2017

Low-complexity cross-validation design of a linear estimator

Jun Tong
University of Wollongong, jtong@uow.edu.au

Jiangtao Xi
University of Wollongong, jiangtao@uow.edu.au

Qinghua Guo
University of Wollongong, qguo@uow.edu.au

Yanguang Yu
University of Wollongong, yanguang@uow.edu.au

Low-complexity cross-validation design of a linear estimator

Abstract

Linear signal estimators have extensive applications. Under the minimum mean squared error (MMSE) criterion, the linear MMSE (LMMSE) estimator is optimal but requires knowledge of the covariance matrices. The sample matched filter generally performs worse but requires less a priori knowledge. A composite estimator that combines the sample LMMSE estimator and matched filter is studied, which may lead to noticeable improvements in performance. It is shown that such a gain can be achieved by low-complexity parameter tuning methods based on cross-validation using training or out-oftraining data. Numerical results are provided to demonstrate the effectiveness of the proposed approaches.

Disciplines

Engineering | Science and Technology Studies

Publication Details

J. Tong, J. Xi, Q. Guo & Y. Yu, "Low-complexity cross-validation design of a linear estimator," Electronics Letters, vol. 53, (18) pp. 1252-1254, 2017.

Low-complexity cross-validation design of a linear estimator

J. Tong, J. Xi, Q. Guo, and Y. Yu

Linear signal estimators have extensive applications. Under the minimum mean squared error (MMSE) criterion, the linear MMSE (LMMSE) estimator is optimal but requires knowledge of the covariance matrices. The sample matched filter generally performs worse but requires less a priori knowledge. In this letter, we study a composite estimator that combines the sample LMMSE estimator and matched filter, which may lead to noticeable improvements in performance. We show that such a gain can be achieved by low-complexity parameter tuning methods based on cross-validation using training or out-of-training data. Numerical results are provided to demonstrate the effectiveness of our proposed approaches.

Introduction: Consider the problem of linearly estimating an unknown, zero-mean, unit-variance signal x from an observation $\mathbf{y} \in \mathbb{C}^N$ under the minimum mean squared error (MMSE) criterion. The linear MMSE estimator [1] is given by $\widehat{x} = \mathbf{w}^\dagger \mathbf{y}$, where

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{r}, \tag{1}$$

$\mathbf{R} \triangleq \mathrm{E}[\mathbf{y}\mathbf{y}^\dagger]$ denotes the covariance matrix of $\mathbf{y}$, $\mathbf{r} \triangleq \mathrm{E}[\mathbf{y}x^*]$ the cross-covariance of $\mathbf{y}$ and x, and $(\cdot)^\dagger$ conjugate transpose. Note that a scalar signal x is assumed here for simplicity but the techniques discussed in this letter can be directly applied to estimating entries of a vector signal from $\mathbf{y}$.

In practice, $\mathbf{R}$ and $\mathbf{r}$ are unknown and may be estimated from a set of training data. Using the sample covariance matrix (SCM) approach, (1) is approximated by

$$\mathbf{w} = \widehat{\mathbf{R}}^{-1}\widehat{\mathbf{r}}, \tag{2}$$

where the sample covariance matrices are computed as

$$\widehat{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t^\dagger, \qquad \widehat{\mathbf{r}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t x_t^*, \tag{3}$$

T is the number of samples, and $(x_t, \mathbf{y}_t)$ are the t-th pair of training samples. When T is large, such a sample LMMSE estimator may lead to a low mean squared error (MSE) of signal estimation:

$$\mathrm{MSE}_x \triangleq \mathrm{E}_x[|x - \widehat{x}|^2], \tag{4}$$

where $\mathrm{E}_x[\cdot]$ denotes expectation with respect to x. One key challenge is that when the training data are limited, i.e., T is small relative to the dimensionality N of $\mathbf{y}$, the SCM $\widehat{\mathbf{R}}$ can be ill-conditioned or even singular, and the resulting sample estimator may result in a high MSE. In this case, a scaled sample matched filter which does not rely on $\widehat{\mathbf{R}}$ may perform better, as will be illustrated later in Fig. 1.

