THE EXTENDED LEAST-SQUARES AND THE JOINT MAXIMUM-A-POSTERIORI - MAXIMUM-LIKELIHOOD ESTIMATION CRITERIA

Arie Yeredor and Ehud Weinstein

Department of Electrical Engineering - Systems
Tel-Aviv University, Tel-Aviv, 69978, ISRAEL
e-mail: arie@eng.tau.ac.il udi@eng.tau.ac.il

ABSTRACT

Approximate model equations often relate given measurements to unknown parameters whose estimate is sought. The Least-Squares (LS) estimation criterion assumes the measured data to be exact, and seeks parameters which minimize the model errors. Existing extensions of LS, such as the Total LS (TLS) and Constrained TLS (CTLS) take the opposite approach, namely assume the model equations to be exact, and attribute all errors to measurement inaccuracies. We introduce the Extended LS (XLS) criterion, which accommodates both error sources. We define 'pseudo-linear' models, with which we provide an iterative algorithm for minimization of the XLS criterion. Under certain statistical assumptions, we show that XLS coincides with a statistical criterion, which we term the 'joint Maximum-A-Posteriori - Maximum-Likelihood' (JMAP-ML) criterion. We identify the differences between the JMAP-ML and ML criteria, and explain the observed superiority of JMAP-ML over ML under non-asymptotic conditions.

1. INTRODUCTION

In a vast variety of problems in engineering it is desired to estimate an unknown vector of parameters $\boldsymbol{\theta}$ from a vector of data measurements $\mathbf{x}$. The parameters and the data can generally be related by an approximate set of model equations:

$$\mathbf{g}(\mathbf{x},\boldsymbol{\theta}) \approx \mathbf{0}. \tag{1}$$

The inconsistency of the set (1) may often be attributed to two possibly distinct mechanisms: model mismatch and measurement inaccuracies. Model mismatch encompasses inaccuracies that exist in (1) even when exact measurements are used. Measurement inaccuracies, on the other hand, account for possible deviations of the measured data $\mathbf{x}$ from the true (unknown) data, say $\tilde{\mathbf{x}}$, with which (1) is exact (in the absence of model mismatch). In other words, (1) may be broken down into two distinct (in)equalities:

$$\mathbf{g}(\tilde{\mathbf{x}},\boldsymbol{\theta}) \approx \mathbf{0} \tag{2a}$$

$$\mathbf{x}-\tilde{\mathbf{x}} \approx \mathbf{0} \tag{2b}$$

where $\tilde{\mathbf{x}}$ is a vector of "presumed" ("accurate") data, so that (2a) accounts only for model mismatch, while (2b) accounts only for measurement inaccuracies. Of course, measurement inaccuracies can always be attributed to model errors. However, they are often caused by mechanisms that are essentially unrelated to the underlying model. In such cases the distinction in (2a), (2b) is justified.

The well-known Least Squares (LS) estimation approach seeks parameters $\boldsymbol{\theta}$, which, together with the given measurements $\mathbf{x}$ minimize the (possibly weighted) Euclidean norm of $\mathbf{g}(\mathbf{x},\boldsymbol{\theta})$:

$$\min_{\boldsymbol{\theta}}\{\mathbf{g}^T(\mathbf{x},\boldsymbol{\theta})\mathbf{W}\mathbf{g}(\mathbf{x},\boldsymbol{\theta})\} \Rightarrow \hat{\boldsymbol{\theta}}_{\text{LS}} \tag{3}$$

where $\mathbf{W}$ is some pre-specified symmetric positive-definite weight matrix. $\hat{\boldsymbol{\theta}}_{\text{LS}}$ is actually the value that brings $\mathbf{g}(\mathbf{x},\boldsymbol{\theta})$ as close as possible to its presumed value of $\mathbf{0}$. However, it does not allow substitution of the observed data, thus implying that (2b) is completely satisfied (with equality).

On the other hand, in the context of pseudo-linear models (to be defined immediately), some well-known modifications of LS can be regarded as taking the opposite approach. Specifically, they attempt to find a vector of 'presumed' data $\tilde{\mathbf{x}}$ and an associated parameters vector $\boldsymbol{\theta}$, with which the model equation (2a) is satisfied (with equality), while keeping to minimum the deviation of the 'presumed' data $\tilde{\mathbf{x}}$ from the observed data $\mathbf{x}$.

Such is, e.g., the Total LS (TLS) method, which was originally addressed by Golub and Van-Loan [1]. Various aspects of TLS have since been thoroughly explored by Van-Huffel and Vandewalle (e.g. [2], [3]), Dowling and DeGroat (e.g. [4]) and others. The main drawback of the TLS approach in many engineering applications is its inability to account for structural limitations of the pseudo-linear model. Modifications of TLS in that respect were proposed by Abatzoglou and Mendel [5], Cadzow et al. [6] (termed Constrained TLS (CTLS)), as well as DeMoor et al. [7] (termed Structured TLS (STLS)).

