

Linear Shrinkage Estimation of Covariance Matrices Using Low-Complexity Cross-Validation

Jun Tong, Rui Hu, Jiangtao Xi, Zhitao Xiao, Qinghua Guo and Yanguang Yu

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.

Index Terms—Covariance matrix, cross-validation, linear shrinkage, ordinary least squares, sample covariance matrix.

I. 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 Σ . 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 Σ , 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.

J. Tong, R. Hu, J. Xi, Q. Guo and Y. Yu are with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia. Email: rh546@uowmail.edu.au, {jtong, jiangtao, qguo, yanguang}@uow.edu.au.

Z. Xiao is with School of Electronic and Information Engineering, Tianjin Polytechnic University, China.

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

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

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 Σ 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-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 Σ 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], [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.

II. 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., $\hat{\Sigma}_{\rho,\tau} = \rho\mathbf{R} + \tau\mathbf{T}_0$. The following assumptions are made:

- 1) The true covariance matrix Σ , 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.,

$$\rho \geq 0, \quad \tau \geq 0. \quad (2)$$

Assumption 3 follows the treatments in [2]-[4] and is sufficient but not necessary to guarantee that the shrinkage estimate $\hat{\Sigma}_{\rho,\tau}$ is PSD when Assumption 1 holds¹. 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].

A. 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 Σ , \mathbf{R} and \mathbf{T}_0 , the oracle shrinkage coefficients minimize

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

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

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

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

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

where $\text{tr}(\cdot)$ 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

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

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

$$\frac{\text{tr}(\mathbf{R}\mathbf{T}_0)}{\text{tr}(\mathbf{T}_0\mathbf{\Sigma})}\rho + \frac{\text{tr}(\mathbf{T}_0^2)}{\text{tr}(\mathbf{T}_0\mathbf{\Sigma})}\tau = 1. \quad (8)$$

The oracle shrinkage coefficients can be obtained by solving (7) and (8):

$$\begin{bmatrix} \rho_O^* \\ \tau_O^* \end{bmatrix} = \mathbf{A}_O^{-1} \mathbf{b}_O. \quad (9)$$

Note that (9) may produce negative shrinkage coefficients, which may not lead to a positive-definite estimate of the covariance matrix. In this case, we clip the negative coefficient to zero and then find the other coefficient using (7) or (8) to guarantee the positive definiteness, for $\tau = 0$ or $\rho = 0$, respectively. This treatment is similar to [2]-[5] and provides a suboptimal yet simple solution. The oracle estimator requires knowledge of $\mathbf{\Sigma}$, which is unavailable in real applications, but the result serves as an upper bound of the performance given the linear shrinkage structure.

B. LOOCV Choice for General Cases

Let $\widehat{\mathbf{\Sigma}}$ denote a positive-definite, Hermitian matrix. It can be easily verified that the following cost

$$J_S(\widehat{\mathbf{\Sigma}}) \triangleq \mathbb{E}[\|\widehat{\mathbf{\Sigma}} - \mathbf{y}\mathbf{y}^\dagger\|_F^2] \quad (10)$$

is minimized when $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma}$, where the expectation is taken over \mathbf{y} . In this paper, we apply LOOCV [38] to produce an estimate of $J_S(\widehat{\mathbf{\Sigma}})$ as the proxy for measuring the accuracy of $\widehat{\mathbf{\Sigma}}$, based on which the shrinkage coefficients can be selected. With the LOOCV method, the length- T training data $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T]$ is repeatedly split into two sets with respect to time. For the t -th split, where $1 \leq t \leq T$, $T-1$ samples in \mathbf{Y}_t (with the t -th column \mathbf{y}_t omitted from \mathbf{Y}) are used for producing a covariance matrix estimate \mathbf{R}_t and the remaining sample \mathbf{y}_t is spared for parameter validation. In total, T splits of the training data \mathbf{Y} are used and all the training samples are used for validation once. Assuming shrinkage estimation with given shrinkage coefficients (ρ, τ) , we construct from each \mathbf{Y}_t a shrinkage covariance matrix estimator as

$$\widehat{\mathbf{\Sigma}}_{t,\rho,\tau} = \rho \mathbf{R}_t + \tau \mathbf{T}_0. \quad (11)$$

We propose to use the following LOOCV cost function

$$J_{CV}(\rho, \tau) = \frac{1}{T} \sum_{t=1}^T \|\widehat{\mathbf{\Sigma}}_{t,\rho,\tau} - \mathbf{y}_t \mathbf{y}_t^\dagger\|_F^2 \quad (12)$$

$$= \frac{1}{T} \sum_{t=1}^T \|\rho \mathbf{R}_t + \tau \mathbf{T}_0 - \mathbf{y}_t \mathbf{y}_t^\dagger\|_F^2 \quad (13)$$

to approximate the cost in (10) when $\widehat{\mathbf{\Sigma}}$ is chosen as $\widehat{\mathbf{\Sigma}}_{t,\rho,\tau}$. For notational simplicity, define

$$\mathbf{S}_t \triangleq \mathbf{y}_t \mathbf{y}_t^\dagger. \quad (14)$$

After some manipulations, the above cost function can be written similarly to (4) as

$$J_{CV}(\rho, \tau) = \begin{bmatrix} \rho \\ \tau \end{bmatrix}^T \mathbf{A}_{CV} \begin{bmatrix} \rho \\ \tau \end{bmatrix} - 2 \begin{bmatrix} \rho \\ \tau \end{bmatrix}^T \mathbf{b}_{CV} + \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{S}_t^2), \quad (15)$$

where

$$\mathbf{A}_{CV} = \begin{bmatrix} \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t^2) & \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{T}_0) \\ \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{T}_0) & \text{tr}(\mathbf{T}_0^2) \end{bmatrix}, \quad (16)$$

$$\mathbf{b}_{CV} = \begin{bmatrix} \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{S}_t) \\ \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{T}_0 \mathbf{S}_t) \end{bmatrix}. \quad (17)$$

The shrinkage coefficients can then be found by solving the above bivariate, constant-coefficient quadratic program. Analytical solutions can be obtained under different conditions, as shown below.

1) *Unconstrained shrinkage:* For unconstrained (ρ, τ) , setting the partial derivatives $\frac{\partial J_{CV}(\rho, \tau)}{\partial \rho} = \frac{\partial J_{CV}(\rho, \tau)}{\partial \tau} = 0$ yields

$$\frac{\sum_{t=1}^T \text{tr}(\mathbf{R}_t^2)}{\sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{S}_t)} \rho + \frac{\sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{T}_0)}{\sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{S}_t)} \tau = 1, \quad (18)$$

$$\frac{\sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{T}_0)}{\sum_{t=1}^T \text{tr}(\mathbf{T}_0 \mathbf{S}_t)} \rho + \frac{T \text{tr}(\mathbf{T}_0^2)}{\sum_{t=1}^T \text{tr}(\mathbf{T}_0 \mathbf{S}_t)} \tau = 1. \quad (19)$$

Solving (18) and (19) produces the unconstrained solution

$$\begin{bmatrix} \rho_{CV}^* \\ \tau_{CV}^* \end{bmatrix} = \mathbf{A}_{CV}^{-1} \mathbf{b}_{CV}. \quad (20)$$

We choose (20) as the optimal shrinkage coefficients if both ρ_{CV}^* and τ_{CV}^* are nonnegative. Otherwise, we consider the optimal choices on the boundary of $\rho \geq 0, \tau \geq 0$ specified by (18) or (19) for $\tau = 0$ or $\rho = 0$ as

$$\rho_{CV}^* = \frac{\sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{S}_t)}{\sum_{t=1}^T \text{tr}(\mathbf{R}_t^2)}, \quad \tau_{CV}^* = 0, \quad (21)$$

or

$$\rho_{CV}^* = 0, \quad \tau_{CV}^* = \frac{\sum_{t=1}^T \text{tr}(\mathbf{T}_0 \mathbf{S}_t)}{T \text{tr}(\mathbf{T}_0^2)}. \quad (22)$$

2) *Constrained shrinkage:* For the more parsimonious design using convex linear combination, the following constraint is imposed:

$$\rho = 1 - \tau. \quad (23)$$

By plugging (23) into the cost function (12) and taking the minimizer, we can also easily find the optimal shrinkage coefficients using

$$\rho_{CV}^* = \frac{\sum_{t=1}^T (\text{tr}(\mathbf{T}_0^2) - \text{tr}(\mathbf{R}_t \mathbf{T}_0) - \text{tr}(\mathbf{T}_0 \mathbf{S}_t) + \text{tr}(\mathbf{S}_t \mathbf{R}_t))}{\sum_{t=1}^T (\text{tr}(\mathbf{R}_t^2) - 2\text{tr}(\mathbf{R}_t \mathbf{T}_0) + \text{tr}(\mathbf{T}_0^2))}. \quad (24)$$

In case a negative shrinkage coefficient is produced, we set it to zero and let the other be one according to (23). Note that although the closed-form solution involves multiple matrix operations, the quantities involved need to be computed only once. Furthermore, the computational complexity may be greatly reduced given a specific method of covariance matrix estimation. In the following two subsections, we will show the simplified solutions for SCM- and OLS-based covariance matrix estimation.

