigma}}_{\rho,\tau}$ is PSD when Assumption 1 holds.[1] Two classes of shrinkage targets will be considered in this paper. One is constructed independent of the training samples $\{\mathbf{y}_t\}$ for generating $\mathbf{R}$, similarly to the knowledge-aided targets as considered in [3]. The other is constructed from $\{\mathbf{y}_t\}$, but is highly structured with significantly fewer free parameters as compared to $\mathbf{R}$. Examples of the second class include those constructed using only the diagonal entries of $\mathbf{R}$ [4,20] and the Toeplitz approximations of $\mathbf{R}$ [17].

2.1. Oracle choice

Different criteria may be used for evaluating the covariance matrix estimators. In this paper, we use the squared Frobenius norm of the estimation error as the performance measure. Given $\mathbf{\Sigma}$, $\mathbf{R}$ and $\mathbf{T}_0$, the oracle shrinkage coefficients minimize

$$J_O(\rho,\tau) = ||\widehat{\mathbf{\Sigma}}_{\rho,\tau} - \mathbf{\Sigma}||_F^2 = ||\rho\mathbf{R} + \tau\mathbf{T}_0 - \mathbf{\Sigma}||_F^2, \tag{3}$$

where $||\cdot||_F$ denotes the Frobenius norm. The cost function in (3) can then be rewritten as a quadratic function of the shrinkage coefficients:

$$J_O(\rho,\tau) = \begin{bmatrix}\rho\\ \tau\end{bmatrix}^T \mathbf{A}_O \begin{bmatrix}\rho\\ \tau\end{bmatrix} - 2\begin{bmatrix}\rho\\ \tau\end{bmatrix}^T \mathbf{b}_O + \mathrm{tr}(\mathbf{\Sigma}^2), \tag{4}$$

$$\mathbf{A}_O = \begin{bmatrix}\mathrm{tr}(\mathbf{R}^2) & \mathrm{tr}(\mathbf{R}\mathbf{T}_0)\\ \mathrm{tr}(\mathbf{R}\mathbf{T}_0) & \mathrm{tr}(\mathbf{T}_0^2)\end{bmatrix}, \tag{5}$$

$$\mathbf{b}_O = \begin{bmatrix}\mathrm{tr}(\mathbf{R}\mathbf{\Sigma})\\ \mathrm{tr}(\mathbf{T}_0\mathbf{\Sigma})\end{bmatrix}, \tag{6}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. As $\mathbf{A}_O$ is positive-definite, we can find the minimizer of $J_O(\rho,\tau)$ by solving the above bivariate convex optimization problem. We can also apply the Karush–Kuhn–Tucker (KKT) conditions to find the solution analytically. From (4), letting $\frac{J_O(\rho,\tau)}{\partial\rho} = \frac{J_O(\rho,\tau)}{\partial\tau} = 0$ leads to

$$\frac{\mathrm{tr}(\mathbf{R}^2)}{\mathrm{tr}(\mathbf{R}\mathbf{\Sigma})}\rho + \frac{\mathrm{tr}(\mathbf{R}\mathbf{T}_0)}{\mathrm{tr}(\mathbf{R}\mathbf{\Sigma})}\tau = 1, \tag{7}$$

[1] Imposing Assumption 3 may introduce performance loss. Alternatively, one may remove the constraint $\rho \geq 0$, $\tau \geq 0$ and impose a constraint that $\widehat{\mathbf{\Sigma}}_{\rho,\tau}$ is PSD, similar to a treatment in [5].