Steven Kay and Yonina C. Eldar

Rethinking Biased Estimation

In this lecture note we discuss methods to improve the accuracy of unbiased estimators used in many signal processing problems. Our approach is based on introducing a bias as a means of reducing the mean-squared error (MSE). The important aspect of our framework is that the reduction in MSE is guaranteed for all values of the unknown parameter.

It is customary in signal processing to seek unbiased estimators that perform well. This is typically accomplished by determining the minimum variance unbiased (MVU) estimator, using the theory of sufficient statistics or the attainment of the Cramér-Rao lower bound [1]. A more desirable estimator, however, is one that minimizes the MSE, which is a direct measure of estimation error. In these notes we revisit the problem of determining a minimum MSE (MMSE) estimator for a parameter that is deterministic but unknown. We indicate how biased estimators can be found that outperform the MVU estimator in terms of MSE.

RELEVANCE

Biased estimation is already a mainstream approach [2]–[6]. The importance of this subject is that estimators can be derived that outperform existing approaches especially for short data records and/or low signal-to-noise ratios (SNRs). Applications include the design of estimation algorithms for sonar, radar, and communications as well as a myriad of other disciplines that rely heavily on precise measurement of parameters.

Courses that may benefit from this lecture note include statistical signal processing, digital communications, information theory, and modern control theory.

PREREQUISITES

The reader is assumed to be familiar with basic classical estimation theory as it is presented in [1].

PROBLEM STATEMENT

The determination of an MVU estimator of a deterministic scalar parameter θ is a pervasive goal in signal processing applications. However, in some cases an unbiased estimator may not exist [1], or the unbiasedness requirement can produce nonsensical results [7]. But perhaps the most important objection to the constraint of unbiasedness is that it produces estimators $\hat{\theta}$ whose optimality is based on the difference between $\hat{\theta}$ and its *average value*, not $\hat{\theta}$ and the *true* value θ, as measured by the MSE. It is the latter that is actually of prime importance in an estimation problem. We next indicate how one can determine and implement estimators of deterministic parameters with smaller MSE than the MVU estimator for all values of θ. The key to be able to reduce the MSE is by scaling an unbiased estimator by a number between zero and one, producing a so-called "shrinkage estimator."

Suppose we wish to estimate the value of a deterministic scalar parameter θ based on the available data $\{x_1, x_2, \ldots, x_N\}$. It is assumed that the MVU estimator $\hat{\theta}_u$ and its variance $\text{var}(\hat{\theta}_u) = \text{MSE}(\hat{\theta}_u)$ are known. To reduce the MSE, we will bias the MVU estimator by scaling it towards zero [3], [8]. Specifically, the biased estimator is given by

$$\hat{\theta}_b = (1+m)\hat{\theta}_u\,, \tag{1}$$

where m will be chosen to minimize the MSE $E[(\hat{\theta}_b - \theta)^2]$. The MSE can be written as the sum of the variance and the squared bias by

$$\begin{aligned}\text{MSE}(\hat{\theta}_b) &= E[(\hat{\theta}_b - \theta)^2] \\ &= \text{var}(\hat{\theta}_b) + (E[\hat{\theta}_b] - \theta)^2\,.\end{aligned} \tag{2}$$

Using the fact that $E[\hat{\theta}_u] = \theta$, the MSE of $\hat{\theta}_b$ becomes

$$\text{MSE}(\hat{\theta}_b) = (1+m)^2\text{var}(\hat{\theta}_u) + m^2\theta^2\,. \tag{3}$$

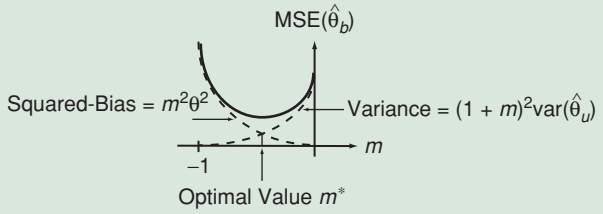
Our goal is to choose m so that $\text{MSE}(\hat{\theta}_b)$ is less than the MSE of the original unbiased estimator $\hat{\theta}_u$, which is its variance $\text{var}(\hat{\theta}_u)$, *for all values of* θ. (For clarity, we have begun our discussion with the scalar case; the results are generalized to multiple parameters later in this column.)

