ADAPTIVE LINEAR ESTIMATORS, USING BIASED CRAMÉR-RAO BOUND

Kamal Shahtalebi and Saeed Gazor

Department of Electrical and Computer Engineering,
Queen's University, Kingston, ON, Canada

ABSTRACT

In this paper, the biased Cramér-Rao Lower Bound (BCRLB) is used to derive the estimate of unknown parameters in a linear model with an arbitrary known additive noise probability density function (PDF).We show that the derived linear estimators (not unique) are linear functions of the observations. Examples are included to illustrate their performances. We show that a biased estimator obtained by optimization of BCRLB is not necessary satisfactory in a general case; therefore, additional considerations must be taken into account. If the Fisher information matrix (FIM) is singular, we use the method of singular value decomposition (SVD) to extract the parameter estimate of linear model. For example we show that in a linear model, parameter estimation based on single observation leads to the normalized least mean square (NLMS) algorithm. In this example using BCRLB optimization, we find the relation between the step size of the NLMS algorithm and bound of bias gradient matrix.

1. INTRODUCTION

The Cramér Rao bound (CRB) which is the inverse of Fisher information matrix (FIM), provides a lower bound on the covariance matrix of any unbiased estimate of a non random parameter [1,2]. Lower bounds on the covariance matrix are also achievable using biased estimators [3–5]. Hero *et al*, developed a lower bound on biased estimators [3, 4]. Following them, Eldar [5] derived bounds on the total variance of any estimate $\hat{W}$ of W with bias gradient matrix whose norm is bounded by a constant. In [3–5], the FIM is treated as a nonsingular matrix.

In this paper, we study the biased estimators of the parameters of a linear model either with nonsingular or singular FIM with the criterion of attaining the lower bound. We show that any estimator attaining the lower bound has two components, the first component is a linear function of observations. The second component is invariant with the parameter. We give some examples to compare the performances of these estimators. All of them have the same total variance and bias gradient matrix. However, it does not mean that all of them are satisfactory estimators. Since, they have different weight estimate errors, we conclude that BCRLB is not always a convenient criteria.

In many applications, the rank r of the signal covariance is often smaller than the number of observations or the filter order. In such applications the FIM is singular. In such cases, there is no unbiased estimator with finite variance, except under unusual conditions [6]. We extend the biased Cramer Rao lower bound (BCRLB) on the total variance when the FIM is singular. Similar to the nonsingular case, we apply the results to a linear regressive model using singular value decomposition (SVD) and derive the optimum biased estimators. In linear models, parameter estimation based on single observation is an example of singular FIM that leads to the well known normalized least mean square (NLMS) algorithm, where its step-size is a function of bias gradient matrix bound.

Accordingly the paper is organized as follows. In section 2 the biased Cramér Rao lower bound (BCRLB) optimization is reviewed. In section 3 the BCRLB estimate of a linear time invariant model is extracted. Assuming an arbitrary additive noise with known PDF, we show that the result is a linear function of observations. Some examples are given to compare the performances of the estimators. Section 4 describes the problem of singular FIM. In this section we find the BCRLB when the FIM is singular. In section 5 we use the new results to extract the estimate of parameter in a linear time invariant system, based on SVD method. The conclusions of the paper are drawn in Section 6.

2. PARAMETER ESTIMATION BASED ON BIASED CRLB (BCRLB)-NONSINGULAR CASE

The Biased CRLB is summarized in the following theorem [1–5].

Theorem 1 *Consider the problem of estimating an unknown deterministic parameter vector $W \in \mathbb{C}^m$ using given measurements $Y \in \mathbb{C}^n$, where the relationship between Y and W is described by the conditional probability density function (PDF) of Y, $f(Y|W)$. We assume that $f(Y|W)$ is twice differentiable with respect to W. Let $\hat{W}$ denote an*

estimation of W with bias:

$$b(W) = E[\hat{W}] - W \tag{1}$$

and the following covariance

$$\boldsymbol{C} = E[(\hat{W} - E[\hat{W}])(\hat{W} - E[\hat{W}])^H], \tag{2}$$

where H stands for complex transpose. Then the covariance matrix $\boldsymbol{C}$ must satisfy [7, 8]:

$$\boldsymbol{C} \geq (\boldsymbol{I} + \boldsymbol{D})\boldsymbol{J}^{-1}(\boldsymbol{I} + \boldsymbol{D})^H, \tag{3}$$

where $\boldsymbol{I}$ is $(m \times m)$ identity matrix[1]*, if the $m \times m$ Fisher information matrix defined by:*

$$\boldsymbol{J} = E\left[\left(\frac{\partial \log f(Y|W)}{\partial W}\right)^H \left(\frac{\partial \log f(Y|W)}{\partial W}\right)\right], \tag{4}$$

is nonsingular and the bias gradient $m \times m$ matrix $\boldsymbol{D}$, is defined by

$$\boldsymbol{D} = \frac{\partial b(W)}{\partial W}. \qquad \blacksquare \tag{5}$$