In order to adapt the estimation approach to the distinction between the two error sources, the following Extended LS (XLS) criterion may be considered:

Forming the concatenation of (2a) and (2b), we get

$$\tilde{\mathbf{g}}(\mathbf{x},\tilde{\mathbf{x}},\boldsymbol{\theta}) \triangleq \begin{bmatrix} \mathbf{g}(\tilde{\mathbf{x}},\boldsymbol{\theta}) \\ \mathbf{x}-\tilde{\mathbf{x}} \end{bmatrix} \approx \mathbf{0} \tag{4}$$

to which we may now apply the LS criterion, yielding the XLS estimate of $\boldsymbol{\theta}$. However, since $\tilde{\mathbf{x}}$ is obviously unknown, we have to minimize with respect to $\tilde{\mathbf{x}}$ as well, obtaining as a by-product the XLS estimate of the presumed data:

$$\min_{\tilde{\mathbf{x}},\boldsymbol{\theta}}\{\tilde{\mathbf{g}}^T(\mathbf{x},\tilde{\mathbf{x}},\boldsymbol{\theta})\tilde{\mathbf{W}}\tilde{\mathbf{g}}(\mathbf{x},\tilde{\mathbf{x}},\boldsymbol{\theta})\} \Rightarrow \hat{\boldsymbol{\theta}}_{\text{XLS}} \;\; (+\hat{\mathbf{x}}_{\text{XLS}}) \tag{5}$$

where $\tilde{\mathbf{W}}$ is the extended weight matrix. Often a block-diagonal $\tilde{\mathbf{W}}$ would be chosen, $\tilde{\mathbf{W}} = \text{diag}(\mathbf{W}_g, \mathbf{W}_x)$, where $\mathbf{W}_g$ and $\mathbf{W}_x$ fit the dimensions of $\mathbf{g}(\tilde{\mathbf{x}},\boldsymbol{\theta})$ and $\mathbf{x}$, respectively (which need not be the same), so that the XLS cost-function may assume the following form:

$$C_{\text{XLS}}(\tilde{\mathbf{x}},\boldsymbol{\theta}) = \mathbf{g}^T(\tilde{\mathbf{x}},\boldsymbol{\theta})\mathbf{W}_g\mathbf{g}(\tilde{\mathbf{x}},\boldsymbol{\theta}) + (\mathbf{x}-\tilde{\mathbf{x}})^T\mathbf{W}_x(\mathbf{x}-\tilde{\mathbf{x}}) \tag{6}$$

to be minimized with respect to both $\tilde{\mathbf{x}}$ and $\boldsymbol{\theta}$, given $\mathbf{x}$.

It can be easily observed, that when $\mathbf{W}_g \ll \mathbf{W}_x$, the minimization with respect to $\tilde{\mathbf{x}}$ is dominated by the second term, and is obviously attained near $\tilde{\mathbf{x}} \approx \mathbf{x}$. This means, that the minimization with respect to $\boldsymbol{\theta}$ simply minimizes the first term with $\tilde{\mathbf{x}}$ replaced by $\mathbf{x}$. Obviously, this coincides with the LS estimate. We therefore identify the LS approach as a limiting case which rules out deviations of the measured data from the true data, by heavily penalizing them with $\mathbf{W}_x$. Thus the LS estimate "blames" all the inconsistency in (1) on model mismatch.

When, on the other hand, $\mathbf{W}_x \ll \mathbf{W}_g$, minimization is attained when the first term is nearly zeroed out. That means that in the second term the minimal perturbation of the measured data is sought, for which the model equations can be completely satisfied with some value of $\boldsymbol{\theta}$. This approach is, in a sense, the opposite of LS, as it blames all the inconsistency on measurement inaccuracies. It therefore generates its own version of presumed data, which is perfectly consistent with some $\boldsymbol{\theta}$.

Careful selection of $\mathbf{W}_g$ and $\mathbf{W}_x$ would normally reflect the optimal sharing of inconsistency between model mismatch and measurement errors. When statistical assumptions as to the nature of the model mismatch and measurement noises are incorporated, specific selection of weights reflects specific statistical interpretations for the obtained estimate. These interpretations are discussed in section 4.