CAN A SCALING FACTOR BE FOUND?

To reduce the MSE we see immediately from (3) that the variance component must be reduced more than the squared-bias component is increased. The MSE is plotted in Figure 1 versus m over the range $-1 \leq m < 0$ (for which the variance is decreased). Evidently, there is a value of m, which we denote by m^*, that minimizes the overall MSE, trading off an increase in bias for a decrease in variance. The key issue is whether the optimal $m = m^*$ depends on the unknown value of θ. If it does, then the biased estimator $\hat{\theta}_b$ will not be realizable and cannot be implemented. However, even in this case it is still sometimes possible to find an m for which the MSE can be reduced, as we will see shortly. We next examine how to determine m^* to see if it is dependent on θ.

Digital Object Identifier 10.1109/MSP.2008.918027

1053-5888/08/$25.00©2008IEEE



[FIG1] Trading off bias for variance in reduction of MSE. The biased estimator is $\hat{\theta}_b = (1+m)\hat{\theta}_u$, which is a scaled version of the unbiased MVU estimator.

FINDING THE OPTIMAL SCALING FACTOR

The value of m that minimizes the MSE given by (3) is easily found. Differentiating the MSE, which is quadratic in m, with respect to m and setting the result equal to zero produces the optimal value as

$$m^* = -\frac{\text{var}(\hat{\theta}_u)}{\text{var}(\hat{\theta}_u)+\theta^2} = -\frac{1}{1+\theta^2/\text{var}(\hat{\theta}_u)} \quad (4)$$

which unfortunately appears to depend on θ. However, if

$$\rho = \frac{\theta^2}{\text{var}(\hat{\theta}_u)} \quad (5)$$

is independent of θ, then so is m^*. The latter occurs when $\text{var}(\hat{\theta}_u)$ is proportional to θ^2.

EXAMPLE 1—SCALE PARAMETER FOR EXPONENTIAL PDF

Assume that we have N independent and identically distributed (IID) observations of an exponential random variable with probability density function (PDF) $p_X(x) = (1/\theta)\exp(-x/\theta)$ for $x \geq 0$ and 0 otherwise, where $\theta > 0$. It is well known that the MVU estimator of θ is the sample mean $\hat{\theta}_u = \bar{x} = (1/N)\sum_{i=1}^{N} x_i$ with a variance of $\text{var}(\hat{\theta}_u) = \theta^2/N$. Clearly, the variance of the unbiased estimator is proportional to θ^2. As a result, the MMSE estimator can be realized. From (4) we have with $m^* = -1/(1+N)$

$$\begin{aligned}\hat{\theta}_b &= (1+m^*)\hat{\theta}_u \\ &= \left(1-\frac{1}{1+N}\right)\hat{\theta}_u \\ &= \frac{N}{N+1}\bar{x} \\ &= \frac{1}{N+1}\sum_{i=1}^{N} x_i\,.\end{aligned}$$

The MSE of $\hat{\theta}_b$ is given by (3) and is equal to

$$\begin{aligned}\text{MSE}(\hat{\theta}_b) &= \left(\frac{N}{N+1}\right)^2\frac{\theta^2}{N} + \frac{1}{(N+1)^2}\theta^2 \\ &= \frac{\theta^2}{N+1} \\ &< \frac{\theta^2}{N} \\ &= \text{var}(\hat{\theta}_u) \\ &= \text{MSE}(\hat{\theta}_u)\,.\end{aligned}$$

The greatest reduction in MSE occurs for short data record lengths.

We next examine two cases of increasing complexity in which the variance is *not* proportional to θ^2. First $\text{var}(\hat{\theta}_u)$ is assumed to be constant with θ and then the more general problem in which $\text{var}(\hat{\theta}_u)$ may depend on θ in an arbitrary manner is described.