In general, the above shows that using a biased estimator allows further reduction in the covariance, and biased estimators are proposed, exploiting the tradeoff between the bias and covariance by bounding a norm of the gradient of the bias [3, 4]. Assuming $\mathbf{J}$ is nonsingular, Eldar proposed minimizing the total variance,

$$\mathcal{C} = \text{tr}[(\mathbf{I} + \mathbf{D})\mathbf{J}^{-1}(\mathbf{I} + \mathbf{D})^H], \tag{6}$$

where tr[.] stands for trace, subject to a bounded bias gradient matrix, i.e., tr$[\mathbf{D}^H\mathbf{D}\mathbf{A}] \leq \gamma$, where $\mathbf{A}$ is a non-negative hermitian matrix and $\gamma > 0$ [5]. For simplicity we assume $A = \frac{\gamma}{m\rho^2}\mathbf{I}$, (i.e., tr$[\mathbf{D}^H\mathbf{D}] \leq m\rho^2$ where ρ^2 is a positive constant). The result of this constrained minimization is [5]

$$\mathbf{D} = -\mathbf{I} + \alpha(\mathbf{I} + \alpha\mathbf{J})^{-1}\mathbf{J}, \qquad \alpha \geq 0 \tag{7}$$

and

$$\mathcal{C} = \alpha^2 \text{tr}[(\mathbf{I} + \alpha\mathbf{J})^{-2}\mathbf{J}], \quad \alpha \geq 0. \tag{8}$$

where $\alpha \geq 0$ is the (unique) positive real solution of the following equation

$$m\rho^2 \;=\; \text{tr}\left[(\mathbf{I} + \alpha\mathbf{J})^{-2}\right] = \sum_{k=1}^{m} \frac{1}{(1 + \alpha\lambda_k)^2}, \tag{9}$$

and $\{\lambda_k\}_{k=1}^m$ are the eigenvalues of $\mathbf{J}$. If $\rho \in [0, 1]$, there exists a unique nonnegative solution for equation (9), since $\sum_{k=1}^m \frac{1}{(1+\alpha\lambda_k)^2}$ is a decreasing function of $\alpha \geq 0$ and takes values of m and 0 for $\alpha = 0$ and $\alpha \to \infty$, respectively. Note that, for $\rho = 1$ (or equivalently $\alpha = 0$) the optimum solution is $\mathbf{D} = -\mathbf{I}$ which represents the case in which the covariance is zero while the estimator has considerable gradient bias. The other extreme case is where an unbiased estimator is used; this case is represented by $\rho = 0$ (or equivalently $\alpha \to \infty$).

3. LINEAR TIME INVARIANT MODEL-NONSINGULAR FIM

We use the above results in parameter estimation of a linear model described by[2]

$$y_k = X_k^T W + v_k, \quad k = n_0, n_0 + 1, \cdots, n \tag{10}$$

where T stands for transpose, $W \in R^m$ is unknown vector of parameters, $\{y_k\}_{k=n_0}^n$ are $n - n_0 + 1$ observations, $\{X_k\}_{k=n_0}^n$ are known $m \times 1$ input vectors, and $\{v_k\}_{k=n_0}^n$ is an i.i.d additive noise sequence with PDF $f_{\mathbf{v}}(.)$. Rewriting the model in matrix form, we get,

$$Y_n = \mathbf{X}_n^T W + V_n, \tag{11}$$

where $Y_n = [y_{n_0}, \cdots, y_n]^T$, $\mathbf{X}_n = [X_{n_0}, \cdots, X_n]$, and $V_n = [v_{n_0}, \cdots, v_n]^T$. Given the observation Y_n and input matrix $\mathbf{X}_n$, for the model (11) the FIM $\mathbf{J}_n$ is given by (See Appendix)

$$\mathbf{J}_n = \theta\mathbf{X}_n\mathbf{X}_n^T, \tag{12}$$

where

$$\theta = E\left[\left(\frac{f_{\mathbf{v}}^{'}(v)}{f_{\mathbf{v}}(v)}\right)^2\right] = \int_{-\infty}^{\infty} \frac{{f_{\mathbf{v}}^{'}}^2(x)}{f_{\mathbf{v}}(x)}dx, \tag{13}$$

and $f_{\mathbf{v}}^{'}(v) = \frac{d}{dv}f_{\mathbf{v}}(v)$. From (7), (8) and (12), for a given α_n the bias gradient matrix $\mathbf{D}_n$ and the constrained minimized total variance $\mathcal{C}_n$ are respectively given by