C. LOOCV Choice for SCM-Based Estimation

We consider in this subsection that \mathbf{R} is the SCM estimate of Σ . In this case,

$$\mathbf{R} = \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t \mathbf{y}_t^\dagger = \frac{1}{T} \sum_{t=1}^T \mathbf{R}_t = \frac{1}{T} \sum_{t=1}^T \mathbf{S}_t, \quad (25)$$

which is a sufficient statistic for Gaussian-distributed data when the mean vector is the zero vector. For the t -th split, the SCM constructed from all the samples except the t -th is

$$\mathbf{R}_t = \frac{1}{T-1} \sum_{j \neq t} \mathbf{y}_j \mathbf{y}_j^\dagger = \frac{T}{T-1} \mathbf{R} - \frac{1}{T-1} \mathbf{S}_t. \quad (26)$$

We can then verify the following expressions for quickly computing the relevant quantities in (16) and (17):

$$\frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t^2) = \frac{T(T-2)}{(T-1)^2} \text{tr}(\mathbf{R}^2) - \frac{1}{T(T-1)^2} \sum_{t=1}^T \|\mathbf{y}_t\|_F^4, \quad (27)$$

$$\frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{S}_t) = \frac{T}{T-1} \text{tr}(\mathbf{R}^2) - \frac{1}{T(T-1)} \sum_{t=1}^T \|\mathbf{y}_t\|_F^4, \quad (28)$$

$$\frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t \mathbf{T}_0) = \text{tr}(\mathbf{R} \mathbf{T}_0), \quad (29)$$

$$\frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{S}_t \mathbf{T}_0) = \text{tr}(\mathbf{R} \mathbf{T}_0). \quad (30)$$

Plugging these into (16) and (17) and after some manipulations, we can rewrite the LOOCV cost function (15) as

$$\begin{aligned} J_{\text{CV}}(\rho, \tau) &= \frac{\rho T(\rho T - 2\rho - 2T + 2)}{(T-1)^2} \text{tr}(\mathbf{R}^2) \\ &\quad + 2\tau(\rho - 1) \text{tr}(\mathbf{R} \mathbf{T}_0) + \tau^2 \text{tr}(\mathbf{T}_0^2) \\ &\quad + \frac{1}{T} \left(\frac{\rho}{T-1} + 1 \right)^2 \sum_{t=1}^T \|\mathbf{y}_t\|_F^4. \end{aligned} \quad (31)$$

The optimal shrinkage coefficients can then be obtained analytically from the SCM \mathbf{R} , the shrinkage target \mathbf{T}_0 , and the training samples $\{\mathbf{y}_t\}$, as discussed below.

1) *Unconstrained shrinkage*: It can be verified from (19) and (30) that the optimal shrinkage coefficients (ignoring the nonnegative constraint $\rho \geq 0, \tau \geq 0$) satisfy

$$\tau = (1 - \rho) \frac{\text{tr}(\mathbf{R} \mathbf{T}_0)}{\text{tr}(\mathbf{T}_0^2)}. \quad (32)$$

The closed-form solution to ρ is given by

$$\rho_{\text{CV,SCM}}^* = \frac{\frac{T \text{tr}(\mathbf{R}^2)}{T-1} - \frac{(\text{tr}(\mathbf{R} \mathbf{T}_0))^2}{\text{tr}(\mathbf{T}_0^2)} - \frac{\sum_{t=1}^T \|\mathbf{y}_t\|_F^4}{T(T-1)}}{\frac{(T^2 - 2T) \text{tr}(\mathbf{R}^2)}{(T-1)^2} - \frac{(\text{tr}(\mathbf{R} \mathbf{T}_0))^2}{\text{tr}(\mathbf{T}_0^2)} + \frac{\sum_{t=1}^T \|\mathbf{y}_t\|_F^4}{T(T-1)^2}}. \quad (33)$$

In case $\rho_{\text{CV,SCM}}^* > 1$ or $\rho_{\text{CV,SCM}}^* < 0$, we apply (21) or (22), respectively, to determine the solution, using the expressions in (27)-(30).

Note that for the typical shrinkage target $\mathbf{T}_0 = \frac{\text{tr}(\mathbf{R})}{N} \mathbf{I}$, (32) results in $\tau_{\text{CV,SCM}}^* = 1 - \rho_{\text{CV,SCM}}^*$. This provides another justification for the convex linear combination design with an identity target, which has been widely adopted in the literature, e.g., [4]. This also shows that for such a special target the unconstrained solution is equivalent to the constrained solution, which does not hold for more general shrinkage targets.

2) *Constrained shrinkage*: For the widely considered convex linear combination with constraint $\rho + \tau = 1$, the optimal ρ (ignoring the nonnegative constraint) is computed as

$$\rho_{\text{CV,SCM}}^* = \frac{\frac{T \text{tr}(\mathbf{R}^2)}{T-1} - 2 \text{tr}(\mathbf{R} \mathbf{T}_0) + \text{tr}(\mathbf{T}_0^2) - \frac{\sum_{t=1}^T \|\mathbf{y}_t\|_F^4}{T(T-1)}}{\frac{(T^2 - 2T) \text{tr}(\mathbf{R}^2)}{(T-1)^2} - 2 \text{tr}(\mathbf{R} \mathbf{T}_0) + \text{tr}(\mathbf{T}_0^2) + \frac{\sum_{t=1}^T \|\mathbf{y}_t\|_F^4}{T(T-1)^2}}. \quad (34)$$

Similarly, in case a negative shrinkage coefficient is obtained, we set it to zero and let the other be one.

The above results show that the optimal shrinkage coefficients for the covariance matrix estimate (1) can be computed directly from the samples and shrinkage target, without the need of specifying any user parameters. The constrained shrinkage design may lead to certain performance loss as compared to the unconstrained one.

D. LOOCV Choice for OLS-Based Covariance Estimation

One advantage of the LOOCV method is that it can be applied to different covariance matrix estimators. In this subsection, we discuss the LOOCV method for OLS-based covariance matrix estimation. Note that most existing analytical solutions for choosing the shrinkage coefficients assume SCM and specific shrinkage targets and need to be re-derived for general cases. Also, in contrast to general applications of LOOCV which require a grid search of the parameters and thus a high computational complexity, we have shown that for SCM, fast analytical solutions can be obtained for choosing the shrinkage coefficients. This will also be the case for the OLS-based covariance matrix estimation.

Consider the case with observation $\mathbf{y} \in \mathbb{C}^N$ modeled as

$$\mathbf{y} = \mathbf{H} \mathbf{x} + \mathbf{z}, \quad (35)$$

where $\mathbf{H} \in \mathbb{C}^{N \times M}$ is a deterministic channel matrix and $\mathbf{z} \in \mathbb{C}^N$ a zero-mean, white noise with covariance matrix $\sigma^2 \mathbf{I}$, which is uncorrelated with the zero-mean input signal $\mathbf{x} \in \mathbb{C}^M$ with covariance matrix \mathbf{I} . If both training samples of \mathbf{x} and \mathbf{y} are known, we may first estimate the channel matrix \mathbf{H} and the covariance matrix of the noise \mathbf{z} using the ordinary

least squares (OLS) approach. Let the block of training data be (\mathbf{X}, \mathbf{Y}) , where the input signal \mathbf{X} can be designed to have certain properties such as being orthogonal. The OLS estimates of the channel matrix and noise variance are then obtained as

$$\widehat{\mathbf{H}} = \mathbf{Y}\mathbf{X}^\dagger (\mathbf{X}\mathbf{X}^\dagger)^{-1}, \quad (36)$$

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{TN} \left\| \mathbf{Y} - \widehat{\mathbf{H}}\mathbf{X} \right\|_F^2 \\ &= \frac{1}{TN} \text{tr} \left((\mathbf{Y} - \widehat{\mathbf{H}}\mathbf{X})(\mathbf{Y} - \widehat{\mathbf{H}}\mathbf{X})^\dagger \right) \\ &= \frac{1}{TN} \text{tr} \left(\mathbf{Y}(\mathbf{I} - \mathbf{X}^\dagger(\mathbf{X}\mathbf{X}^\dagger)^{-1}\mathbf{X})\mathbf{Y}^\dagger \right), \end{aligned} \quad (37)$$

where $\widehat{(\cdot)}$ denotes the estimate of a quantity. In this case, the covariance matrix of \mathbf{y} can be estimated as

$$\mathbf{R} = \widehat{\mathbf{H}}\widehat{\mathbf{H}}^\dagger + \widehat{\sigma}^2\mathbf{I}. \quad (38)$$