2. PSEUDO-LINEAR MODELS

In the context of this paper we now reduce our discussion to models termed 'pseudo-linear' models, which are linear in the measurements given the parameters, and vice versa. Such models are necessarily of the form

$$\mathbf{g}(\tilde{\mathbf{x}}, \boldsymbol{\theta}) = \mathbf{A}(\tilde{\mathbf{x}})\mathbf{v}(\boldsymbol{\theta}) \tag{7}$$

where each element of the matrix $\mathbf{A}(\tilde{\mathbf{x}})$ is a linear function of the elements of $\tilde{\mathbf{x}}$, and where the vector $\mathbf{v}(\boldsymbol{\theta})$ contains the vector $\boldsymbol{\theta}$ itself and possibly some additional constant. Explicitly stated, if $\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_N \end{bmatrix}^T$, then $\mathbf{A}(\tilde{\mathbf{x}})$ can be expressed as:

$$\mathbf{A}(\tilde{\mathbf{x}}) = \mathbf{A}_0 + \sum_{n=1}^{N} \tilde{x}_n \mathbf{A}_n \tag{8a}$$

where $\mathbf{A}_0, \mathbf{A}_1, \ldots \mathbf{A}_N$ are a set of constant matrices; also,

$$\mathbf{v}(\boldsymbol{\theta}) = \begin{bmatrix} v_0 \\ \boldsymbol{\theta} \end{bmatrix} \tag{8b}$$

where v_0 is some constant. The role of the free matrix A_0 and the free constant v_0 is to provide optional terms that are purely linear in $\boldsymbol{\theta}$ and $\tilde{\mathbf{x}}$ (respectively). In the absence of such terms in the model, $\mathbf{A}_0$ or v_0 are set to 0. Without loss of generality we may set $v_0 = 1$ when v_0 is non-zero. In that way all the scaling is taken care of in the $\mathbf{A}(\tilde{\mathbf{x}})$ term. Note that if $v_0 = 0$, the model equations can be made exact with a trivial choice $\boldsymbol{\theta} = \mathbf{0}$, which renders such cases uninteresting, as the XLS criterion (5) would be minimized (zeroed out) by setting in addition $\tilde{\mathbf{x}} = \mathbf{x}$. Since we are only interested in interesting cases, we assume $v_0 = 1$.

Pseudo-linear models can be used with the XLS criterion in a variety of problems. For example, in estimating the parameters of an Auto-Regressive (AR) process contaminated by additive noise, the process' model equations $\tilde{x}_n = -\sum_{k=1}^{p} \theta_k \tilde{x}_{n-k} + u_n$ (where $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \cdots \ \theta_p]^T$ are the unknown parameters and u_n is the driving noise) can be expressed as:

$$\mathbf{g}(\tilde{\mathbf{x}}, \boldsymbol{\theta}) = \mathbf{A}(\tilde{\mathbf{x}}) \begin{bmatrix} 1 \\ \boldsymbol{\theta} \end{bmatrix} = \left[\mathbf{A}_0 + \sum_{n=1}^{N} \tilde{x}_n \mathbf{A}_n \right] \begin{bmatrix} 1 \\ \boldsymbol{\theta} \end{bmatrix} \approx \mathbf{0} \tag{9}$$

where $\tilde{\mathbf{x}} = [\tilde{x}_1 \ \tilde{x}_2 \ \cdots \ \tilde{x}_N]^T$ are (unavailable) N samples of the underlying process, $\mathbf{A}_n \ n = 1, 2, \ldots N$ are $N \times (p+1)$ Toeplitz matrices with a diagonal of 1-s at the $(n, 1)$ thru $(n+p, 1+p)$ entries (and zeros elsewhere), and $\mathbf{A}_0$ is also mostly zeros $N \times (p+1)$, whose upper right triangle reflects known non-zero initial conditions, if any.

The $\approx$ sign in (9) actually conceals the driving noise sequence u_n. The N available samples comprising the data vector $\mathbf{x}$ are noisy versions of $\tilde{\mathbf{x}}$, such that the $\approx$ sign of (2b) conceals the observation noise. We refrain from introducing statistical assumptions pertaining to the two noise sources, since our approach is purely deterministic at this stage.

In the context of this problem, the ordinary LS approach would ignore the inequality $\mathbf{x} \approx \tilde{\mathbf{x}}$, and treat $\mathbf{x}$ as a noiseless realization of $\tilde{\mathbf{x}}$, leading to the well-known Yule-Walker type equations for estimating $\boldsymbol{\theta}$. On the other hand, CTLS or STLS ignore the inequality in (9), and are therefore only suitable for estimating $\boldsymbol{\theta}$ from a noisy realization of an AR process with zero innovations, such as a linear system's zero-input response to known initial conditions - see e.g. De-Moor [7]. However, neither LS nor STLS (or CTLS) are suited to this noisy AR process problem, which is better approached by XLS.