$$\mathbf{D}_n \;=\; -\mathbf{I} + (\alpha_n^{-1}\theta^{-1}\mathbf{I} + \mathbf{X}_n\mathbf{X}_n^T)^{-1}\mathbf{X}_n\mathbf{X}_n^T, \tag{14}$$

$$\mathcal{C}_n \;=\; \alpha_n^2 \text{tr}[(\mathbf{I} + \alpha_n\theta\mathbf{X}_n\mathbf{X}_n^T)^{-2}\theta\mathbf{X}_n\mathbf{X}_n^T], \tag{15}$$

where α_n is computed from

$$m\rho_n^2 \;=\; \text{tr}[\mathbf{D}_n^T\mathbf{D}_n] = \text{tr}[(\mathbf{I} + \alpha_n\theta\mathbf{X}_n\mathbf{X}_n^T)^{-2}]. \tag{16}$$

For the model (11), we consider the following linear estimator[3]:

$$W_n = \mathbf{G}_n Y_n + H_n = [\mathbf{G}_n \; H_n] \begin{bmatrix} Y_n \\ 1 \end{bmatrix}, \tag{17}$$

[1] In this paper $\mathbf{I}$ is used for identity matrix. The perception of its dimension is taken from other matrices that added to it.

[2] Without loss of generality we assumed the model is real. It is because a linear complex equation can be represented by two linear real equations.

[3] The estimator (17) is a linear function of $[Y_n^T, 1]^T$.

Table 1. The recursion for computing W_n.

Initial values: $\hat{W}_{n_0} = 0$, $P_{n_0} = \alpha_n \theta \mathbf{I}$,
For $k = n_0, \cdots, n$,

$$\begin{aligned} e_k &= y_k - X_k^T \hat{W}_{k-1}, \\ P_k &= P_{k-1} - \frac{P_{k-1} X_n X_n^T P_{k-1}}{1 + X_k^T P_{k-1} X_n}, \\ \hat{W}_k &= \hat{W}_{k-1} + P_k X_k e_k, \\ W_n &= \hat{W}_n + H_n. \end{aligned}$$

where H_n is an arbitrary known vector independent of W and show that W_n can attain the minimum total variance while satisfying the constraint on the norm of the biased gradient matrix. For the above estimator using (11), the bias gradient matrix is given by

$$\mathbf{D}_n = \frac{\partial b(W)}{\partial W} = -\mathbf{I} + \mathbf{G}_n \mathbf{X}_n^T. \tag{18}$$

In order to determine $\mathbf{G}_n$, we compare (14) and (18) and get

$$\mathbf{G}_n = (\alpha_n^{-1}\theta^{-1}\mathbf{I} + \mathbf{X}_n \mathbf{X}_n^T)^{-1} \mathbf{X}_n. \tag{19}$$

Equivalently, the estimator is given by

$$W_n = (\alpha_n^{-1}\theta^{-1}\mathbf{I} + \mathbf{X}_n \mathbf{X}_n^T)^{-1} \mathbf{X}_n Y_n + H_n \tag{20}$$

Hence BCRLB is achievable using any linear estimator of the form (20). It is to be notified that if W_{n-1} is any given a priori estimate of W and if we choose H_n as follows

$$H_n = W_{n-1} - (\alpha_n^{-1}\theta^{-1}\mathbf{I} + \mathbf{X}_n \mathbf{X}_n^T)^{-1} \mathbf{X}_n \mathbf{X}_n^T W_{n-1}, \tag{21}$$

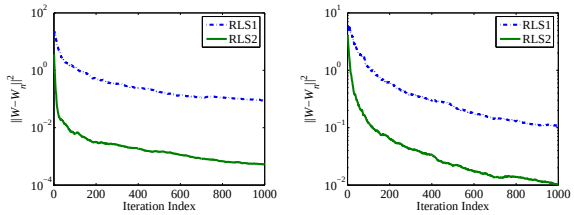
then from (20) we get

$$W_n = W_{n-1} + (\alpha_n^{-1}\theta^{-1}\mathbf{I} + \mathbf{X}_n \mathbf{X}_n^T)^{-1} \mathbf{X}_n \mathcal{E}_n, \tag{22}$$

where $\mathcal{E}_n = Y_n - \mathbf{X}_n^T W_{n-1}$. The above can be viewed as an innovation based method in parameter estimation. By using matrix inversion lemma [8] in computing $(\alpha_n^{-1}\theta^{-1}\mathbf{I} + \mathbf{X}_n \mathbf{X}_n^T)^{-1}$ iteratively as in Table 1, we observe that the above algorithm looks similar to the recursive least square (RLS) algorithm. In particular, if $\alpha_n \theta$ is constant and $H_n = 0$ the above algorithm will reduce to the RLS. We however note that appropriate choice of H_n and/or α_n improves the performance.