Such OLS-based covariance matrix estimation may be useful for designing signal estimation schemes in wireless communications. We can apply the linear shrinkage design (1) to enhance its accuracy and apply the LOOCV method (12) to choose the shrinkage coefficients. Note that in this case, in the t -th split, we generate the covariance matrix estimate \mathbf{R}_t by applying the OLS estimate to the leave-one-out samples $(\mathbf{X}_t, \mathbf{Y}_t)$ which are the subset of (\mathbf{X}, \mathbf{Y}) with the pair $(\mathbf{x}_t, \mathbf{y}_t)$ omitted. The LOOCV cost is the same as (15). In this case, the leave-one-out estimate of the covariance matrix for the t -th data split is

$$\mathbf{R}_t = \widehat{\mathbf{H}}_t\widehat{\mathbf{H}}_t^\dagger + \widehat{\sigma}_t^2\mathbf{I}, \quad (39)$$

where $\widehat{\mathbf{H}}_t$ and $\widehat{\sigma}_t^2$ denote the channel matrix and noise variance estimated from $(\mathbf{X}_t, \mathbf{Y}_t)$, respectively. A direct computation of (16) and (17) for evaluating the LOOCV cost performs OLS estimation for T times, which incurs significant complexity. The complexity can be greatly reduced by observing that the leave-one-out OLS estimate of the channel matrix is related to the OLS channel matrix estimate $\widehat{\mathbf{H}}$ in (36) by a rank-one update:

$$\widehat{\mathbf{H}}_t = \mathbf{Y}_t\mathbf{X}_t^\dagger \left(\mathbf{X}_t\mathbf{X}_t^\dagger \right)^{-1} = \widehat{\mathbf{H}} - \mathbf{e}_t\mathbf{f}_t^\dagger, \quad (40)$$

where

$$\mathbf{e}_t \triangleq \mathbf{y}_t - \widehat{\mathbf{H}}\mathbf{x}_t, \quad (41)$$

$$\mathbf{f}_t \triangleq \frac{1}{1 - \Phi_t} (\mathbf{X}\mathbf{X}^\dagger)^{-1}\mathbf{x}_t. \quad (42)$$

In the above,

$$\Phi_t \triangleq \mathbf{x}_t^\dagger (\mathbf{X}\mathbf{X}^\dagger)^{-1}\mathbf{x}_t \quad (43)$$

is the t -th diagonal entry of

$$\Phi = \mathbf{X}^\dagger (\mathbf{X}\mathbf{X}^\dagger)^{-1}\mathbf{X}. \quad (44)$$

Similarly, the leave-one-out estimate of the noise variance can be updated as

$$\begin{aligned} \widehat{\sigma}_t^2 &= \frac{1}{N(T-1)} \text{tr} \left((\mathbf{Y}_t - \widehat{\mathbf{H}}_t\mathbf{X}_t)(\mathbf{Y}_t - \widehat{\mathbf{H}}_t\mathbf{X}_t)^\dagger \right) \\ &= \widehat{\sigma}^2 - \delta_t, \end{aligned} \quad (45)$$

where

$$\delta_t = \frac{\|\mathbf{e}_t\|_F^2}{N(T-1)(1-\Phi_t)} - \frac{\widehat{\sigma}^2}{T-1}. \quad (46)$$

Note that both updates can be achieved with low complexity when a few matrices are computed in advance and reused. In this way, the covariance matrix estimate can be computed as

$$\mathbf{R}_t = \mathbf{R} - \delta_t\mathbf{I} - \mathbf{e}_t\phi_t^\dagger - \psi_t\mathbf{e}_t^\dagger, \quad (47)$$

where ϕ_t and ψ_t are defined as

$$\phi_t = \widehat{\mathbf{H}}\mathbf{f}_t, \quad (48)$$

$$\psi_t = \phi_t - \|\mathbf{f}_t\|_F^2\mathbf{e}_t. \quad (49)$$

This shows that the leave-one-out OLS covariance matrix estimate can be obtained from \mathbf{R} by corrections involving a scaled identity matrix and two rank-one updates. Eqn. (47) can be exploited to compute the closed-form LOOCV solution quickly. From (47), the most involved computation for finding the solution of the optimization problem (15) can be implemented as

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t^2) &= \text{tr}(\mathbf{R}^2) + \frac{N \sum_{i=1}^T \delta_i^2}{T} - \frac{2 \sum_{i=1}^T \delta_i}{T} \text{tr}(\mathbf{R}) \\ &+ \frac{1}{T} \sum_{i=1}^T \|\mathbf{e}_i\|_F^2 (\|\phi_i\|_F^2 + \|\psi_i\|_F^2) \\ &- \frac{2}{T} \sum_{i=1}^T \text{R}(\mathbf{e}_i^\dagger (\mathbf{R} - \delta_i\mathbf{I})(\phi_i + \psi_i) - \mathbf{e}_i^\dagger \psi_i \mathbf{e}_i^\dagger \phi_i), \end{aligned} \quad (50)$$

where $\text{R}(\cdot)$ denotes the real part of a scalar. When \mathbf{R} is already computed, the right-hand side of (50) can be evaluated using inner products and matrix-vector products. The terms $\text{tr}(\mathbf{T}_0^2)$ and $\text{tr}(\mathbf{T}_0\mathbf{S}_t)$ are the same as those for SCM. For the other two terms, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t\mathbf{S}_t) &= \frac{1}{T} \text{tr}(\mathbf{R}\mathbf{Y}\mathbf{Y}^\dagger) \\ &- \frac{1}{T} \sum_{t=1}^T \text{R}(\delta_t \|\mathbf{y}_t\|_F^2 + \mathbf{y}_t^\dagger \mathbf{e}_t \phi_t^\dagger \mathbf{y}_t + \mathbf{y}_t^\dagger \psi_t \mathbf{e}_t^\dagger \mathbf{y}_t), \end{aligned} \quad (51)$$

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t\mathbf{T}_0) &= \text{tr}(\mathbf{R}\mathbf{T}_0) - \frac{\sum_{t=1}^T \delta_t}{T} \text{tr}(\mathbf{T}_0) \\ &- \frac{1}{T} \sum_{t=1}^T \text{R}(\phi_t^\dagger \mathbf{T}_0 \mathbf{e}_t + \mathbf{e}_t^\dagger \mathbf{T}_0 \psi_t). \end{aligned} \quad (52)$$

Note that the computational complexities of (50)-(52) are low because the major operations are matrix-vector products and inner products.

E. Comparisons with Alternative Choices of Linear Shrinkage Coefficients

In the above, we have introduced LOOCV methods with analytical solutions for choosing the coefficients for linear

shrinkage covariance matrix estimators. We now discuss several alternative techniques which have received considerable attentions recently and compare them with the LOOCV methods proposed in this paper.

In 2004, Ledoit and Wolf (LW) [2] studied estimators that shrink SCM toward an identity target, i.e., $\mathbf{T}_0 = \mathbf{I}$. Such estimators do not alter the eigenvectors but shrink eigenvalues of the SCM, which is well supported by the fact that sample eigenvalues tend to be more spread than population eigenvalues. The optimal shrinkage coefficients under the MMSE criterion (3) can be written as

$$\rho^* = \frac{\alpha^2}{\delta^2}, \quad \tau^* = \frac{\beta^2}{\delta^2} \mu, \quad (53)$$

where the parameters $\mu \triangleq \frac{\text{tr}(\boldsymbol{\Sigma})}{N}$, $\delta^2 \triangleq \mathbb{E}[\|\mathbf{R} - \mu\mathbf{I}\|_F^2]$, $\beta^2 = \mathbb{E}[\|\boldsymbol{\Sigma} - \mathbf{R}\|_F^2]$, and $\alpha^2 = \|\boldsymbol{\Sigma} - \mu\mathbf{I}\|_F^2$ depend on the true covariance matrix $\boldsymbol{\Sigma}$ and other unknown statistics. Ref. [2] shows that $\delta^2 = \alpha^2 + \beta^2$ and proposes to approximate these quantities by their asymptotic estimates under $T \rightarrow \infty, N \rightarrow \infty, N/T \rightarrow c < \infty$, as

$$\widehat{\mu} = \frac{\text{tr}(\mathbf{R})}{N}, \quad \widehat{\delta}^2 = \|\mathbf{R} - \widehat{\mu}\mathbf{I}\|_F^2, \quad (54)$$

$$\widehat{\beta}^2 = \min \left(\widehat{\delta}^2, \frac{1}{T^2} \sum_{t=1}^T \|\mathbf{y}_t \mathbf{y}_t^\dagger - \mathbf{R}\|_F^2 \right), \quad \widehat{\alpha}^2 = \widehat{\delta}^2 - \widehat{\beta}^2, \quad (55)$$

which can all be computed from the training samples. By substituting these into (53), estimators that significantly outperform SCM are obtained, which also approach the oracle estimators when the training length is large enough.

The above LW estimator is extended by Stoica *et al* in 2008 [3] for complex-valued signals with general shrinkage targets \mathbf{T}_0 , with applications to knowledge-aided space-time adaptive processing (KA-STAP) in radar applications. Several estimators with similar performance are derived there. For the general linear combination (GLC) design of [3], it is shown that the oracle shrinkage coefficients for (1) satisfy

$$\rho^* = 1 - \frac{\tau^*}{\nu}, \quad (56)$$

where

$$\nu = \frac{\text{tr}(\mathbf{T}_0 \boldsymbol{\Sigma})}{\|\mathbf{T}_0\|_F^2}, \quad \tau^* = \nu \frac{\beta^2}{\mathbb{E}[\|\mathbf{R} - \nu \mathbf{T}_0\|_F^2]}. \quad (57)$$

The quantity β^2 is estimated in the same way as (55), and a computationally efficient expression for $\widehat{\beta}^2$ is given by

$$\widehat{\beta}^2 = \frac{1}{T^2} \sum_{t=1}^T \|\mathbf{y}_t\|_F^4 - \frac{1}{T} \|\mathbf{R}\|_F^2. \quad (58)$$

Furthermore, ν and $\mathbb{E}[\|\mathbf{R} - \nu \mathbf{T}_0\|_F^2]$ are estimated as $\widehat{\nu} = \frac{\text{tr}(\mathbf{T}_0 \mathbf{R})}{\|\mathbf{T}_0\|_F^2}$ and $\|\mathbf{R} - \widehat{\nu} \mathbf{T}_0\|_F^2$, respectively. This leads to the result given by Eqns. (34) and (35) of [3], which can recover the LW estimator [2] when the identity shrinkage target $\mathbf{T}_0 = \mathbf{I}$ is assumed.