Furthermore, the estimator below, which is a linear combination of the sample LMMSE estimator and matched filter,

$$\mathbf{w}_{\rho,\tau} = \rho\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{r}} + \tau\widehat{\mathbf{r}} \tag{5}$$

may significantly outperform (2), as shown in [2]. However, the coefficients (ρ, τ) must be carefully chosen to optimize performance. In [2], low-complexity methods for choosing (ρ, τ) for cases with perfect knowledge of $\mathbf{r}$, i.e., $\widehat{\mathbf{r}} = \mathbf{r}$, were derived using the random matrix theory (RMT). Other similar schemes have been discussed in [3]-[8]. In particular, [2]-[4] assume $\widehat{\mathbf{r}} = \mathbf{r}$, while the schemes of [5]-[8] rely on a grid search to tune the parameters, which incurs higher complexities.

In this letter, we derive cross-validation (CV) schemes to choose (ρ, τ) for (5) using training and out-of-training data, respectively. The schemes do not assume perfect knowledge of $\mathbf{r}$ and avoid the grid search required in [5]-[8]. When training samples of both $\mathbf{y}$ and x are available, we introduce a CV scheme for choosing (ρ, τ) that directly minimizes the signal estimation error. For cases where x is unobservable, we propose a scheme that optimizes a performance proxy based on a prediction analysis of the out-of-training data. Our proposed methods give analytical solutions to (ρ, τ) and have low implementation complexities. Their effectiveness is illustrated using numerical examples.

CV based on training data: The MSE of estimating x using (5) can be computed as

$$\mathrm{MSE}_x = 1 - \mathbf{r}^\dagger\mathbf{R}^{-1}\mathbf{r} + (\mathbf{w}_{\rho,\tau} - \mathbf{w})^\dagger\mathbf{R}(\mathbf{w}_{\rho,\tau} - \mathbf{w}). \tag{6}$$

It can be shown that the parameters for (5) that minimize the above MSE, referred to as the oracle parameter choice, are given by

$$\begin{bmatrix}\rho^\star\\ \tau^\star\end{bmatrix} = \begin{bmatrix}\widehat{\mathbf{r}}^\dagger\widehat{\mathbf{R}}^{-1}\mathbf{R}\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{r}} & \widehat{\mathbf{r}}^\dagger\widehat{\mathbf{R}}^{-1}\mathbf{R}\widehat{\mathbf{r}}\\ \widehat{\mathbf{r}}^\dagger\mathbf{R}\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{r}} & \widehat{\mathbf{r}}^\dagger\mathbf{R}\widehat{\mathbf{r}}\end{bmatrix}^{-1}\begin{bmatrix}\widehat{\mathbf{r}}^\dagger\widehat{\mathbf{R}}^{-1}\mathbf{r}\\ \widehat{\mathbf{r}}^\dagger\mathbf{r}\end{bmatrix}. \tag{7}$$

Unfortunately, such oracle parameters can not be obtained in practice because $\mathbf{R}$ and $\mathbf{r}$ are unknown. We now derive a practical scheme based on leave-one-out cross-validation (LOOCV) to approximate the oracle parameters. Similarly to [8], the length-T training block $(\mathbf{x}, \mathbf{Y})$ consisting of $\{x_t, \mathbf{y}_t\}$ is repeatedly split into two sets with respect to time. For the t-th split, $T-1$ pairs of training symbols are used for covariance matrix estimation and the remaining one pair $(x_t, \mathbf{y}_t)$ is reserved for parameter validation. In total, T different splits are obtained and the parameter that minimizes the average squared error of estimating x_t is chosen. Specifically, for the t-th split, the estimator is constructed using SCM

$$\widehat{\mathbf{R}}_t = \frac{1}{T-1}\sum_{i\neq t}\mathbf{y}_i\mathbf{y}_i^\dagger, \quad \widehat{\mathbf{r}}_t = \frac{1}{T-1}\sum_{i\neq t}\mathbf{y}_i x_i^* \tag{8}$$

as

$$\widehat{\mathbf{w}}_t = \rho\widehat{\mathbf{R}}_t^{-1}\widehat{\mathbf{r}}_t + \tau\widehat{\mathbf{r}}_t. \tag{9}$$

The error of predicting x_t using $\widehat{\mathbf{w}}_t^\dagger\mathbf{y}_t$ is found as

$$\zeta_t \triangleq x_t - \widehat{\mathbf{w}}_t^\dagger\mathbf{y}_t. \tag{10}$$