The following examples are given to illustrate the usefulness of the above discussions.

Example1 Consider the linear model in (10) where $X_k = [x_k, x_{k-1}]^T$, and x_k is randomly distributed in $\{\pm 1\}$. The additive noise v_n is Gaussian with mean γ and variance σ_v^2. We assumed $W = [-1, 2]^T$. Two RLS algorithms were executed. The first one used y_k as the desired signal and the second one used $y_k - \gamma$, i.e., we removed the known mean of the additive noise. Since γ is independent of W both algorithms are solutions of BCRLB optimization, i.e., both of them attain the same total variance $\mathcal{C}_n$ and bias gradient matrix $\mathbf{D}_n$. Figure 1 compares the average of the error squared of parameter estimate $\|W - W_n\|^2$, over 50 runs. As it shows the performance of the second RLS algorithm has improved. Actually the difference between these two estimators can be viewed as using $H_n = 0$ for the first one while $H_n = -(\alpha^{-1}\theta^{-1}\mathbf{I} + \mathbf{X}_n \mathbf{X}_n^H)^{-1} \mathbf{X}_n [\gamma, \cdots, \gamma]^T$ for the second.

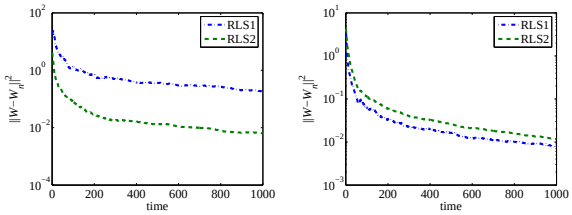


(a) (b)

Fig. 1. Comparison of $||W_n - W||^2$ for two RLS algorithm (a) $\gamma = 4, \sigma_v^2 = 0.1, \alpha\theta = 5$ (b) $\gamma = 4, \sigma_v^2 = 1.5, \alpha\theta = 0.5$

The following example illustrates that even if the additive noise has zero mean, the performance of the estimator may be improved by choosing proper H_n.

Example2 In the above example assume that the additive noise is $v_k = u_k + b_k$ where u_k and b_k, two independent iid sequences, u_k is a zero mean Gaussian noise with variance σ_v^2 and b_k is equal to either γ or $-\gamma$ with probability $\frac{1}{2}$, i.e., $f_{\mathbf{v}}(v) = \frac{1}{2\sqrt{2\pi}\sigma_v} \left(\exp(-\frac{(v-\gamma)^2}{2\sigma_v^2}) + \exp(-\frac{(v+\gamma)^2}{2\sigma_v^2}) \right)$. Obviously v_k has zero mean and for $\gamma \gg \sigma_v$ we have $\theta \simeq \sigma_v^{-2}$. The RLS algorithm which is the solution of BCRLB optimization with $H_n = 0$ and $\alpha_n = \alpha = \text{cte}$ is a candidate for estimating $W = [w_0, w_1]^T = [-1, 2]$. In order to improve the performance of the algorithm, we also add a known sequence to y_k and use the resulting sequence in the second RLS algorithm. For $\gamma \gg \sigma_v$ with a high probability we have $\gamma \text{sign}(v_k) \simeq b_k$. Therefore, $v_k - \gamma \text{sign}(v_k)$ is almost a zero mean Gaussian noise with variance σ_v^2. However, since $\text{sign}(v_k)$ is not available, we subtract $\gamma \text{sign}(y_k - X_k^T W_{n-1})$ as an estimate of $\gamma \text{sign}(v_k)$ from y_n and use $\tilde{y}_k = y_k - \gamma \text{sign}(y_k - X_k^T W_{n-1}) = X_k^T W + \tilde{v}_k$ as the desired signal, where $\tilde{v}_k = v_k - \gamma \text{sign}(\epsilon_k)$, $\epsilon_k = y_k - X_k^T W_{n-1}$, and W_{n-1} is the a prior estimate of the second RLS algorithm at time instant $n-1$. In this example, we assume $\sigma_v^2 = 1 \simeq \theta^{-1}$, and $\alpha = 2.5$. Figure 2 shows the average of weight error squared $\|W - W_n\|^2$ of both RLS algorithms over 50 runs for $\gamma = 5$ and $\gamma = 0.5$. For the case $\gamma = 5$, as Figure 2a shows, significant improvement is achieved by using the second algorithm. Note that at time instant n, the difference between parameter estimate of the two algorithms is given by $-(\alpha^{-1}\theta^{-1} + \mathbf{X}_n \mathbf{X}_k^T)^{-1} \sum_{k=n_0}^{n} X_k \gamma \text{sign}(\epsilon_k)$. For sufficiently large γ, $\text{sign}(\epsilon_k) \simeq \text{sign}(v_k)$ and therefore the



(a) (b)

Fig. 2. Comparison of the weight square error of two RLS algorithm, $\sigma_v^2 = 1$, $\alpha = 2.5$ (a) $\gamma = 5$ (b) $\gamma = 0.5$

above relation (i.e. H_n in the second RLS algorithm) is independent of W. For $\gamma = 0.5$ the first algorithm has better performance. Note that in this case the second algorithm is not a BCRLB optimized algorithm because the assumption $\text{sign}(\epsilon_n) \simeq \text{sign}(v_n)$ is invalid and the above relation is a function of W.