More recently, Chen *et al* [4] derived the oracle approximating shrinkage (OAS) estimator, which assumes SCM, real-valued Gaussian samples, and scaled identity target with

$\mathbf{T}_0 = \frac{\text{tr}(\mathbf{R})}{N} \mathbf{I}$ and $\rho = 1 - \tau$. They first derive the oracle shrinkage coefficients for SCM obtained from i.i.d. Gaussian samples, which is determined by $N, T, \text{tr}(\boldsymbol{\Sigma})$ and $\text{tr}(\boldsymbol{\Sigma}^2)$. Then, they propose an iterative procedure to approach the oracle estimator. In the iterations, $\text{tr}(\boldsymbol{\Sigma}^2)$ and $\text{tr}(\boldsymbol{\Sigma})$ are estimated by $\text{tr}(\widehat{\boldsymbol{\Sigma}}_j \mathbf{R})$ and $\text{tr}(\widehat{\boldsymbol{\Sigma}}_j)$, respectively, where $\widehat{\boldsymbol{\Sigma}}_j$ is the covariance matrix estimate at the j -th iteration. It is further proved that $\widehat{\boldsymbol{\Sigma}}_j$ converges to the OAS estimator with the following analytical expression for τ :

$$\tau_{\text{OAS}}^* = \min \left(1, \frac{(1 - \frac{2}{N}) \text{tr}(\mathbf{R}^2) + (\text{tr}(\mathbf{R}))^2}{(T + 1 - \frac{2}{N}) [\text{tr}(\mathbf{R}^2) - \frac{(\text{tr}(\mathbf{R}))^2}{N}]} \right). \quad (59)$$

This approach achieves superior performance for (scaled) identity target and Gaussian data and dominates the LW estimator [2] when T is small. It was later generalized by Senneret *et al* [20] to a shrinkage target chosen as the diagonal entries of the SCM. Other related techniques include [14], which also assumes SCM, Gaussian data, and identity/diagonal shrinkage targets.

All the above techniques provide analytical solutions and achieve near-oracle performance when the underlying assumptions (e.g., large dimensionality, large size of training data, identity/diagonal shrinkage targets) hold. However, they also have limitations. A common restriction is that all these analytical solutions assume SCM and are not optimized for other types of covariance matrix estimators such as model-based estimators. In particular, the LW and GLC methods [2], [3], which employ asymptotic approximations, may exhibit a noticeable gap to the oracle choice when the sample support is low, which may be relevant in some applications. The OAS method [4] assumed identity target, but its extensions to more general cases, e.g., with multiple/general shrinkage targets, are not trivial. By contrast, the LOOCV method proposed in this paper allows different designs and achieves near-oracle performance in general.

Cross-validation has also been applied previously for choosing shrinkage coefficients for covariance matrix estimation. The key issues for applying this generic tool include finding appropriate predictive metrics for scoring the different estimators and fast computation schemes. In [10], [13], the Gaussian likelihood was chosen as such a proxy. The computations with likelihood are generally involved as multiple matrix inverses/determinants are required, and a grid search is required for finding the optimal parameters. In this paper, we use the distribution-free, Frobenius norm loss in (12) as the metric, which leads to analytical solutions and is computationally more tractable.

III. MULTI-TARGET SHRINKAGE

In Section 2, we have considered linear shrinkage designs with a single target. Multiple shrinkage targets may be used to further enhance performance, which may be obtained from a priori knowledge, e.g., a past covariance matrix estimate from older training samples or from neighboring frequencies. We can easily extend our proposed LOOCV method to multiple targets.

A. Oracle choice of shrinkage coefficients

Consider the multi-target shrinkage design

$$\widehat{\Sigma}_{\rho, \tau} = \rho \mathbf{R} + \sum_{k=1}^K \tau_k \mathbf{T}_k, \quad (60)$$

where all the shrinkage coefficients are nonnegative to guarantee PSD covariance matrix estimates, i.e.,

$$\rho \geq 0; \quad \tau_k \geq 0, \forall k. \quad (61)$$

The oracle multi-target shrinkage minimizes the squared Frobenius norm of the estimation error

$$J_{O,MT}(\rho, \tau) = \left\| \rho \mathbf{R} + \sum_{k=1}^K \tau_k \mathbf{T}_k - \Sigma \right\|_F^2, \quad (62)$$

which can be rewritten as

$$J_{O,MT}(\rho, \tau) = \begin{bmatrix} \rho \\ \tau \end{bmatrix}^T \mathbf{A}_{O,MT} \begin{bmatrix} \rho \\ \tau \end{bmatrix} - 2 \begin{bmatrix} \rho \\ \tau \end{bmatrix}^T \mathbf{b}_{O,MT} + \text{tr}(\Sigma^2), \quad (63)$$

where $\tau = [\tau_1, \tau_2, \dots, \tau_K]^T$,

$$\mathbf{A}_{O,MT} = \begin{bmatrix} \text{tr}(\mathbf{R}^2) & \text{tr}(\mathbf{R}\mathbf{T}_1) & \cdots & \text{tr}(\mathbf{R}\mathbf{T}_K) \\ \text{tr}(\mathbf{T}_1\mathbf{R}) & \text{tr}(\mathbf{T}_1^2) & \cdots & \text{tr}(\mathbf{T}_1\mathbf{T}_K) \\ \vdots & \vdots & \ddots & \vdots \\ \text{tr}(\mathbf{T}_K\mathbf{R}) & \text{tr}(\mathbf{T}_K\mathbf{T}_1) & \cdots & \text{tr}(\mathbf{T}_K^2) \end{bmatrix}, \quad (64)$$

$$\mathbf{b}_{O,MT} = \begin{bmatrix} \text{tr}(\mathbf{R}\Sigma) \\ \text{tr}(\mathbf{T}_1\Sigma) \\ \vdots \\ \text{tr}(\mathbf{T}_K\Sigma) \end{bmatrix}. \quad (65)$$

The oracle shrinkage coefficients can then be obtained by solving the problem of minimizing the cost function $J_{O,MT}(\rho, \tau)$ of (63), which is a strictly convex quadratic program (SQP) with $K + 1$ variables.

B. LOOCV choice of shrinkage coefficients

We now extend the LOOCV method in Section 2 to the multi-target shrinkage here. Following the same treatment as in Section II-B, in each split of the training data, \mathbf{R}_t and \mathbf{S}_t are constructed to generate and validate the covariance matrix estimate, respectively. The multiple shrinkage coefficients are chosen to minimize the LOOCV cost

$$J_{CV,MT}(\rho, \tau) = \frac{1}{T} \sum_{t=1}^T \left\| \rho \mathbf{R}_t + \sum_{k=1}^K \tau_k \mathbf{T}_k - \mathbf{S}_t \right\|_F^2. \quad (66)$$

The above cost function can be rewritten in a form similar to (15) as

$$J_{CV,MT}(\rho, \tau) = \begin{bmatrix} \rho \\ \tau \end{bmatrix}^T \mathbf{A}_{CV,MT} \begin{bmatrix} \rho \\ \tau \end{bmatrix} - 2 \begin{bmatrix} \rho \\ \tau \end{bmatrix}^T \mathbf{b}_{CV,MT} + \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{S}_t^2) \quad (67)$$

with

$$\mathbf{A}_{CV,MT} = \begin{bmatrix} \frac{\sum_{t=1}^T \text{tr}(\mathbf{R}_t^2)}{T} & \frac{\sum_{t=1}^T \text{tr}(\mathbf{R}_t\mathbf{T}_1)}{T} & \cdots & \frac{\sum_{t=1}^T \text{tr}(\mathbf{R}_t\mathbf{T}_K)}{T} \\ \frac{\sum_{t=1}^T \text{tr}(\mathbf{T}_1\mathbf{R}_t)}{T} & \text{tr}(\mathbf{T}_1^2) & \cdots & \text{tr}(\mathbf{T}_1\mathbf{T}_K) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sum_{t=1}^T \text{tr}(\mathbf{T}_K\mathbf{R}_t)}{T} & \text{tr}(\mathbf{T}_K\mathbf{T}_1) & \cdots & \text{tr}(\mathbf{T}_K^2) \end{bmatrix}, \quad (68)$$

$$\mathbf{b}_{CV,MT} = \begin{bmatrix} \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t\mathbf{S}_t) \\ \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{T}_1\mathbf{S}_t) \\ \vdots \\ \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{T}_K\mathbf{S}_t) \end{bmatrix}. \quad (69)$$

The constant entries of $\mathbf{A}_{CV,MT}$ and $\mathbf{b}_{CV,MT}$ can be computed in the same way as for the single-target case. When K is small, which is typically the case, the solution that minimizes the LOOCV cost can be found quickly using standard optimization tools. Alternatively, we may find first the global optimizer that ignores the nonnegative constraint by

$$\begin{bmatrix} \rho_{CV,MT}^* \\ \tau_{CV,MT}^* \end{bmatrix} = \mathbf{A}_{CV,MT}^{-1} \mathbf{b}_{CV,MT}, \quad (70)$$

and check if the nonnegative condition is satisfied. If a negative shrinkage coefficient is produced, we then consider the boundaries of $\rho \geq 0, \tau_k \geq 0, k = 1, 2, \dots, K$, which are equivalent to removing a certain number of shrinkage targets from the shrinkage design. The solution can be found in exactly the same way as (70) but with fewer targets.