We use this error to estimate the MSE of signal estimation and choose parameters (ρ, τ) to minimize the cost function

$$J(\rho, \tau) = \frac{1}{T}\sum_{t=1}^{T}|\zeta_t|^2. \tag{11}$$

We can show that the above cost function can be rewritten as

$$J(\rho, \tau) = \|\mathbf{x} - \rho\mathbf{x}\mathbf{P} - \tau\mathbf{x}\mathbf{Q}\|_F^2\,, \tag{12}$$

where $\|\cdot\|_F$ denotes Frobenius norm,

$$\mathbf{P} \triangleq (\mathbf{B} - \mathbf{D}_\mathbf{B})(\mathbf{I} - \mathbf{D}_\mathbf{B})^{-1}, \tag{13}$$

$$\mathbf{B} \triangleq \frac{1}{T}\mathbf{Y}^\dagger\widehat{\mathbf{R}}^{-1}\mathbf{Y}, \tag{14}$$

$$\mathbf{Q} \triangleq \mathbf{C} - \mathbf{D}_\mathbf{C}, \tag{15}$$

$$\mathbf{C} \triangleq \frac{1}{T-1}\mathbf{Y}^\dagger\mathbf{Y}, \tag{16}$$

and $\mathbf{D}_\mathbf{X}$ denotes the diagonal matrix that shares the diagonal entries of $\mathbf{X}$. The optimal ρ and τ are finally analytically calculated by

$$\begin{bmatrix}\rho^\star\\ \tau^\star\end{bmatrix} = \begin{bmatrix}||\mathbf{x}\mathbf{P}||_F^2 & \mathbf{x}\mathbf{P}\mathbf{Q}\mathbf{x}^\dagger\\ \mathbf{x}\mathbf{Q}\mathbf{P}^\dagger\mathbf{x}^\dagger & ||\mathbf{x}\mathbf{Q}||_F^2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}\mathbf{P}\mathbf{x}^\dagger\\ \mathbf{x}\mathbf{Q}\mathbf{x}^\dagger\end{bmatrix}. \tag{17}$$

Note that this parameter choice is fully data-driven as the right-hand side of (17) depends only on the training data $(\mathbf{x}, \mathbf{Y})$. It has a low complexity as no grid search is required. We note that [2] provides a RMT-based analytical solution to choose the composite filter (5) for the case $\widehat{\mathbf{r}} = \mathbf{r}$. Here, an alternative approach based on CV is designed, which is applicable when $\widehat{\mathbf{r}}$ is estimated from SCM.

As a byproduct of (17), we can obtain the optimal choices of ρ and τ for special cases with $\tau = 0$ and $\rho = 0$, respectively. This leads to the following scaled sample LMMSE estimator

$$\mathbf{w} = \frac{\mathbf{x}\mathbf{P}\mathbf{x}^\dagger}{||\mathbf{x}\mathbf{P}||_F^2}\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{r}}, \tag{18}$$

and scaled macthed filter

$$\mathbf{w} = \frac{\mathbf{x}\mathbf{Q}\mathbf{x}^\dagger}{||\mathbf{x}\mathbf{Q}||_F^2}\widehat{\mathbf{r}}, \tag{19}$$

which improve their original form $\mathbf{w} = \widehat{\mathbf{R}}^{-1}\widehat{\mathbf{r}}$ and $\mathbf{w} = \widehat{\mathbf{r}}$, respectively.

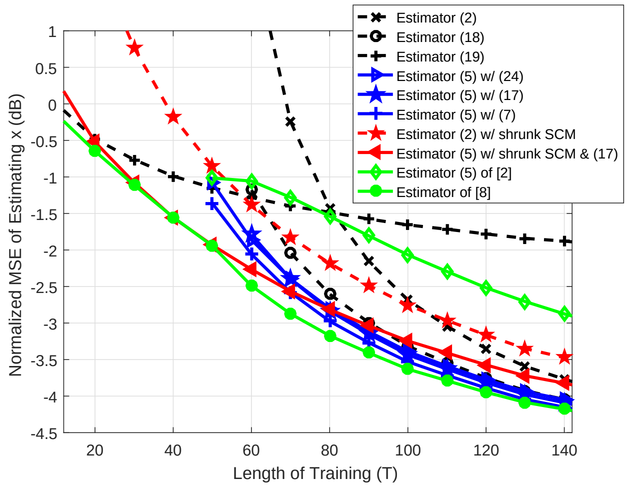

Fig. 1 *Performance comparison of different estimator designs. The LMMSE estimator (1) with perfect knowledge of* $(\mathbf{R}, \mathbf{r})$ *achieves a normalized MSE of about* -5.7 *dB.*