4. PARAMETER ESTIMATION BASED ON BIASED CRLB-SINGULAR CASE

In this section we assume that the FIM is singular. In this case, the covariance matrix $\mathbf{C}$ must satisfy [8]

$$\mathbf{C} \geq (\mathbf{I}+\mathbf{D})\mathbf{J}^{\dagger}(\mathbf{I}+\mathbf{D})^H \tag{23}$$

Where $\mathbf{J}^{\dagger}$ is Moore–Penrose pseudoinverse of $\mathbf{J}$. It is to be notified that for singular $\mathbf{J}$, there is no unbiased estimator with finite covariance. Assuming biased estimator, the regular condition of finite variance on each element of the estimator requires [6]

$$\mathbf{I}+\mathbf{D} = (\mathbf{I}+\mathbf{D})\mathbf{J}\mathbf{J}^{\dagger} \tag{24}$$

If we utilize the eigenvector/eigenvalue representation of $\mathbf{J}$, i.e.,

$$\mathbf{J} = [\mathcal{U} \;\; \mathcal{V}] \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{U}^H \\ \mathcal{V}^H \end{bmatrix} \tag{25}$$

where $[\mathcal{U} \;\; \mathcal{V}]$ is orthonormal, $\Lambda \in \mathbb{C}^{r\times r}$ is diagonal and positive definite, and r is the rank of $\mathbf{J}$, then equation (24) implies that

$$\mathbf{I}+\mathbf{D} = \mathbf{B}\mathcal{U}^H \tag{26}$$

where each column of $m \times r$ matrix $\mathcal{U}$ is an eigenvector corresponding to one of nonzero eigenvalues of $\mathbf{J}$ and $\mathbf{B}$ is an arbitrary $m \times r$ matrix. In the following theorem, minimizing $\text{tr}[(\mathbf{I}+\mathbf{D})\mathbf{J}^{\dagger}(\mathbf{I}+\mathbf{D})^H]$ with the constraint $\text{tr}[\mathbf{D}^H\mathbf{D}] \leq m\rho^2$ for a positive $\sqrt{\frac{m-r}{m}} \leq \rho \leq 1$ is our aim.

Theorem 2 *Consider inequality (23) the bias gradient matrix $\boldsymbol{D}$ which minimizes*

$$\mathcal{C} = tr[(\boldsymbol{I}+\boldsymbol{D})\boldsymbol{J}^{\dagger}(\boldsymbol{I}+\boldsymbol{D})^H] \tag{27}$$

with the constraint

$$tr[\boldsymbol{D}^H\boldsymbol{D}] \leq m\rho^2, \tag{28}$$

for a positive $\sqrt{\frac{m-r}{m}} \leq \rho \leq 1$, is given by

$$\boldsymbol{D} = -\boldsymbol{I} + \alpha\mathcal{U}(\Lambda^{-1}+\alpha\boldsymbol{I})^{-1}\mathcal{U}^H \tag{29}$$

where $\alpha \geq 0$ is the unique solution of the following equation

$$tr[(\boldsymbol{I}+\alpha\Lambda)^{-2}] = m\rho^2 - m + r, \tag{30}$$

and the minimum of $\mathcal{C}$ is given by

$$\mathcal{C} = \alpha^2 tr[(\Lambda^{-1}+\alpha\boldsymbol{I})^{-2}]. \tag{31}$$

Proof: For $\rho = 1$, obviously, $\alpha = 0$, $\mathbf{D}_n = -\mathbf{I}$, and $\mathcal{C} = 0$. For $\sqrt{\frac{m-r}{m}} \leq \rho < 1$, minimizing $\mathcal{C}$ under the constraint $\text{tr}(\mathbf{D}^H\mathbf{D}) \leq m\rho^2$ is done by the Lagrangian method [8]

$$L = \mathcal{C} + \alpha(\text{tr}[\mathbf{D}^H\mathbf{D}] - m\rho^2) \tag{32}$$

where α must be nonnegative [9]. From (25) we have

$$\mathbf{J}^{\dagger} = \mathcal{U}\Lambda^{-1}\mathcal{U}^H. \tag{33}$$

Substituting (26) and (33) in (27) and the result in (32) leads to the following strictly convex function of $\mathbf{B}\mathcal{U}^H$,