Similarly to the single-target case, we may also consider a constrained case, where the shrinkage targets $\{\mathbf{T}_k\}$ have the same trace as the estimated covariance matrix \mathbf{R} , and

$$\rho + \sum_{k=1}^K \tau_k = 1. \quad (71)$$

Then the LOOCV cost function can be rewritten as

$$J_{CV,MT}(\tau) = \frac{1}{T} \sum_{t=1}^T \left\| \sum_{k=1}^K \tau_k \mathbf{A}_{kt} + \mathbf{B}_t \right\|_F^2, \quad (72)$$

where

$$\mathbf{A}_{kt} \triangleq \mathbf{T}_k - \mathbf{R}_t, 1 \leq k \leq K, 1 \leq t \leq T, \quad (73)$$

$$\mathbf{B}_t \triangleq \mathbf{R}_t - \mathbf{S}_t, 1 \leq t \leq T. \quad (74)$$

The optimal shrinkage coefficients can be found similarly as for the unconstrained case by minimizing

$$J_{CV,MT}(\tau) = \tau^T \mathbf{A}'_{CV,MT} \tau - 2 \tau^T \mathbf{b}'_{CV,MT} + \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{B}_t^2), \quad (75)$$

where the entries of $\mathbf{A}'_{CV,MT}$ and $\mathbf{b}'_{CV,MT}$ are defined by

$$[\mathbf{A}'_{CV,MT}]_{mn} \triangleq \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{A}_{mt} \mathbf{A}_{nt}), 1 \leq m, n \leq K, \quad (76)$$

$$[\mathbf{b}'_{CV,MT}]_k \triangleq \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{A}_{kt} \mathbf{B}_t), 1 \leq k \leq K. \quad (77)$$

These entries may also be evaluated quickly. For example, with SCM,

$$[\mathbf{A}'_{CV,MT}]_{mn} = \text{tr}(\mathbf{T}_m \mathbf{T}_n) - \text{tr}((\mathbf{T}_m + \mathbf{T}_n) \mathbf{R}) + \eta, \quad (78)$$

$$[\mathbf{b}'_{CV,MT}]_k = \frac{T}{T-1} \text{tr}(\mathbf{R}^2) - \frac{1}{T(T-1)} \sum_{t=1}^T \|y_t\|_F^4 - \eta, \quad (79)$$

where

$$\eta = \frac{1}{T} \sum_{t=1}^T \text{tr}(\mathbf{R}_t^2)$$

can be computed using (27). The solution to τ can be found as

$$\tau_{CV,MT}^* = \mathbf{A}'_{CV,MT}{}^{-1} \mathbf{b}'_{CV,MT} \quad (80)$$

if the nonnegative condition is satisfied. Otherwise, find the solution in a similar way as for the unconstrained case.

Note that for multi-target shrinkage, Lancewicki and Adjem [17] recently introduced another method for finding the shrinkage coefficients. They assume SCM and shrinkage targets which belong to a set that can be characterized by Eqn. (21) of [17]. Then, they follow the Ledoit-Wolf (LW) framework [2] to derive unbiased estimates of the unknown coefficients needed for minimizing the expectation of the cost in (62), based on which $\{\rho, \tau\}$ can be optimized. By contrast, our approach resorts to a LOOCV estimate of the cost in (10), which does not rely on the aforementioned assumptions in [17]. As will be shown later, the LOOCV method can achieve similar performance as [17] for the shrinkage targets considered there. However, it can be applied to general estimators other than SCM and shrinkage targets which are not covered by Eqn. (21) of [17], offering wider applicability.

IV. NUMERICAL EXAMPLES

In this section, we present numerical examples to demonstrate the effectiveness of the proposed shrinkage design and compare it with alternative methods. The quality of covariance matrix estimation is measured by the MSE normalized by the average of the squared Frobenius norm $\|\Sigma\|_F^2$, i.e.,

$$\text{NMSE}_{\Sigma} \triangleq \frac{\mathbb{E}[\|\hat{\Sigma}_{\rho,\tau} - \Sigma\|_F^2]}{\mathbb{E}[\|\Sigma\|_F^2]}. \quad (81)$$

We show examples of covariance matrix estimation and its applications in array signal processing. We denote by $\mathcal{N}(\mu, \sigma^2)$ a real-valued Gaussian distribution with mean μ and variance σ^2 .

Example 1: Shrinkage toward an identity target: We first consider a real-valued example with an autoregressive (AR) covariance matrix, whose (i, j) -th entry is given by

$$[\Sigma]_{i,j} = r^{|i-j|}, 1 \leq i, j \leq N, \quad (82)$$

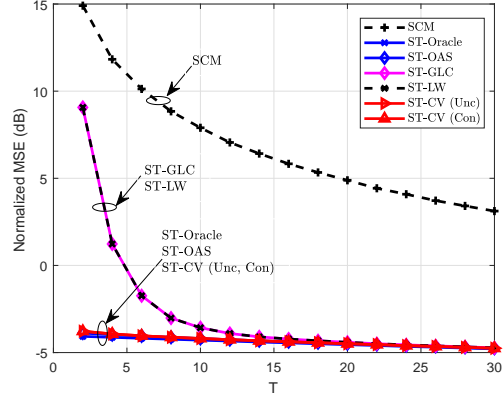


Figure 1. NMSE of single-target (ST) shrinkage estimates of an AR covariance matrix with $N = 100$, $r = 0.5$, $\mathbf{T}_0 = \frac{\text{tr}(\mathbf{R})}{N} \mathbf{I}$. “LW”, “GLC” and “OAS” refer to the methods of [2], [3] and [4], respectively, which are also described in Section II-E; “CV” refers to our proposed LOOCV method; “Oracle” refers to the coefficient choice in Section II-A; and “Con” and “Unc” indicate that the constraint $\rho + \tau = 1$ is imposed or not, respectively.

which has been widely considered for evaluating covariance matrix estimation techniques [4]-[7]. Let $\Sigma^{1/2}$ be the Cholesky factor of Σ . The training samples are randomly generated as $\mathbf{y}_t = \Sigma^{1/2} \mathbf{n}_t$, where \mathbf{n}_t consists of i.i.d. entries drawn from $\mathcal{N}(0, 1)$. The typical shrinkage target $\mathbf{T}_0 = \frac{\text{tr}(\mathbf{R})}{N} \mathbf{I}$ is considered for single-target shrinkage. Our proposed LOOCV method is compared with the widely used alternative methods [2]-[4] for choosing the shrinkage coefficients. The simulation results (averaged over 1000 repetitions for each training length T) in Fig. 1 confirm that the LOOCV methods with and without the constraint $\rho + \tau = 1$ produce the same results for the scaled identity target and they achieve performance almost identical to the OAS estimator [4], which was derived by assuming Gaussian data and identity target. The LW [2] and GLC [3] methods, which are equivalent for the scaled identity target here, do not perform well for very low sample support, but are able to approximate the oracle choice very well when more samples are available, which is consistent with the observations from [4]. All of these shrinkage designs significantly outperform the SCM, confirming the effectiveness of shrinkage for covariance matrix estimation. Recall that these methods were derived using different strategies and assumptions and have different analytical solutions.

Example 2: Shrinkage toward a nondiagonal target: We then consider an example of the linear model given by (35). For each training length, 1000 random realizations of $\Sigma = \mathbf{H}\mathbf{H}^\dagger + \sigma^2 \mathbf{I}$ are generated and estimated through training, where $\sigma^2 = 0.1$. The entries of \mathbf{H} are independently generated from $\mathcal{N}(0, 1)$ and then fixed for the whole training process. Given \mathbf{H} , T training samples are generated by $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}$, with the entries of \mathbf{x} and \mathbf{z} generated independently from $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, \sigma^2)$, respectively. In order to demonstrate the effectiveness of the LOOCV method for general shrinkage targets, we assume a scenario where \mathbf{H} is slowly time-varying and the shrinkage target \mathbf{T}_0 can be constructed as a well-

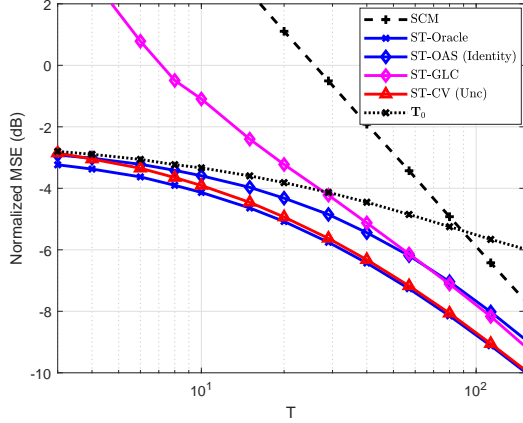


Figure 2. NMSE of single-target (ST) shrinkage estimation of covariance matrix for the linear model (35) with $N = 50$, $M = 50$, $\sigma^2 = 0.1$. The non-diagonal shrinkage target is constructed from the estimate of a past covariance matrix. The result indicated by “ T_0 ” corresponds to estimating Σ as T_0 . “Identity” indicates a scaled identity shrinkage target is used instead. We can show that imposing the constraint (23) leads to negligible change in performance for the proposed LOOCV approach.

conditioned estimate of a past covariance matrix

$$\Sigma^{\text{past}} = \mathbf{H}^{\text{past}} \mathbf{H}^{\text{past}\dagger} + \sigma^2 \mathbf{I}, \quad (83)$$

where

$$\mathbf{H}^{\text{past}} = \mathbf{H} + \Delta, \quad (84)$$

and the entries of Δ are independently drawn from $\mathcal{N}(0, 0.2)$ and are fixed for each repetition. Specifically, we construct T_0 as the shrinkage estimate of Σ^{past} using SCM and the scaled identity target. This construction is similar to the knowledge-aided target considered in [3] and the resulting T_0 is not diagonal. We assume that the numbers of samples used for estimating Σ and T_0 are both equal to T . The simulation results are included in Fig. 2 for the normalized MSE. It can be seen that the LOOCV methods generally achieve near-oracle performance and outperform the GLC method. Also, the non-diagonal shrinkage target achieves better performance than the scaled identity target.