Pseudo-linear models are also useful e.g. in the context of estimating the parameters of an ARX systems from noisy input and output data, and in numerous array-processing problems.

3. MINIMIZATION ALGORITHMS

Several minimization algorithms of the XLS criterion for pseudo-linear models are outlined in [8]. In this paper we outline the most straightforward minimization strategy, namely the 'Alternating Coordinates Minimization' (ACM). The basic idea exploits the availability of closed-form solutions both for minimization with respect to $\boldsymbol{\theta}$ with $\tilde{\mathbf{x}}$ fixed, and vice-versa. Thus, beginning with an intelligent guess for $\boldsymbol{\theta}$ or $\tilde{\mathbf{x}}$, the ACM algorithm alternates between minimization with respect to $\boldsymbol{\theta}$ treating $\tilde{\mathbf{x}}$ as constant, and minimization with respect to $\tilde{\mathbf{x}}$ treating $\boldsymbol{\theta}$ as constant.

Due to the pseudo-linear structure, each minimization phase involves a quadratic minimization, which results in a unique global minimum (assuming the other coordinates fixed). Thus, in each iteration the value of $C_{\text{XLS}}(\tilde{\mathbf{x}}, \boldsymbol{\theta})$ is guaranteed not to increase (usually to decrease). Since $C_{\text{XLS}}(\tilde{\mathbf{x}}, \boldsymbol{\theta})$ is bounded below (e.g. by zero), convergence of the algorithm to a (possibly local) minimum is guaranteed [1].

An explicit algorithm follows (a detailed derivation is given in [8]), starting with some initial guess $\hat{\tilde{\mathbf{x}}}^{[0]}$ for $\tilde{\mathbf{x}}$, possibly (but not necessarily) $\hat{\tilde{\mathbf{x}}}^{[0]} = \mathbf{x}$:

[1] To be precise, the constant decrease and boundedness guarantee convergence; The convergence point is guaranteed to be a (local) minimum because the only (local) minima can be stationary points of the alternating minimization operation.

Algorithm:

For $k = 1, 2, \ldots$ until convergence:

I. Minimize with respect to $\boldsymbol{\theta}$:

Construct

$$\mathbf{A}^{[k]} = \mathbf{A}_0 + \sum_{n=1}^{N} \hat{\tilde{x}}_n^{[k-1]} \mathbf{A}_n$$

Form the partition:

$$\begin{bmatrix} b^{[k]} & \mathbf{b}^{[k]^T} \\ \mathbf{b}^{[k]} & \mathbf{B}^{[k]} \end{bmatrix} = \mathbf{A}^{[k]^T} \mathbf{W}_g \mathbf{A}^{[k]}$$

Obtain $\hat{\boldsymbol{\theta}}^{[k]}$:

$$\hat{\boldsymbol{\theta}}^{[k]} = -\mathbf{B}^{[k]^{-1}} \mathbf{b}^{[k]}$$

II. Minimize with respect to $\tilde{\mathbf{x}}$:

Let

$$\mathbf{v}^{[k]} = \begin{bmatrix} 1 & \hat{\boldsymbol{\theta}}^{[k]^T} \end{bmatrix}^T$$

Construct

$$\mathbf{t}_0^{[k]} = \mathbf{A}_0 \mathbf{v}^{[k]}$$

$$\mathbf{T}^{[k]} = [\mathbf{A}_1 \mathbf{v}^{[k]} \vdots \mathbf{A}_2 \mathbf{v}^{[k]} \vdots \cdots \vdots \mathbf{A}_N \mathbf{v}^{[k]}]$$

Obtain $\hat{\tilde{\mathbf{x}}}^{[k]}$:

$$\hat{\tilde{\mathbf{x}}}^{[k]} = \left[\mathbf{T}^{[k]^T} \mathbf{W}_g \mathbf{T}^{[k]} + \mathbf{W}_x\right]^{-1} \cdot \cdot \left[\mathbf{W}_x \mathbf{x} - \mathbf{T}^{[k]^T} \mathbf{W}_g \mathbf{t}_0^{[k]}\right]$$

Upon convergence ($k = K$), set $\hat{\boldsymbol{\theta}}_{\text{XLS}} = \hat{\boldsymbol{\theta}}^{[K]}$.

As a by-product, $\hat{\mathbf{x}}_{\text{XLS}} = \hat{\tilde{\mathbf{x}}}^{[K]}$.

The ACM algorithm exhibits slow convergence rates in many cases, depending strongly on the relative scale between the weight matrices $\mathbf{W}_x$ and $\mathbf{W}_g$. Alternative algorithms (based on extending the CTLS [6] and STLS [7] methods) with significantly faster convergence rates and moderate computational costs are presented in [8]. Note that although the ACM algorithm is reminiscent of the Estimate-Maximize (EM) algorithm (e.g. [9]) for computing the Maximum-Likelihood (ML) estimate, the two algorithms (and respective estimators) are essentially different - as discussed in the next section and in [8].