CV based on out-of-training data: We now propose an alternative method to choose (ρ, τ) based on a prediction analysis of a set of out-of-training samples $\boldsymbol{\Upsilon}$ of length D. Let $\mathbf{y}$ be a column of $\boldsymbol{\Upsilon}$, y_n its n-th entry, and $\mathbf{y}_{\sim n}$ the vector obtained by excluding y_n from $\mathbf{y}$. The sample covaraince matrix $\widehat{\mathbf{R}}_{\sim n}$ of $\mathbf{y}_{\sim n}$ is a submatrix of $\widehat{\mathbf{R}}$ with its n-th row and n-th column both excluded. Similarly to [8], we choose (ρ, τ) to minimize the average squared error of predicting y_n from $\mathbf{y}_{\sim n}$

$$J'(\rho, \tau) = \frac{1}{ND} \sum_{\{\mathbf{y}\}} \sum_{n=1}^{N} |y_n - \widehat{y}_{n,\rho,\tau}|^2 , \tag{20}$$

where the first summation is over the D columns $\{\mathbf{y}\}$ of $\boldsymbol{\Upsilon}$. The prediction of y_n from $\mathbf{y}_{\sim n}$ follows the LMMSE principle, replacing the true inverse covariance matrix of $\mathbf{y}_{\sim n}$ as $\rho \widehat{\mathbf{R}}_{\sim n}^{-1} + \tau \mathbf{I}$ and the cross-covariance of $\mathbf{y}_{\sim n}$ and y_n as the n-th column of $\widehat{\mathbf{R}}$ with its n-th entry excluded. After some manipulations using a block matrix version of the Woodbury matrix identity that relates $\widehat{\mathbf{R}}_{\sim n}^{-1}$ and $\widehat{\mathbf{R}}^{-1}$, we can show that

$$J'(\rho, \tau) = \frac{1}{ND} ||\boldsymbol{\Upsilon} - \rho \mathbf{A} \boldsymbol{\Upsilon} - \tau \mathbf{S} \boldsymbol{\Upsilon}||_F^2 , \tag{21}$$

where

$$\mathbf{A} = \mathbf{I} - \left(\mathbf{D}_{\widehat{\mathbf{R}}^{-1}}\right)^{-1} \widehat{\mathbf{R}}^{-1}, \tag{22}$$

$$\mathbf{S} = \widehat{\mathbf{R}} - \mathbf{D}_{\widehat{\mathbf{R}}}. \tag{23}$$

The optimal (ρ, τ) that minimize (21) are found as

$$\begin{bmatrix} \rho^\star \\ \tau^\star \end{bmatrix} = \begin{bmatrix} \mathrm{tr}(\mathbf{ATA}^\dagger) & \mathcal{R}(\mathrm{tr}(\mathbf{ATS}^\dagger)) \\ \mathcal{R}(\mathrm{tr}(\mathbf{STA}^\dagger)) & \mathrm{tr}(\mathbf{STS}^\dagger) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{tr}(\mathbf{AT}) \\ \mathrm{tr}(\mathbf{ST}) \end{bmatrix} , \tag{24}$$

where $\mathcal{R}(\cdot)$ denotes the real part of a complex number and $\mathbf{T} = \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\dagger$. Note that (24) can be computed when the training data $\mathbf{x}$ is not observable and the cost function of (12) can not be computed.

Generalization: The discussion above assumes that the SCM $\widehat{\mathbf{R}}$ is positive-definite, i.e., the length T of training data is not smaller than the dimensionality N of $\mathbf{y}$. For singular or ill-conditioned SCM $\widehat{\mathbf{R}}$, we may replace $\widehat{\mathbf{R}}$ by a shrunk version $\beta_1 \widehat{\mathbf{R}} + \beta_2 \mathbf{I}$ which is better conditioned than $\widehat{\mathbf{R}}$ if (β_1, β_2) are properly chosen. The closed-form solutions to (β_1, β_2) from [9]-[11], which do not involve grid search, can be used. Then (ρ, τ) are found by directly plugging the shrunk estimator into (14), (22) and (23), respectively, when the training and out-of-training data are used for tuning the parameters.