$$L = \text{tr}[\mathbf{B}\Lambda^{-1}\mathbf{B}^H] + \alpha(\text{tr}[(\mathbf{B}\mathcal{U}^H - \mathbf{I})(\mathbf{B}\mathcal{U}^H - \mathbf{I})^H] - m\rho^2)$$

where $\mathbf{B}$ and $\alpha > 0$ should be find. Differentiating L with respect to $\mathbf{B}$ and setting the result to zero, leads to

$$\mathbf{B} = \alpha\mathcal{U}(\Lambda^{-1}+\alpha\mathbf{I})^{-1} \tag{34}$$

Replacing (34) in (26) leads to equation (29). Under constraint $\text{tr}[(\mathbf{B}\mathcal{U}^H - \mathbf{I})(\mathbf{B}\mathcal{U}^H - \mathbf{I})^H] = \text{tr}[D^H D] \leq m\rho^2$, the unique minimum of the strictly convex function L is given by satisfying (28)(in which $D = -I + \mathbf{B}\mathcal{U}^H$) with equality [9]. i.e.

$$\text{tr}[(\mathbf{B}\mathcal{U}^H - \mathbf{I})(\mathbf{B}\mathcal{U}^H - \mathbf{I})^H] = m\rho^2. \tag{35}$$

Replacing (34) in (35) and with some algebraic computations we get (30). ■

It is not difficult to show that for $\rho < \rho_{\min} = \sqrt{\frac{m-r}{m}}$ there is no appropriate solution. It means that for singular case, as we mentioned before the estimator definitely will be biased.

5. LINEAR TIME INVARIANT MODEL-SINGULAR FIM

Consider again the linear model (10) where the FIM $\mathbf{J}_n$ is singular with rank r, i.e.,

$$\mathbf{J}_n = \theta\mathbf{X}_n\mathbf{X}_n^T = [\mathcal{U}_n \;\; \mathcal{V}_n] \begin{bmatrix} \Lambda_n & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{U}_n^T \\ \mathcal{V}_n^T \end{bmatrix} \tag{36}$$

To get the estimator in this case, we replace $\mathbf{X}_n$ by its skinny SVD in (11). The skinny SVD of $\mathbf{X}_n$ is given by [10]

$$\mathbf{X}_n = [\mathcal{U}_n \;\; \mathcal{V}_n] \begin{bmatrix} \theta^{-\frac{1}{2}}\Lambda_n^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{\mathcal{U}}_n^T \\ \bar{\mathcal{V}}_n^T \end{bmatrix}, \tag{37}$$

where $[\bar{\mathcal{U}}_n \;\; \bar{\mathcal{V}}_n]$ is the orthonormal matrix of left singular vectors of $\mathbf{X}_n$. Using (37) in (11), we get

$$Y_n = \theta^{-\frac{1}{2}}\bar{\mathcal{U}}_n\Lambda_n^{\frac{1}{2}}\mathcal{U}_n^T W + V_n. \tag{38}$$

Using Theorem 2 and (38), the optimum gradient matrix and the optimum total variance of model are given by

$$\mathbf{D} = -\mathbf{I} + \alpha_n\mathcal{U}_n(\Lambda_n^{-1} + \alpha_n\mathbf{I})^{-1}\mathcal{U}_n^T \tag{39}$$

$$\mathcal{C}_n = \alpha_n^2 \text{tr}[(\Lambda_n^{-1} + \alpha_n\mathbf{I})^{-2}\Lambda_n^{-1}], \tag{40}$$

where for a given $\rho_n \in [\sqrt{\frac{m-r}{m}}, 1]$ the value of α_n is the smallest positive solution of

$$\text{tr}[(\mathbf{I} + \alpha_n\Lambda_n)^{-2}] = m\rho_n^2 - m + r \tag{41}$$

We find the class of linear estimators of the form $W_n = \mathbf{G}_nY_n + H_n$ where their constrained total variances reach the BCRLB. The gradient matrix of this estimator is given by

$$\frac{\partial b(W)}{\partial W} = -\mathbf{I} + \mathbf{G}_n\mathbf{X}_n^T = -\mathbf{I} + \theta^{-\frac{1}{2}}\mathbf{G}_n\bar{\mathcal{U}}_n\Lambda_n^{\frac{1}{2}}\mathcal{U}_n^T. \tag{42}$$

Comparing (39) and (42), we get

$$\theta^{\frac{-1}{2}}\mathbf{G}_n\bar{\mathcal{U}}_n = \alpha_n\mathcal{U}_n(\Lambda_n^{-1} + \alpha_n\mathbf{I})^{-1}\Lambda_n^{-\frac{1}{2}}. \tag{43}$$