4. STATISTICAL INTERPRETATIONS

The LS criterion often has a statistical interpretation, as it coincides with the ML criterion, e.g. when the model inaccuracies are modeled as a Gaussian noise. In this section we present the statistical interpretation of the XLS criterion under certain similar assumptions. We show that it coincides with a different statistical criterion, which we term 'Joint Maximum-A-Posteriori - Maximum-Likelihood' (JMAP-ML).

4.1. The JMAP - ML estimation criterion

Consider the following statistical model: Let $\tilde{\mathbf{x}}$ and $\mathbf{x}$ be two zero-mean jointly Gaussian random vectors (r.v.'s), whose joint covariance matrix is known up to an unknown parameters vector $\boldsymbol{\theta}$,

$$\begin{bmatrix} \tilde{\mathbf{x}} \\ \mathbf{x} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}(\boldsymbol{\theta}) \triangleq \begin{bmatrix} \boldsymbol{\Lambda}_{\tilde{x}\tilde{x}}(\boldsymbol{\theta}) & \boldsymbol{\Lambda}_{\tilde{x}x}(\boldsymbol{\theta}) \\ \boldsymbol{\Lambda}_{x\tilde{x}}(\boldsymbol{\theta}) & \boldsymbol{\Lambda}_{xx}(\boldsymbol{\theta}) \end{bmatrix}\right). \quad (10)$$

Assume that we are given a realization of $\mathbf{x}$ (but have no access to $\tilde{\mathbf{x}}$), and wish to estimate $\boldsymbol{\theta}$. The classical ML approach would estimate $\boldsymbol{\theta}$ as the maximizer of the (marginal) probability density function (pdf) of $\mathbf{x}$, $f(\mathbf{x};\boldsymbol{\theta})$ (or its log), or, equivalently, as the minimizer of

$$C_1(\boldsymbol{\theta}) \triangleq \mathbf{x}^T \boldsymbol{\Lambda}_{xx}^{-1}(\boldsymbol{\theta}) \mathbf{x} + \log|\boldsymbol{\Lambda}_{xx}(\boldsymbol{\theta})|. \quad (11)$$

An alternative approach would be to involve an estimate of the unobserved $\tilde{\mathbf{x}}$, by maximizing the joint pdf of the observation $\mathbf{x}$ and the unknown $\tilde{\mathbf{x}}$ with respect to both $\tilde{\mathbf{x}}$ and $\boldsymbol{\theta}$. This yields what we would term the 'joint MAP-ML' estimate of $\boldsymbol{\theta}$, since it involves a joint MAP estimate of $\tilde{\mathbf{x}}$ and an ML estimate of $\boldsymbol{\theta}$. For brevity we shall denote that estimate $\hat{\boldsymbol{\theta}}_{\text{J}}$:

$$\max_{\boldsymbol{\theta}} \left\{ \max_{\tilde{\mathbf{x}}} \{\log f(\mathbf{x}, \tilde{\mathbf{x}}; \boldsymbol{\theta})\} \right\} \Rightarrow \hat{\boldsymbol{\theta}}_{\text{J}}. \quad (12)$$

The value of $\tilde{\mathbf{x}}$ that maximizes $\log f(\mathbf{x}, \tilde{\mathbf{x}}; \boldsymbol{\theta})$ is given by the same value that maximizes the conditional $\log f(\tilde{\mathbf{x}}|\mathbf{x}; \boldsymbol{\theta})$, namely - the well-known MAP estimate $\hat{\tilde{\mathbf{x}}}_{\text{MAP}} = \boldsymbol{\Lambda}_{\tilde{x}x}(\boldsymbol{\theta}) \boldsymbol{\Lambda}_{xx}^{-1}(\boldsymbol{\theta}) \mathbf{x}$, which, when substituted into (12) eliminates $\tilde{\mathbf{x}}$ from the minimization. The problem can thus be reduced to maximization with respect to $\boldsymbol{\theta}$ of $\log f(\mathbf{x}, \hat{\tilde{\mathbf{x}}}_{\text{MAP}}; \boldsymbol{\theta})$, which translates into minimization of the following cost-function:

$$C_2(\boldsymbol{\theta}) \triangleq \mathbf{x}^T [\mathbf{Q}^T(\boldsymbol{\theta}) \boldsymbol{\Lambda}^{-1}(\boldsymbol{\theta}) \mathbf{Q}(\boldsymbol{\theta})] \mathbf{x} + \log|\boldsymbol{\Lambda}(\boldsymbol{\theta})| \quad (13)$$

where

$$\mathbf{Q}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{I} \\ \boldsymbol{\Lambda}_{\tilde{x}x}(\boldsymbol{\theta}) \boldsymbol{\Lambda}_{xx}^{-1}(\boldsymbol{\theta}) \end{bmatrix}. \quad (14)$$