Example 3: Shrinkage with multiple targets: A multi-target example is illustrated in Fig. 3. An AR covariance matrix is estimated by shrinking SCM with three targets which can be represented by Eqn. (21) of [17]: $T_1 = \frac{\text{tr}(\mathbf{R})}{N} \mathbf{I}$, $T_2 = \text{Diag}(\mathbf{R})$, and T_3 is a symmetric, Toeplitz matrix which was considered in [17]:

$$\mathbf{T}_3 = \frac{\text{tr}(\mathbf{R})}{N} \mathbf{I} + \sum_{i=1}^{N-1} \frac{\text{tr}(\mathbf{C}_i \mathbf{R})}{2(N-i)} \mathbf{C}_i, \quad (85)$$

where \mathbf{C}_i is a symmetric, Toeplitz matrix with unit entries on the i -th sub- and super-diagonals and zeros elsewhere. It is seen that multi-target shrinkage can significantly outperform single-target shrinkage with $T_0 = \frac{\text{tr}(\mathbf{R})}{N} \mathbf{I}$ when the number of samples is large enough. For the oracle parameter choices, the unconstrained shrinkage design, which allows a larger set of shrinkage factors to be chosen, can noticeably outperform the design constrained by (71). However, when the proposed

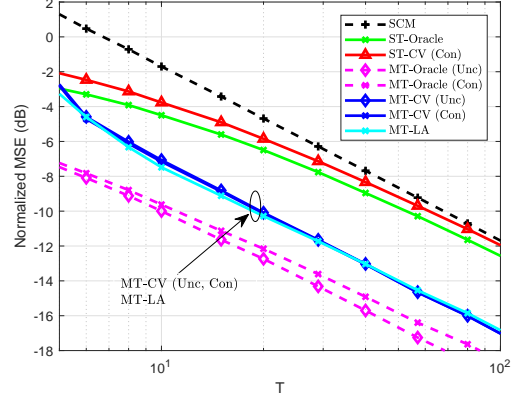


Figure 3. MSE of covariance matrix estimation with multi-target (MT) shrinkage and LOOCV parameter choices. AR covariance matrix with $N = 50$, $r = 0.9$ is assumed. “LA” refers to the method proposed by Lancelwicki and Aladjem [17]. Note that the LOOCV methods and the LA method achieve similar performance for this example.

LOOCV methods are used, the gap is significantly reduced. We can show that when the number of samples is small, using a more parsimonious design with constrained shrinkage coefficients or fewer shrinkage targets may achieve better performance. It is seen that the multi-target shrinkage method of [17] (indicated by “MT-LA” in Fig. 3) performs similarly to the LOOCV method for this example. Note that the method of [17] assumes SCM and shrinkage targets satisfying certain structures and does not apply directly to model-based covariance matrix estimation or more general shrinkage targets.

Example 4: Application to MMSE estimation of MIMO channels. A potential application of the proposed technique is the design of MMSE estimator of MIMO channels. Consider a point-to-point MIMO system with N_t transmitting antennas and N_r receiving antennas. Let B be the length of the pilot sequence. The received signal matrix during the training stage is modelled as

$$\mathbf{Y} = \mathbf{H} \mathbf{P} + \mathbf{N}, \quad (86)$$

where $\mathbf{Y} \in \mathbb{C}^{N_r \times B}$ is the received signal matrix, $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ the channel matrix, $\mathbf{P} \in \mathbb{C}^{N_t \times B}$ the pilot matrix, and $\mathbf{N} \in \mathbb{C}^{N_r \times B}$ the noise which is uncorrelated with \mathbf{H} . Vectorizing \mathbf{Y} in (86) gives

$$\mathbf{y} = \tilde{\mathbf{P}} \mathbf{h} + \mathbf{n}, \quad (87)$$

where $\mathbf{y} = \text{vec}(\mathbf{Y})$, $\tilde{\mathbf{P}} = \mathbf{P}^T \otimes \mathbf{I}$, $\mathbf{h} = \text{vec}(\mathbf{H})$, $\mathbf{n} = \text{vec}(\mathbf{N})$, $\text{vec}(\cdot)$ denotes vectorization, and \otimes denotes Kronecker product. We assume a Rayleigh fading channel and denote by $\Sigma_{\mathbf{h}} \in \mathbb{C}^{N_t N_r \times N_t N_r}$ the covariance matrix of the channel vector \mathbf{h} . We also assume that the disturbance \mathbf{n} is complex Gaussian-distributed with a zero mean and identity covariance matrix.

Given $\Sigma_{\mathbf{h}}$, the MMSE estimate of \mathbf{h} from \mathbf{y} can be computed as [41]

$$\hat{\mathbf{h}}_{\text{MMSE}} = \Sigma_{\mathbf{h}} \tilde{\mathbf{P}}^\dagger (\tilde{\mathbf{P}} \Sigma_{\mathbf{h}} \tilde{\mathbf{P}}^\dagger + \mathbf{I})^{-1} \mathbf{y}. \quad (88)$$

The covariance matrix $\Sigma_{\mathbf{h}}$, which can be very large, must be estimated in order to compute $\hat{\mathbf{h}}_{\text{MMSE}}$. In communication

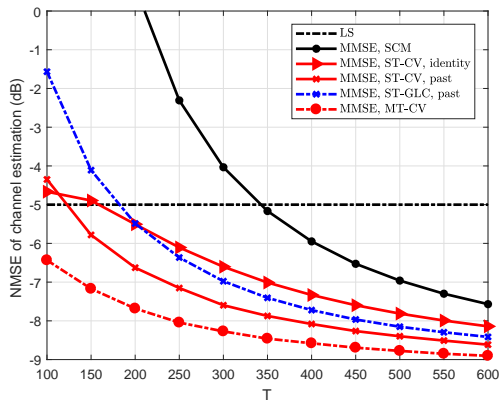


Figure 4. Performance of MMSE estimation of MIMO channels with the channel covariance matrix estimated using different estimators with $N_t = N_r = B = 10$. Pilot-to-noise ratio is 5 dB. “LS” refers to the LS estimator in (89); “MMSE” refers to the MMSE channel estimator (88) constructed using estimated covariance matrices; “identity” and “past” represent shrinkage targets chosen as the scaled identity matrix and the estimate of a past covariance matrix, respectively; “MT-CV” uses both the identity target and the target set as a past estimate.

systems, \mathbf{h} is not directly observable and thus the SCM estimator can not be directly applied to generate $\Sigma_{\mathbf{h}}$. One may estimate $\Sigma_{\mathbf{h}}$ from least squares (LS) estimates of \mathbf{H} , i.e.,

$$\hat{\mathbf{H}}_{\text{LS}} = \mathbf{Y}\mathbf{P}^\dagger(\mathbf{P}\mathbf{P}^\dagger)^{-1}. \quad (89)$$

When orthogonal training signal with $\mathbf{P} = \sqrt{P}\mathbf{I}$ is applied, where P determines the power for training signals, it can be shown that

$$\hat{\mathbf{H}}_{\text{LS}} = \frac{1}{\sqrt{P}}\mathbf{Y} = \frac{1}{\sqrt{P}}(\mathbf{H}\mathbf{P} + \mathbf{N}) = \mathbf{H} + \frac{1}{\sqrt{P}}\mathbf{N}. \quad (90)$$

Denote by $\hat{\mathbf{h}}_{\text{LS}}$ the vectorization of $\hat{\mathbf{H}}_{\text{LS}}$. It can be shown that the covariance matrix of $\hat{\mathbf{h}}_{\text{LS}}$ is

$$\Sigma_{\hat{\mathbf{h}}_{\text{LS}}} \triangleq \text{E}[\hat{\mathbf{h}}_{\text{LS}}\hat{\mathbf{h}}_{\text{LS}}^\dagger] = \Sigma_{\mathbf{h}} + \frac{1}{P}\mathbf{I}. \quad (91)$$

Therefore, if $\Sigma_{\hat{\mathbf{h}}_{\text{LS}}}$ is estimated as $\hat{\Sigma}_{\hat{\mathbf{h}}_{\text{LS}}}$, we can then use (91) to estimate $\Sigma_{\mathbf{h}}$ as

$$\hat{\Sigma}_{\mathbf{h}} = \hat{\Sigma}_{\hat{\mathbf{h}}_{\text{LS}}} - \frac{1}{P}\mathbf{I},$$

which can be used in (88). The estimation of $\Sigma_{\hat{\mathbf{h}}_{\text{LS}}}$ can be achieved using the different shrinkage estimators introduced in this paper.