Hence, the (linear) estimate of W based on BCRLB optimization is given by

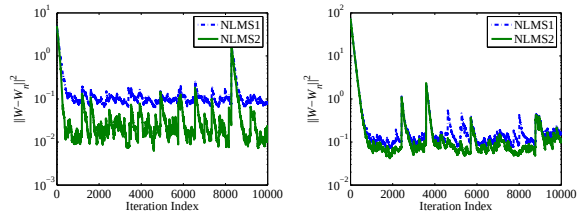
$$W_n = \theta^{\frac{1}{2}}\alpha_n\mathcal{U}_n(\Lambda_n^{-1} + \alpha_n\mathbf{I})^{-1}\Lambda_n^{-\frac{1}{2}}\bar{\mathcal{U}}_n^T\mathbf{Y}_n + H_n. \tag{44}$$

Assuming $\alpha_n = \text{cte} > 0$, and $H_n = 0$, there are some methods to compute W_n from observations iteratively (for instance see [13] for a class of low-rank adaptive filters). While full-rank RLS algorithm has a complexity of $\mathcal{O}(m^2)$, these algorithms, require only $\mathcal{O}(mr)$ operations per time-step. One interesting case is parameter estimation based on a single observation $n_0 = n$, i.e., $Y_n = y_n$ and $\mathbf{X}_n = X_n$. In this case we have $\mathcal{U}_n = \frac{X_n}{||X_n||}$, $\bar{\mathcal{U}}_n = 1$, $\Lambda_n = \theta\|X_n\|^2$, and $\alpha_n = \frac{1-\rho_n}{\rho_n\theta\|X_n\|^2}$; therefore from (44), we get

$$W_n = \frac{1-\rho_n}{\|X_n\|^2}X_ny_n + H_n \tag{45}$$

Assuming H_n is independent of θ, the parameter estimate W_n is also independent of θ. It means that the above estimator is robust against the noise PDF. If we choose,

$$H_n = W_{n-1} - \frac{1-\rho_n}{\|X_n\|^2}X_nX_n^TW_{n-1} \tag{46}$$



(a) (b)

Fig. 3. Comparison of the weight square error of two NLMS algorithm (a) $\sigma_v^2 = 0.01$ (b) $\sigma_v^2 = 0.3$

where W_{n-1} is the prior estimate of W, we get the following normalized least mean squared (NLMS) algorithm

$$W_n = W_{n-1} + \frac{1-\rho_n}{\|X_n\|^2}X_ne_n, \tag{47}$$

where $\mu_n = 1-\rho_n$ is the step size and $e_n = y_n - X_n^TW_{n-1}$. The maximum value of the step-size that keeps the algorithm to be a BCRLB minimization, is given by $\mu_{\max} = 1 - \rho_{\min} = 1 - \sqrt{\frac{m-1}{m}}$. Hence for NLMS algorithm to be a BCRLB minimization method we must have $\mu_n \in (0, 1 - \sqrt{\frac{m-1}{m}}]$. We also note that $\rho_n = 1$ does not necessary mean an estimator with undesirable bias, e.g., see the following example.

example3 Consider Example 2 where, two NLMS algorithm are executed. For the first one (NLMS1) we assume $\rho_n = 0.99$ and for the second one (NLMS2), we use the following,

$$\rho_n = \begin{cases} 1 & \text{if } \; |e_n| \le \gamma \\ 0.99 & \text{if } \; |e_n| > \gamma. \end{cases} \tag{48}$$

We also assumed $\gamma = 1$. Figure 3 shows the average of weight error squared $\|W - W_n\|^2$ of both NLMS algorithms over 50 runs for $\sigma_v^2 = 0.01$ and $\sigma_v^2 = 0.3$. For the case $\sigma_v^2 = 0.01$, as Figure 3a shows, significant improvement is achieved by the second algorithm. By using a counter in program, we noticed that for the second algorithm about %50 of time $\rho_n = 1$.

6. CONCLUSION

In this paper, we used the BCRLB to find the optimum estimate of an unknown parameter vector of a linear model. We shown that for a linear model, the optimum solution of BCRLB is achievable by a class of linear estimators. Our examples illustrated that while all estimators in this class have the same total variance and the same gradient matrix, their performances are not necessary the same. We concluded that the BCRLB is not a sufficient criteria to design optimal estimators. We also demonstrated that when the FIM is singular the input vectors must be projected onto a

dominant signal subspace of reduced ranks. We showed that the NLMS algorithm is optimal by means of the BCRLB criterion for estimating the unknown parameter, where only one single observation is used at a time. An example is also given to illustrate that large gradient matrix does not necessary mean an estimator with an undesirable bias.