Using four-blocks inverse notation for $\boldsymbol{\Lambda}^{-1}$,

$$\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \mathbf{P} & -\mathbf{P} \boldsymbol{\Lambda}_{\tilde{x}x} \boldsymbol{\Lambda}_{xx}^{-1} \\ -\boldsymbol{\Lambda}_{xx}^{-1} \boldsymbol{\Lambda}_{x\tilde{x}} \mathbf{P} & \boldsymbol{\Lambda}_{xx}^{-1} + \boldsymbol{\Lambda}_{xx}^{-1} \boldsymbol{\Lambda}_{x\tilde{x}} \mathbf{P} \boldsymbol{\Lambda}_{\tilde{x}x} \boldsymbol{\Lambda}_{xx}^{-1} \end{bmatrix} \quad (15)$$

where $\mathbf{P} = [\boldsymbol{\Lambda}_{\tilde{x}\tilde{x}} - \boldsymbol{\Lambda}_{\tilde{x}x} \boldsymbol{\Lambda}_{xx}^{-1} \boldsymbol{\Lambda}_{x\tilde{x}}]^{-1}$, and applying some cumbersome but straightforward algebraic manipulations, we obtain

$$\mathbf{Q}^T(\boldsymbol{\theta}) \boldsymbol{\Lambda}^{-1}(\boldsymbol{\theta}) \mathbf{Q}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_{xx}^{-1}(\boldsymbol{\theta}) \quad (16)$$

rendering identical the first terms of both $C_1(\boldsymbol{\theta})$ and $C_2(\boldsymbol{\theta})$. Note that only these (first) terms in each cost function carry the direct dependence on the measurements $\mathbf{x}$.

As for the deterministic part, using a 2×2 block-triangular factorization for $\boldsymbol{\Lambda}$, we can obtain (see [8] for details) $|\boldsymbol{\Lambda}| = |\boldsymbol{\Lambda}_{xx}| \cdot |\boldsymbol{\Lambda}_{\tilde{x}\tilde{x}} - \boldsymbol{\Lambda}_{\tilde{x}x} \boldsymbol{\Lambda}_{xx}^{-1} \boldsymbol{\Lambda}_{x\tilde{x}}|$, so that $C_2(\boldsymbol{\theta})$ can be written in terms of $C_1(\boldsymbol{\theta})$:

$$C_2(\boldsymbol{\theta}) = C_1(\boldsymbol{\theta}) + \log|\boldsymbol{\Lambda}_{\tilde{x}\tilde{x}}(\boldsymbol{\theta}) - \boldsymbol{\Lambda}_{\tilde{x}x}(\boldsymbol{\theta}) \boldsymbol{\Lambda}_{xx}^{-1}(\boldsymbol{\theta}) \boldsymbol{\Lambda}_{x\tilde{x}}(\boldsymbol{\theta})|. \quad (17)$$

The additional term is a deterministic function of $\boldsymbol{\theta}$ (independent of the measurements), which has the following interesting property: For each value of $\boldsymbol{\theta}$, consider the hypothetical problem of estimating $\tilde{\mathbf{x}}$ from $\mathbf{x}$ when $\boldsymbol{\theta}$ is known. The additional term of (17) coincides with the determinant of the estimation error covariance in that hypothetical problem. It can therefore be expected, that in the original problem of estimating $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{\text{J}}$ (minimizer of C_2) should tend away from $\hat{\boldsymbol{\theta}}_{\text{ML}}$ (minimizer of C_1) in the direction of values of $\boldsymbol{\theta}$ with which the hypothetical estimation of $\tilde{\mathbf{x}}$ from $\mathbf{x}$ would attain a *smaller* covariance.