FINDING A SCALING FACTOR—VARIANCE OF MVU ESTIMATOR IS CONSTANT

We now assume that $\text{var}(\hat{\theta}_u) = V$, a constant. Then $\rho = \theta^2/V$ is dependent on θ and hence, $\hat{\theta}_b$ cannot be implemented. Nonetheless it may still be possible to find a biased estimator that achieves a lower MSE than the MVU estimator. To do so we require that θ be restricted to a given range of values. In particular, suppose that $|\theta| \leq \theta_0$ for some $\theta_0 > 0$, a reasonable assumption for many practical problems. Then from (3) we have that $\text{MSE}(\hat{\theta}_b) = (1+m)^2V + m^2\theta^2$, which is shown in Figure 2 as a function of θ, along with $\text{MSE}(\hat{\theta}_u)$. It is clear that if we can find an m so that $\text{MSE}(\hat{\theta}_b)$ for $\theta = \theta_0$ is less than $\text{MSE}(\hat{\theta}_u)$, then the MSE of the biased estimator will be less than that of the unbiased estimated *for all* $|\theta| \leq \theta_0$. Hence, we would like to choose an m so that

$$\text{MSE}(\hat{\theta}_b) = (1+m)^2V + m^2\theta_0^2 < V \quad (6)$$

for all $|\theta| \leq \theta_0$ or equivalently

$$\max_{|\theta|\leq\theta_0}\{\text{MSE}(\hat{\theta}_b) - \text{MSE}(\hat{\theta}_u)\} = (1+m)^2V + m^2\theta_0^2 - V < 0\,. \quad (7)$$

This produces the allowable range of m to be

$$1+m > \frac{\theta_0^2 - V}{\theta_0^2 + V}\,. \quad (8)$$

Any estimator of the form (1) with m satisfying (8) will have lower MSE than the unbiased estimator for all $|\theta| \leq \theta_0$. As our goal is to reduce the MSE as much as possible, we choose m to minimize $(1+m)^2V + m^2\theta_0^2$. Hence, again differentiating (6) and setting it equal to zero produces

$$1+m^* = \frac{\theta_0^2}{\theta_0^2+V}\,, \quad (9)$$

which is easily seen to satisfy (8). Thus, the biased estimator that *minimizes the maximum MSE* over $|\theta| \leq \theta_0$ is

$$\hat{\theta}_b = (1+m^*)\hat{\theta}_u = \frac{\theta_0^2}{\theta_0^2+V}\bar{x}\,. \quad (10)$$

Interestingly, as $\theta_0 \to \infty$, the unbiased and biased estimators coincide. Using (3) and (9) the resulting minimum MSE can be shown to be

$$\mathrm{MSE}(\hat{\theta}_b) = V\left[\frac{\theta_0^4 + \theta^2 V}{(\theta_0^2 + V)^2}\right]. \qquad (11)$$

For $|\theta| \le \theta_0$ the term in brackets is less than or equal to $\theta_0^2/(\theta_0^2 + V)$. Thus, a sizable reduction in the MSE results if $\theta_0^2/V \ll 1$.

EXAMPLE 2—MEAN OF GAUSSIAN PDF

As a specific example, suppose we have N IID observations of X, which have the Gaussian PDF $p_X(x) = (1/\sqrt{2\pi\sigma^2})\exp[-(x-\theta)^2/(2\sigma^2)]$. Our goal is to estimate the mean θ. The MVU estimator of θ is the sample mean $\hat{\theta}_u = \bar{x}$, whose variance is the constant $V = \sigma^2/N$. Our previous results therefore apply and the MSE can be reduced by using a biased estimator. From (10) it is

$$\hat{\theta}_b = (1 + m^*)\hat{\theta}_u = \frac{\theta_0^2}{\theta_0^2 + \sigma^2/N}\,\bar{x}. \qquad (12)$$

The condition for a sizable reduction in MSE becomes $\theta_0^2/(\sigma^2/N) \ll 1$, which is equivalent to a short data record and/or low SNR.