An example is shown in Fig. 4. The covariance matrix is assumed to be

$$\Sigma_{\mathbf{h}} = \Sigma_t \otimes \Sigma_r, \quad (92)$$

where Σ_t and Σ_r are, respectively, the transmitter side and receiver side covariance matrix, with entries given by

$$[\Sigma_t]_{i,j} = \begin{cases} r_t^{|i-j|}, & i \geq j \\ (r_t^*)^{|i-j|}, & i < j \end{cases}, \quad (93)$$

$$[\Sigma_r]_{i,j} = \begin{cases} r_r^{|i-j|}, & i \geq j \\ (r_r^*)^{|i-j|}, & i < j \end{cases}, \quad (94)$$

$r_t = 0.7e^{-j0.9349\pi}$ and $r_r = 0.9e^{-j0.9289\pi}$. While applying shrinkage to estimate $\Sigma_{\hat{\mathbf{h}}_{\text{LS}}}$, two shrinkage targets are tested: the identity matrix and the shrinkage estimate (with a scaled identity target) of a past covariance matrix. The second is considered based on the assumption that $\Sigma_{\mathbf{h}}$ is slowly varying in time and a well-conditioned estimate of a past covariance matrix $\Sigma_{\mathbf{h}}^{\text{past}}$ can be available. In our simulations, $\Sigma_{\mathbf{h}}^{\text{past}}$ is modeled by randomly perturbing r_t and r_r in (93) and (94) by δ_t and δ_r whose real and imaginary parts are both randomly and uniformly generated from $[-\frac{1}{10\sqrt{2}}, \frac{1}{10\sqrt{2}}]$. The normalized MSE of channel estimation is defined as

$$\text{NMSE}_{\mathbf{h}} \triangleq \frac{\text{E}[||\hat{\mathbf{h}}_{\text{MMSE}} - \mathbf{h}||_F^2]}{\text{E}[||\mathbf{h}||_F^2]}, \quad (95)$$

where $\hat{\mathbf{h}}_{\text{MMSE}}$ is the MMSE channel estimate obtained from (88) with the true channel covariance matrix replaced by its shrinkage estimate.

From the simulation results in Fig. 4, when the number of samples T of channel estimates is small, the MMSE channel estimator constructed using the SCM estimate of $\Sigma_{\mathbf{h}}$ is poorer than the LS estimator which does not require any knowledge of $\Sigma_{\mathbf{h}}$. Therefore, an accurate estimate of the covariance matrix is necessary to exploit the potential of the MMSE channel estimator. Shrinkage with LOOCV choice of the shrinkage coefficients improves the performance of the MMSE channel estimator by providing a better estimate of $\Sigma_{\mathbf{h}}$. Two-target shrinkage can further enhance performance. Note that the multi-target method of [17] is not directly applicable to the shrinkage target used here. Similarly to [42], we do not exploit the Kronecker product structure in (92) and the exponential modeling of (93) and (94) while estimating the covariance matrix and similar trends can be observed when the channel covariance matrix follows different models such as those in [43], [44].

Example 5: Application to LMMSE signal estimation: Another example application is the design of linear minimum mean squared error (LMMSE) estimator [45], [46] for estimating the transmitted signal \mathbf{x} in MIMO communications. The received signal is modeled by (35) and the LMMSE estimate of \mathbf{x} is obtained as

$$\hat{\mathbf{x}} = \mathbf{H}^\dagger \Sigma_y^{-1} \mathbf{y}, \quad (96)$$

where we have assumed that \mathbf{x} has identity covariance matrix and Σ_y is the covariance matrix of \mathbf{y} . The OLS-based covariance matrix estimation in Section II-D can be used to estimate Σ_y in (96). In Fig. 5, we show an example where the shrinkage target \mathbf{T}_0 is chosen as the diagonal matrix of the OLS estimate (38) of the covariance matrix. This results in a shrunk LMMSE signal estimator

$$\hat{\mathbf{x}} = \hat{\mathbf{H}}^\dagger (\rho(\hat{\mathbf{H}}\hat{\mathbf{H}}^\dagger + \hat{\sigma}^2\mathbf{I}) + \tau\mathbf{T}_0)^{-1} \mathbf{y}. \quad (97)$$

Orthogonal training of length T constructed from the discrete Fourier transform (DFT) matrix is assumed for the OLS channel estimate and finding the shrinkage coefficients using our proposed LOOCV method is achieved at a low complexity.

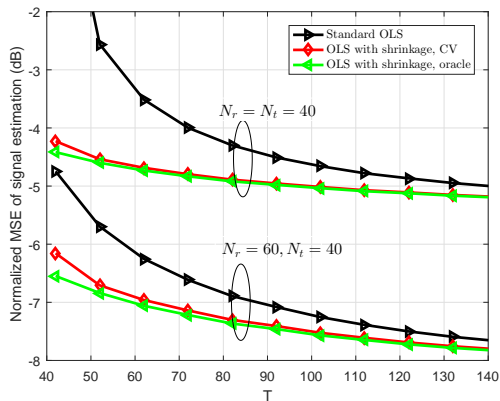


Figure 5. Performance of the LMMSE signal estimator with channel matrix and received signal’s covariance matrices estimated using OLS and shrinkage. The entries of \mathbf{H} are independently generated from complex Gaussian distribution with zero mean and variance $1/40$, and the noise variance $\sigma^2 = 0.1$.

The normalized MSE of signal estimation is defined as

$$\text{NMSE}_{\mathbf{x}} \triangleq \frac{\mathbb{E}[||\hat{\mathbf{x}} - \mathbf{x}||_F^2]}{\mathbb{E}[||\mathbf{x}||_F^2]}. \quad (98)$$

Fig. 5 presents the simulation results averaged over 1000 random realizations of \mathbf{H} for each T . It can be seen that the shrinkage estimate of the covariance matrix can lead to noticeable improvement of the MSE performance of signal estimation. The resulting performance can approach the oracle choice of (ρ, τ) that minimizes the MSE of estimating \mathbf{x} [35]. Note that in contrast to the cross-validation methods in [35] and [36] which choose shrinkage factors by a grid search for optimizing the signal estimation performance, the method proposed in this paper has an analytical solution and optimizes covariance matrix estimation. It also differs from [47] which targets the design of a signal estimator that shrinks the sample LMMSE filter toward the matched filter.

Example 6: Application to MVDR beamforming: Finally, we show an example application to minimum variance distortionless response (MVDR) beamforming [31], [33]. We assume a $N = 30$ -element uniform linear array (ULA) with half-wavelength spacing between neighboring antennas. As in [33], we assume that the desired complex Gaussian signal has an angle of arrival (AoA) of $\theta_0 = 0^\circ$ and there are 8 complex Gaussian interferences in the directions $\{\theta_m\} = \{8^\circ, -15^\circ, 23^\circ, -21^\circ, 46^\circ, -44^\circ, -85^\circ, 74^\circ\}$, all with an average power 10 dB higher than the desired signal. The noise is assumed to be additive white Gaussian noise (AWGN) with an average power 10 dB lower than the desired signal. The MVDR beamformer is given by

$$\mathbf{w} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{s}}{\mathbf{s}^\dagger \boldsymbol{\Sigma}^{-1}\mathbf{s}}, \quad (99)$$

where \mathbf{s} is the steering vector of the desired signal and $\boldsymbol{\Sigma}$ is the covariance matrix of the received signal. We consider a practical scenario where the desired signal’s steering vector suffers from an AoA error uniformly distributed in $[-5^\circ, 5^\circ]$ and $\boldsymbol{\Sigma}$ is estimated from the training samples by shrinking the SCM \mathbf{R} toward the scaled identity matrix $\frac{\text{tr}(\mathbf{R})}{N}\mathbf{I}$. We

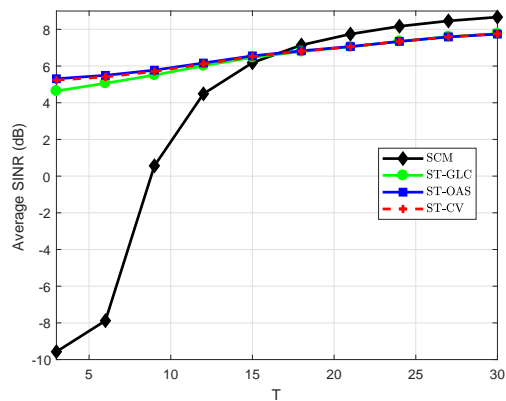


Figure 6. Average output SINR for a MVDR beamformer with AoA mismatch and the estimated covariance matrix. The results labeled by “SCM” is obtained by replacing $\boldsymbol{\Sigma}^{-1}$ in (99) with the pseudo-inverse of the SCM. Note that the LOOCV and OAS methods achieve almost the same performance, which is slightly better than the GLC method when T is very small.

focus on the low-sample-support case and compare the result with an approach that uses the pseudo-inverse of the SCM for computing \mathbf{w} . The output signal-to-interference-and-noise ratio (SINR) averaged over 1000 repetitions are plotted in Fig. 6. It is seen that though the proposed approach targets covariance matrix estimation only and is not optimized for beamformer designs, it still provides noticeable gains as compared to the pseudo-inverse approach in the low-sample-support regime.