For example, note that this tendency explains observations made e.g. in [10], [11]: when the two methods are applied to the estimation of the poles of noisy speech (modeled as an all-poles process),

the poles estimated using the criterion we termed 'JMAP-ML' tend more towards the unit-circle relative to their ML estimate counterparts. This is because for poles closer to the unit-circle, estimating the underlying speech process ($\tilde{\mathbf{x}}$) from the noisy measurement ($\mathbf{x}$) would attain a smaller error covariance (under similar noise conditions)[2] - which means a smaller second term in (17); therefore an estimate of $\boldsymbol{\theta}$ using $C_2(\boldsymbol{\theta})$ would tend more towards that constellation than an estimate based on $C_1(\boldsymbol{\theta})$

Under asymptotic ($N \to \infty$) conditions, $\hat{\boldsymbol{\theta}}_{\text{ML}}$ is well-known to be unbiased with minimum variance (attaining the Cramér-Rao Lower Bound). However, under non-asymptotic conditions $\hat{\boldsymbol{\theta}}_{\text{ML}}$ is often biased away from that constellation, so that $\hat{\boldsymbol{\theta}}_{\text{J}}$ may partially correct that bias, and even have a smaller variance (and mean squared error (m.s.e.)) than $\hat{\boldsymbol{\theta}}_{\text{ML}}$. In fact, in [8], a comprehensive error analysis of $\hat{\boldsymbol{\theta}}_{\text{ML}}$ and $\hat{\boldsymbol{\theta}}_{\text{J}}$ in estimating the parameter of a first-order AR process in white noise reveals the advantages of $\hat{\boldsymbol{\theta}}_{\text{J}}$, especially with short data records.

4.2. JMAP-ML and the XLS criterion

We now turn to relate the XLS criterion to the JMAP-ML criterion. The pseudo-linear model (7) may also be written as $\mathbf{g}(\tilde{\mathbf{x}}, \boldsymbol{\theta}) = \mathbf{t}_0(\boldsymbol{\theta}) + \mathbf{T}(\boldsymbol{\theta})\tilde{\mathbf{x}}$, where

$$\mathbf{t}_n(\boldsymbol{\theta}) \stackrel{\triangle}{=} \mathbf{A}_n \mathbf{v}(\boldsymbol{\theta}) \quad n = 0, 1, \ldots N \tag{18}$$

and

$$\mathbf{T}(\boldsymbol{\theta}) = [\mathbf{t}_1(\boldsymbol{\theta}) \vdots \mathbf{t}_2(\boldsymbol{\theta}) \vdots \cdots \vdots \mathbf{t}_N(\boldsymbol{\theta})]. \tag{19}$$

Let $\mathbf{u}$ denote the vector of model errors, namely $\mathbf{u} = \mathbf{g}(\tilde{\mathbf{x}}, \boldsymbol{\theta})$, assumed to be zero-mean Gaussian with covariance $\boldsymbol{\Lambda}_{uu}$.

We now make several further assumptions (some of which can be relaxed, as detailed in [8]):

i. $\mathbf{t}_0(\boldsymbol{\theta}) = \mathbf{0}$ (for all possible values of $\boldsymbol{\theta}$);

ii. $\mathbf{T}(\boldsymbol{\theta})$ is full-rank (for all possible values of $\boldsymbol{\theta}$);

iii. The dimension N of $\tilde{\mathbf{x}}$ equals the dimension p of $\mathbf{g}(\tilde{\mathbf{x}}, \boldsymbol{\theta})$.

Note that by assumptions *(ii)* and *(iii)* $\mathbf{T}(\boldsymbol{\theta})$ is square invertible. Consequently, $\tilde{\mathbf{x}}$ is also a zero-mean Gaussian vector, with covariance $\boldsymbol{\Lambda}_{\tilde{x}\tilde{x}}(\boldsymbol{\theta}) = \mathbf{T}^{-1}(\boldsymbol{\theta})\boldsymbol{\Lambda}_{uu}(\mathbf{T}^{-1}(\boldsymbol{\theta}))^T$, so that

$$\log f(\tilde{\mathbf{x}}; \boldsymbol{\theta}) = c + \log|\mathbf{T}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{T}(\boldsymbol{\theta})\tilde{\mathbf{x}})^T \boldsymbol{\Lambda}_{uu}^{-1}(\mathbf{T}(\boldsymbol{\theta})\tilde{\mathbf{x}}) \tag{20}$$

where c is a constant that does not depend on $\boldsymbol{\theta}$.

Let $\boldsymbol{\epsilon} \stackrel{\triangle}{=} \mathbf{x} - \tilde{\mathbf{x}}$ denote the vector of measurement errors, assumed zero-mean Gaussian, independent of $\mathbf{u}$, with covariance $\boldsymbol{\Lambda}_{\epsilon\epsilon}$. Then $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are zero-mean jointly Gaussian, as in (10). The JMAP-ML criterion for estimating $\boldsymbol{\theta}$ from $\mathbf{x}$ can be written as $C_2(\tilde{\mathbf{x}}, \boldsymbol{\theta}) = \log f(\mathbf{x}, \tilde{\mathbf{x}}; \boldsymbol{\theta}) = \log f(\mathbf{x}|\tilde{\mathbf{x}}) + \log f(\tilde{\mathbf{x}}; \boldsymbol{\theta})$, or