FINDING A SCALING FACTOR—VARIANCE OF MVU ESTIMATOR IS DEPENDENT ON θ

The preceding section illustrates that even when the optimal m of (4) depends on θ, we may still be able to reduce the MSE *uniformly over all allowable* θ by employing a *minimax* strategy. In essence, we chose an m that minimized the MSE for a given value of the parameter $\theta = \theta_0$ and argued that the MSE would also be reduced for all other possible values of θ. The value of θ chosen was the one that *maximized* the MSE, i.e., a worst case. Because of the subsequent minimization over m this is a *minimax* approach. To state these results more generally, suppose we have an MVU estimator $\hat{\theta}_u$ of θ with $\mathrm{MSE}(\hat{\theta}_u) = \mathrm{var}(\hat{\theta}_u)$, which now may depend on θ in a general fashion. To reduce the MSE of $\hat{\theta}_u$, we consider biased estimators of the form (1), where we choose m so that $\mathrm{MSE}(\hat{\theta}_b) < \mathrm{MSE}(\hat{\theta}_u)$ for all θ, and such that the reduction in MSE is as large as possible. Mathematically these goals can be achieved by selecting m^* to minimize the largest (negative) difference between the two approaches [8] as

$$m^* = \arg\min_m \max_\theta \{\mathrm{MSE}(\hat{\theta}_b) - \mathrm{MSE}(\hat{\theta}_u)\}. \qquad (13)$$

The solution of (13) may be obtained by using any one of the many known iterative algorithms for minimax problems. An important observation is that even in the absence of constraints on θ, a biased estimator can yield reduced MSE over an unbiased approach. If we have prior deterministic knowledge on θ of the form $\theta \in \mathcal{U}$, where $\mathcal{U}$ is a given constraint set, then we can readily incorporate it into our framework by restricting the inner maximization in (13) over the corresponding set. An example follows with others contained in [8].

EXAMPLE 3: PARAMETER IN THE MVU ESTIMATOR VARIANCE

Assume that $-\infty < \theta < \infty$ and $\mathrm{var}(\hat{\theta}_u) = a + b\theta^2$ for some $a, b > 0$. The variance is not proportional to θ^2, which negates the approach in Example 1. Also, $\mathrm{var}(\hat{\theta}_u)$ depends on θ in such a way that it is unbounded (no parameter constraints here) so that the approach in Example 2 does not apply. Using convex analysis tools, however, (13) can be solved explicitly resulting in [8]

$$m^* = \max\left(-\frac{2b}{b+1}, -1\right).$$

The corresponding biased estimator is

$$\hat{\theta}_b = \begin{cases} \frac{1-b}{1+b}\hat{\theta}_u, & b < 1 \\ 0, & b \ge 1. \end{cases}$$

It can be shown that $\hat{\theta}_b$ has lower MSE than $\hat{\theta}_u$ for all values of $-\infty < \theta < \infty$.

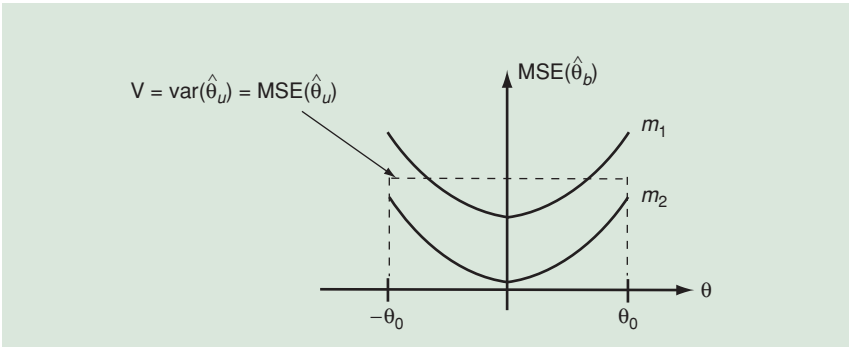
EXTENSION TO MULTIPLE PARAMETERS

All of the results presented so far extend naturally to estimation of a vector parameter $\boldsymbol{\theta}$. Assuming an MVU estimator $\hat{\boldsymbol{\theta}}_u$ exists, we can follow the same approach presented in the previous sections and seek a biased estimator $\hat{\boldsymbol{\theta}}_b = (\mathbf{I} + \mathbf{M})\hat{\boldsymbol{\theta}}_u$ whose total MSE, given by $E[\|\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}\|^2]$, is smaller than that of $\hat{\boldsymbol{\theta}}_u$. To design an appropriate *matrix* $\mathbf{M}$ we solve the vector equivalent of (13)