Numerical Results: We now show an example of the above methods, with $N = 50$

$$\mathbf{y} = \mathbf{h}x + \mathbf{S}\boldsymbol{\xi} + \mathbf{z}, \tag{25}$$

where $\mathbf{h} \in \mathbb{C}^{50\times 1}$ and $\mathbf{S} \in \mathbb{C}^{50\times 49}$ are fixed for each Monte-Carlo trial, x and $\boldsymbol{\xi}$ contain i.i.d. zero-mean Gaussian variables with unit variance, and $\mathbf{z}$ is additive white Gaussian noise (AWGN) with variance 0.1. We also assume that $\mathbf{h}$ and $\mathbf{S}$ consist of i.i.d. Gaussian variables with zero mean and unit variance. For the parameter choice using out-of-training data, we set $D = 10$. For the case where the SCM is shrunk, the Ledoit-Wolf method [9] is applied. The MSE normalized by the average power of x is shown for different estimators in Fig. 1. It is seen that our proposed CV choices of parameters (ρ, τ) for (5) significantly improve the performance compared to the sample LMMSE estimator (2) and can approach the oracle performance achieved with (7). The method of [2], which is based on random matrix theory and can perform very well for perfect knowledge of $\mathbf{r}$, does not work well here with $\mathbf{r}$ directly replaced by $\widehat{\mathbf{r}}$.

Applying the low-complexity shrinkage covariance matrix estimator of [9] allows our methods to be used for low sample support with $T < N = 50$. In this case, the performance achieved by the estimator (5) with our proposed parameter choices is slightly worse than the signal estimator of [8], which has a different form of $\mathbf{w}_{\alpha,\beta} = \alpha(\widehat{\mathbf{R}} + \beta \mathbf{I})^{-1}\widehat{\mathbf{r}}$. Note, however, that the estimator of [8] requires a line search of the parameter β based on a CV criterion and exhibits a significantly higher complexity. Thus the approaches proposed in this letter may be attractive for applications that require a low complexity.

Conclusions: In this letter, we studied the choice of parameters for a composite filter using cross-validation. The proposed designs can enhance the performance of the sample LMMSE estimator for applications with low sample support and approach the optimal performance of the composite structure studied. They also exhibit a low complexity as no line search of the parameter is needed and may be attractive for applications requiring low complexity.

Acknowledgment: This work was supported in part by an International Links grant of University of Wollongong (UOW), Australia.

J. Tong, J. Xi, Q. Guo and Y. Yu (*School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia.*)

Email: jtong@uow.edu.au

References

1 Scharf, L.L.: *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison–Wesley, Bosten, 1991.

2 Serra, J., and Nájar, M.: 'Asymptotically optimal linear shrinkage of sample LMMSE and MVDR filters,' *IEEE Trans. Sig. Process.*, 2014, 62, (14), pp. 3552-3564.

3 Serra, J., and Rubio, F.: 'Bias corrections in linear mmse estimation with large filters,' in *Proc. Eur. Sig. Process. Conf. (EUSIPCO 2010)*, 2010, pp. 23-27.

4 Serra, J., and Nájar, M.: 'Double shrinkage correction in sample LMMSE estimation,' in *Proc. IEEE Sig. Process. Conf.*, 2013.

5 Mestre, X., and Lagunas, M. A., 'Finite sample size effect on minimum variance beamformers: Optimum diagonal loading factor for large arrays,' *IEEE Trans. Sig. Process.*, 2006, 54, (1), pp. 69-82.

6 Wen, C.-K., Chen J.-C., and Ting, P.: 'A shrinkage linear minimum mean square error estimator,' *IEEE Sig. Process. Lett.*, 2013, 20, (12), pp.1179-1182.

7 Zhang, M., Rubio, F., Palomar, D., and Mestre. X, 'Finite-sample linear filter optimization in wireless communications and financial systems,' *IEEE Trans. Sig. Process.*, 2013, 61, (20), pp. 5014-5025.

8 Tong, J., Schreier, P. J., Guo, Q., Tong, S., Xi, J., and Yu, Y.: 'Shrinkage of covariance matrices for linear signal estimation using cross-validation,' *IEEE Trans. Sig. Process.*, 2016, 64, (11), pp. 2965-2975.

9 Ledoit, O., and Wolf, M.: 'A well-conditioned estimator for large-dimensional covariance matrices,' *J. Multivar. Anal.*, 2014, 88, (2), pp. 365-411.

10 Stoica, P., Li, J., Zhu, X., and Guerci, J. R.: 'On using a priori knowledge in space-time adaptive processing,' *IEEE Trans. Sig. Process.*, 2008, 56, (6), pp. 2598-2602.

11 Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O., 'Shrinkage algorithms for MMSE covariance estimation,' *IEEE Trans. Sig. Process.*, 2010, 58, (10), pp. 5016-5029.