V. CONCLUSIONS

In this paper, we have introduced a leave-one-out cross-validation (LOOCV) method for choosing the coefficients for linear shrinkage covariance matrix estimators. By employing a quadratic loss as the LOOCV prediction error, analytical expressions of the optimal shrinkage coefficients are obtained, which do not require a grid search of the parameters. As a result, the coefficients can be computed at low costs for the SCM- and OLS-based estimation of the covariance matrix. The LOOCV method is generic in the sense that it can be applied to different covariance matrix estimation methods and different shrinkage targets. Numerical examples show that it can approximate the oracle parameter choices in general and have wider applications than several existing analytical methods that have been widely applied.

Zero-mean signals have been assumed in this paper. When nonzero-mean signals are considered, our proposed approach may be applied after subtracting an estimate of the mean from the samples. However, the inaccuracy in the mean vector estimate may introduce extra errors to the covariance matrix estimation. Jointly estimating the mean and covariance matrix in a robust manner may be further explored. Other future work includes theoretical study of the properties of the proposed approach and low-complexity cross-validation schemes for choosing shrinkage factors for specific signal processing applications such as beamforming, space-time adaptive processing, correlation analysis, etc.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Antonio Napolitano and the anonymous reviewers for their constructive comments which have greatly improved the paper. This work was supported in part by an International Links grant of University of Wollongong (UOW), Australia, and in part by NSFC under Grant 61601325.

REFERENCES

- [1] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison-Wesley, Boston, 1991.
- [2] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, pp. 365-411, 2004.
- [3] P. Stoica, J. Li, X. Zhu, and J. R. Guerci, "On using a priori knowledge in space-time adaptive processing," *IEEE Trans. Sig. Process.*, vol. 56, no. 6, pp. 2598-2602, 2008.
- [4] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Sig. Process.*, vol. 58, no. 10, pp. 5016-5029, 2010.
- [5] L. Du, J. Li, and P. Stoica, "Fully automatic computation of diagonal loading levels for robust adaptive beamforming," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 449-458, 2010.
- [6] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Statist.*, vol. 36, no. 1, pp. 199-227, 2008.
- [7] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Ann. Statist.*, vol. 36, no. 6, pp. 2577-2604, 2008.
- [8] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proc. Third Berkeley Symp. Math. Statist. Prob.*, 1, pp. 197-206, 1956.
- [9] L. R. Haff, "Empirical Bayes estimation of the multivariate normal covariance matrix," *Ann. Statist.*, vol. 8, no. 3, pp. 586-597, 1980.
- [10] J. P. Hoffbeck and D.A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp.763-767, 1996.
- [11] M. Daniels and R. Kass, "Shrinkage estimators for covariance matrices," *Biometrics*, vol. 57, pp. 1173-1184, 2001.
- [12] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statist. Appl. Genetics Molecular Biol.*, vol. 4, 2005.
- [13] D. I. Warton, "Penalized normal likelihood and ridge regularization of correlation and covariance matrices," *JASA*, vol. 103, no. 481, pp. 340-49, 2008.
- [14] T. J. Fisher and X. Sun, "Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix," *Comp. Statist. Data Analysis*, 55, 1909-1918, 2011.
- [15] X. Chen, Z. Jane Wang, and M. J. McKeown, "Shrinkage-to-tapering estimation of large covariance matrices," *IEEE Trans. Sig. Process.*, vol. 60, pp. 5640-5656, 2012.
- [16] J. Theiler, "The incredible shrinking covariance estimator," *Proc. SPIE.*, vol. 8391, pp. 83910P, 2012.
- [17] T. Lancewicki and M. Aladjem, "Multi-target shrinkage estimation for covariance matrices," *IEEE Trans. Sig. Process.*, vol. 62, no. 24, pp. 6380-6390, 2014.
- [18] Y. Ikeda, T. Kubokawa, and M. S. Srivastava, "Comparison of linear shrinkage estimators of a large covariance matrix in normal and non-normal distributions," *Comput. Stat. Data Anal.*, vol. 95, 95-108, 2016.
- [19] T. Tong, C. Wang, and Y. Wang, "Estimation of variances and covariances for high-dimensional data: a selective review," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 6, no. 4, pp. 255-264, 2014.
- [20] M. Sennere, Y. Malevergne, P. Abry, G. Perrin, and L. Jaffrs, "Covariance versus precision matrix estimation for efficient asset allocation," *IEEE J. Sel. Topics Sig. Process.*, vol. 10, no. 6, pp. 982-993, Sept. 2016.
- [21] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *Econom. J.* 19, no. 1, C1-C32, 2016.
- [22] O. Ledoit and M. Wolf, "Nonlinear shrinkage estimation of large-dimensional covariance matrices," *Ann. Statist.*, vol. 40, no. 2, 1024-1060, 2012.
- [23] C. Lam, "Nonparametric eigenvalue-regularized precision or covariance matrix estimator," *Ann. Statist.*, vol. 44, no. 3, pp. 928-953, 2016.
- [24] A. Aubry, A. De Maio, L. Pallotta, and A. Farina, "Maximum likelihood estimation of a structured covariance matrix with a condition number constraint," *IEEE Trans. Signal Process.*, vol. 60, pp. 3004-3021, 2012.
- [25] J.-H. Won, J. Lim, S.-J. Kim, and B. Rajaratnam, "Condition-number regularized covariance estimation," *J. Roy. Statist. Soc. B*, vol. 75, pp. 427-450, Jun. 2013.
- [26] A. Kourtis, G. Dotsis, and R. N. Markellos, "Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix," *J. Bank. Financ.*, vol. 36, no. 9, pp.2522-2531, 2012.
- [27] M. Zhang, F. Rubio, and D. P. Palomar, "Improved calibration of high-dimensional precision matrices," *IEEE Trans. Sig. Process.*, vol. 61, no. 6, pp. 1509-1519, 2013.
- [28] C. Wang, G. Pan, T. Tong, and L. Zhu, "Shrinkage estimation of large dimensional precision matrix using random matrix theory," *Statistica Sinica*, vol. 25, no. 3, pp. 993-1008, 2015.
- [29] T. Ito and T. Kubokawa, "Linear ridge estimator of high-dimensional precision matrix using random matrix theory," *Technical Report F-995, CIRJE*, Faculty of Economics, University of Tokyo, 2015.
- [30] T. Bodnar, A. K. Gupta, and N. Parolya, "Direct shrinkage estimation of large dimensional precision matrix," *J. Multivar. Anal.*, vol. 146, pp. 223-236, 2016.
- [31] X. Mestre and M. A. Lagunas, "Finite sample size effect on minimum variance beamformers: Optimum diagonal loading factor for large arrays," *IEEE Trans. Sig. Process.*, vol. 54, no. 1, pp. 69-82, 2006.
- [32] C.-K. Wen, J.-C. Chen, and P. Ting, "A shrinkage linear minimum mean square error estimator," *IEEE Sig. Process. Lett.*, vol. 20, no. 12, pp.1179-1182, 2013.
- [33] J. Serra and M. Najar, "Asymptotically optimal linear shrinkage of sample LMMSE and MVDR filters," *IEEE Trans. Sig. Process.*, vol. 62, no. 14, pp. 3552-3564, 2014.
- [34] M. Zhang, F. Rubio, D. Palomar, and X. Mestre, "Finite-sample linear filter optimization in wireless communications and financial systems," *IEEE Trans. Sig. Process.*, vol. 61, no. 20, pp. 5014-5025, 2013.
- [35] J. Tong, P. J. Schreier, Q. Guo, S. Tong, J. Xi, and Y. Yu, "Shrinkage of covariance matrices for linear signal estimation using cross-validation," *IEEE Trans. Sig. Process.*, vol. 64, no. 11, pp. 2965-2975, 2016.
- [36] J. Tong, Q. Guo, J. Xi, Y. Yu, and P. J. Schreier, "Choosing the diagonal loading factor for linear signal estimation using cross validation," in *Proc. IEEE ICASSP 2016*, pp. 3956-3959, 2016.
- [37] J. R. Guerci and E. J. Baranoski, "Knowledge-aided adaptive radar at DARPA: an overview," *IEEE Signal Process. Mag.*, vol. 23, no. 1, pp. 41-50, Jan. 2006.
- [38] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statist. Surv.*, vol. 4, pp. 40-79, 2010.
- [39] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215-223, 1979.
- [40] R. D. Nowak, "Optimal signal estimation using cross-validation," *IEEE Sig. Process. Letters*, vol. 4, no. 1, pp. 23-25, 1997.
- [41] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 18071820, March 2010.
- [42] N. Shariati, E. Björnson, M. Bengtsson and M. Debbah, "Low-complexity polynomial channel estimation in large-scale MIMO with arbitrary statistics," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 815-830, Oct. 2014.
- [43] W. Weichselberger, M. Herdin, H. Ozelik, and E. Bonek, "A stochastic MIMO channel model with joint correlation of both link ends," *IEEE Trans. Wireless Commun.*, vol. 5, no. 1, pp. 90-100, 2006.
- [44] J. Fang, X. Li, H. Li and F. Gao, "Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1935-1947, March 2017.
- [45] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
- [46] N. Kim, Y. Lee and H. Park, "Performance analysis of MIMO system with linear MMSE receiver," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4474-4478, Nov. 2008.
- [47] J. Tong, J. Xi, Q. Guo and Y. Yu, "Low-complexity cross-validation design of a linear estimator," *Electronics Letters*, vol. 53, no. 18, pp. 1252-1254, 2017.