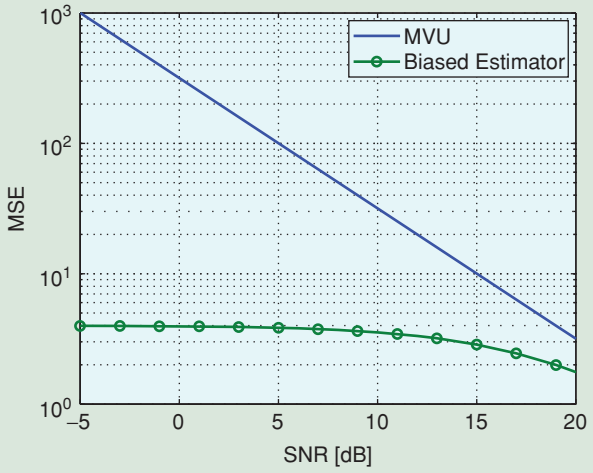
$$\mathbf{M}^* = \arg\min_{\mathbf{M}} \max_{\theta\in\mathcal{U}} \{\mathrm{MSE}(\hat{\boldsymbol{\theta}}_b) - \mathrm{MSE}(\hat{\boldsymbol{\theta}}_u)\}, \qquad (14)$$

where $\mathcal{U}$ is a possible constraint set on $\boldsymbol{\theta}$, although there need not be any constraints, and

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\theta}}_b) &= E[\|\hat{\boldsymbol{\theta}}_b) - \boldsymbol{\theta}\|^2] \\ &= \mathrm{Tr}((\mathbf{I}+\mathbf{M})\mathbf{C}_{\hat{\theta}_u}(\mathbf{I}+\mathbf{M})^T) \\ &\quad + \boldsymbol{\theta}^T\mathbf{M}^T\mathbf{M}\boldsymbol{\theta} \qquad (15)\end{aligned}$$



[FIG2] Choose $\theta = \theta_0$ to guarantee that the MSE will be reduced for all other values of θ. A possible value of m is m_2 but not m_1 since for the latter the MSE is not uniformly reduced.



[FIG3] **MSE in estimating θ in a linear Gaussian model as a function of the SNR using the MVU estimator (16) and the biased estimator (17).**

with $\mathbf{C}_{\hat{\theta}_u} = E[(\hat{\boldsymbol{\theta}}_u - E[\hat{\boldsymbol{\theta}}_u])(\hat{\boldsymbol{\theta}}_u - E[\hat{\boldsymbol{\theta}}_u])^T]$ denoting the covariance matrix of $\hat{\boldsymbol{\theta}}_u$, and $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_u) = \mathrm{var}(\hat{\boldsymbol{\theta}}_u) = \mathrm{Tr}(\mathbf{C}_{\hat{\theta}_u})$ being the sum of the variances of the elements of $\hat{\boldsymbol{\theta}}_u$. Note that the first term of (15) is the variance term and the second is the squared-bias term.

EXAMPLE 4—VECTOR AMPLITUDE PARAMETER OF LINEAR GAUSSIAN MODEL

Consider the linear Gaussian model in which we seek to estimate a $p \times 1$ parameter vector $\boldsymbol{\theta}$ based on an $N \times 1$ observation vector $\mathbf{x}$, which are related through the linear model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}.$$

Here $\mathbf{H}$ is a known $N \times p$ model matrix with $N > p$ and full column rank, and $\mathbf{w}$ is a zero-mean Gaussian random vector with covariance matrix $\mathbf{C} = \sigma^2\mathbf{I}$. This is an extension of Example 2 to the vector setting. The MVU estimator is given by the well known least-squares solution [1]

$$\hat{\boldsymbol{\theta}}_u = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \qquad (16)$$

with covariance matrix $\mathbf{C}_{\hat{\theta}_u} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}$.