Appendix: The FIM for a linear model

Because in (10) the noise process $\{v_k\}$ is an i.i.d random sequence, we have:

$$f(Y_n|W) = \prod_{k=n_0}^{n} f_{\mathbf{v}}(y_k - X_k^T W). \tag{49}$$

Hence

$$\log f(Y_n|W) = \sum_{k=n_0}^{n} \log f_{\mathbf{v}}(y_k - X^T W). \tag{50}$$

Therefore from (10) and above, we get

$$\frac{\partial \log f(Y_n|W)}{\partial W} = \sum_{k=n_0}^{n} \frac{f_{\mathbf{v}}^{'}(v_k)}{f_{\mathbf{v}}(v_k)} X_k^T. \tag{51}$$

From the definition of FIM (4), we find that

$$\mathbf{J}_n = \sum_{k=n_0}^{n} \sum_{l=n_0}^{n} E\left[\frac{f_{\mathbf{v}}^{'}(v_k)}{f_{\mathbf{v}}(v_k)} \frac{f_{\mathbf{v}}^{'}(v_l)}{f_{\mathbf{v}}(v_l)}\right] X_k X_l^T. \tag{52}$$

Since $\{v_k\}$ is an i.i.d. sequence, for $k \neq l$, we have

$$E\left[\frac{f_{\mathbf{v}}^{'}(v_k)}{f_{\mathbf{v}}(v_k)} \frac{f_{\mathbf{v}}^{'}(v_l)}{f_{\mathbf{v}}(v_l)}\right] = \left(E\left[\frac{f_{\mathbf{v}}^{'}(v)}{f_{\mathbf{v}}(v)}\right]\right)^2 = 0. \tag{53}$$

It is because

$$\left(E\left[\frac{f_{\mathbf{v}}^{'}(v)}{f_{\mathbf{v}}(v)}\right]\right)^2 = \left(\int_{-\infty}^{\infty} f_{\mathbf{v}}^{'}(x)dx\right)^2$$

and the right hand side is given by $f_{\mathbf{v}}(\infty) - f_{\mathbf{v}}(-\infty)$, where $f_{\mathbf{v}}(\infty) = f_{\mathbf{v}}(-\infty) = 0$. For $k = l$, we have

$$E\left[\frac{f_{\mathbf{v}}^{'}(v_k)}{f_{\mathbf{v}}(v_k)} \frac{f_{\mathbf{v}}^{'}(v_l)}{f_{\mathbf{v}}(v_l)}\right] = E\left[\left(\frac{f_{\mathbf{v}}^{'}(v)}{f_{\mathbf{v}}(v)}\right)^2\right] \triangleq \theta. \tag{54}$$

Hence substituting (53) and (54) in (52), we obtain

$$\mathbf{J}_n = \theta \sum_{k=n_0}^{n} X_k X_k^T. \tag{55}$$

which is equivalent to (12).

7. REFERENCES

[1] C.R. Rao,"Minimum variance and the estimation of several parameters,"*in Proc. Cambridge Phil. Soc.,* 1946, pp. 280-283

[2] E. L. Lehmann and G. Gasella, *Toery of Point Estimation*, Second ed. New York: Speringer-Verlag 1998.

[3] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the buniform CR bound,"*IEEE Trans. on Signal Processing*, Vol.44, pp.2026-2041, Aug. 1996.

[4] A. O. Hero, "A Cramér-Rao Type Lower Bound for Essentially Unbiased parameter Estimation," Lincoln Lab., Mass.Inst. Technol., Lexington, MA, Tech. Rep. 890, DTIC AD-A246666, 1992.

[5] Y. C. Eldar, "Minimum Variance in Biased Estimation: Bounds and Asymptotically Optimal Estimators,"*IEEE Trans. on Signal Processing*, Vol.52, No. 7. pp.1915-1930, July 2004.

[6] P. Stoica and T. L. Marzetta, "Parameter Estimation Problems with Singular Information Matrices,", *IEEE Trans. on Signal Processing*, no. 1, pp. 87-90, Jan., 2001.

[7] C. R. Rao, *Linear Statistical Inference and Its Applications,* 2nd. ed. New york: Wiley, 1973.

[8] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, New York: Wiely, 1968.

[9] D. P. Bertsekas, *Nonlinear Programming.* Second ed. Belmont, MA: Athena Scientific, 1999

[10] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge, U.K.: Cambridge Univ. Press, 1985.

[11] E. Kreindler and A. Jameson, "Conditions for nonnegativeness of partitioned matrices,"*IEEE Trans. Automat. Contr.,* vol. AC-17, pp. 147-148, Feb. 1972.

[12] S. M. Kay, *Fundamentals of Statistical Signal Processing:Estimation Theory.* Upper Sadle River, NJ: Prentice-Hall, 1993.

[13] P. Strobach,"Low-Rank Adaptive Filters", *IEEE Trans. on Signal Processing*, no. 12, pp. 2932-2947, Dec., 1996.