To reduce the MSE of $\hat{\boldsymbol{\theta}}_u$ suppose that $\boldsymbol{\theta}$ is restricted to lie in a sphere of the form $\mathcal{U} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|^2 \le \theta_0^2\}$ for some known positive scalar $\theta_0^2 > 0$. Solving (14) with the MSE given by (15) yields the biased estimator [9]

$$\hat{\boldsymbol{\theta}}_b = \frac{\theta_0^2}{\theta_0^2 + \mathrm{Tr}((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1})}\hat{\boldsymbol{\theta}}_u. \qquad (17)$$

Note that if $\mathbf{H} = [1\,1\ldots1]^T$, then (17) reduces to (12). As an example of the improvement afforded by $\hat{\boldsymbol{\theta}}_b$, in Figure 3 we compare its MSE to that of $\hat{\boldsymbol{\theta}}_u$ of (16) as a function of the SNR, defined by $10\log_{10}\|\boldsymbol{\theta}\|^2/\sigma^2$. Here $\boldsymbol{\theta} = [1\,1\ldots1]^T$, $\theta_0^2 = 4$ and $\mathbf{H}^T\mathbf{H}$ was generated as a realization of a random matrix. As can be seen, allowing bias in the estimator improves the performance significantly.

WHAT WE HAVE LEARNED

In this lecture note we have illustrated some of the approaches to determining a biased estimator that exhibits smaller MSE than the "optimal" MVU solution for all feasible values of the unknown parameters. Although we have always assumed that the MVU estimator exists and that it is known along with its variance, there are means to derive good biased estimators that rely on bounds such as the Cramér-Rao lower bound. Hence, a practitioner may well be able to determine good biased estimators when the MVU estimator does not exist or cannot be found. The improvement in performance is greatest for short data records and/or low SNRs. Fortuitously, this is exactly the regime in which most signal processing algorithms must operate.

AUTHORS

Steven Kay (kay@ele.uri.edu) is a professor of electrical engineering at the University of Rhode Island, Kingston. His research interests include spectrum analysis, detection and estimation theory, and probability/random processes. He is a Fellow of the IEEE.

Yonina C. Eldar (yonina@ee.technion.ac.il) is an associate professor in the Department of Electrical Engineering at the Technion, Haifa, Israel. Her research interests are in the areas of statistical signal processing, signal processing and computational biology. She is a Member of the IEEE.

REFERENCES

[1] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall 1993.

[2] B. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerosp. Electron.*, vol. 24, no. 3, pp. 397–401, 1988.

[3] J. Bibby and H. Toutenberg, *Prediction and Improved Estimation in Linear Models*. New York: Wiley, 1977.

[4] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Mathematical Statistics Probability*, vol. 1, 1961, pp. 311–319.

[5] B.J.N. Blight, "Some general results on reduced mean square error estimation," *Amer. Statistician*, vol. 25, no. 3, pp. 24–25, June 1971.

[6] M.D. Perlman, "Reduced mean square error estimation for several parameters," *Sankhya*, vol. 34, series B, part I, pp. 89–92, 1972.

[7] D.R. Cox and D.V. Hinkley, *Theoretical Statistics*. London, U.K.: Chapman & Hall, 1974, p. 253.

[8] Y.C. Eldar, "Uniformly improving the Cramer-Rao bound and maximum likelihood estimation," *IEEE Trans. Signal Processing*, vol. 54, pp. 2943–2956, Aug. 2006.

[9] Y.C. Eldar, A. Ben-Tal, and A. Nemirovski, "Robust mean squared error estimation in the presence of model uncertainties," *IEEE Trans. Signal Processing*, vol. 53, pp. 168–181, Jan. 2005. SP