$$\begin{aligned} C_2(\tilde{\mathbf{x}}, \boldsymbol{\theta}) &= c' - \frac{1}{2}(\mathbf{x} - \tilde{\mathbf{x}})^T \boldsymbol{\Lambda}_{\epsilon\epsilon}^{-1}(\mathbf{x} - \tilde{\mathbf{x}}) + \log|\mathbf{T}(\boldsymbol{\theta})| \\ &- \frac{1}{2}(\mathbf{T}(\boldsymbol{\theta})\tilde{\mathbf{x}})^T \boldsymbol{\Lambda}_{uu}^{-1}(\mathbf{T}(\boldsymbol{\theta})\tilde{\mathbf{x}}) \end{aligned} \tag{21}$$

where c' is another constant, independent of $\boldsymbol{\theta}$. If we further assume that $|\mathbf{T}(\boldsymbol{\theta})|$ does not depend on $\boldsymbol{\theta}$ (although $\mathbf{T}(\boldsymbol{\theta})$ certainly does; this is the case in many situations of interest), we identify that maximization of the JMAP-ML criterion is equivalent with the minimization of

$$C_{\text{XLS}}(\tilde{\mathbf{x}}, \boldsymbol{\theta}) = \mathbf{g}^T(\tilde{\mathbf{x}}, \boldsymbol{\theta})\mathbf{W}_g\mathbf{g}(\tilde{\mathbf{x}}, \boldsymbol{\theta}) + (\mathbf{x} - \tilde{\mathbf{x}})^T \mathbf{W}_x(\mathbf{x} - \tilde{\mathbf{x}}) \tag{22}$$

where $\mathbf{W}_g = \boldsymbol{\Lambda}_{uu}^{-1}$ and $\mathbf{W}_x = \boldsymbol{\Lambda}_{\epsilon\epsilon}^{-1}$. This is exactly the block-diagonal weights version (6). However, while this 'natural' choice of weights attributes a statistical interpretation to the XLS criterion, it is not necessarily optimal (in terms of the attained m.s.e.) in a given statistical scenario - as demonstrated in [8].

[2] The ability to obtain a better estimate of $\mathbf{x}$ when the poles are closer to the unit circle dwells on the fact that the narrowed spectrum implies a longer correlation time of $\mathbf{x}$, which enables better exploitation of inter-sample dependence.

5. SUMMARY

The XLS criterion is useful for distinguishing model errors from measurement errors, and can be minimized by the ACM algorithm, and by other, computationally more efficient algorithms developed in [8]. Under some general conditions, the deterministic XLS criterion coincides with the statistical JMAP-ML criterion. The JMAP-ML estimator (the XLS estimator with specific weights) tends towards certain parameters constellations relative to the ML estimator. Thus, with long data records, when ML is unbiased, JMAP-ML is biased. However, it is shown in [8], that with short data records JMAP-ML can outperforms ML in terms of both bias and variance and, moreover, that proper selection of weights for the XLS criterion may outperform JMAP-ML.

6. REFERENCES

[1] G.H. Golub and C.F. Van Loan, "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, vol. 17, no. 4, pp. 883–893, 1979.

[2] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis, Frontiers in Applied Mathematics series*, vol. 9, SIAM, Philadelphia, 1991.

[3] S. Van Huffel, Ed., *Recent Advances in Total Least Squares Techniques and Errors-In-Variables Modeling*, SIAM Proceedings Series. SIAM, Philadelphia, 1997.

[4] E.M. Dowling, R.D. DeGroat, and D.A. Linebarger, "Total least squares with linear constrains," *Proc. ICASSP-92*, vol. 5, pp. 341–344, 1992.

[5] T.J. Abatzoglou, J.M. Mendel, and G.A. Harada, "The constrained total least squares technique and its application to harmonic superresolution," *IEEE Trans. Signal Processing*, vol. vol. 39, no. 5, pp. 1070–1087, 1991.

[6] J.A. Cadzow, "Total least squares, matrix enhancement, and signal processing," *Digital Signal Processing*, vol. 4, pp. 21–39, 1994.

[7] B. De Moor, "Total least squares for affinely structured matrices and the noisy realization problem," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 3104–3113, 1994.

[8] A. Yeredor, *The Extended Least Squares Criterion for Discriminating Measurement Errors from Model Errors: Algorithms, Applications, Analysis*, Ph.D. thesis, Tel-Aviv University, Faculty of Engineering, Dept. of Electrical Engineering - Systems, 1997.

[9] A.P. Dempster, N.M Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Ann. Royal Stat. Soc.*, vol. Ser.3g, pp. 1–38, 1977.

[10] Y. Bar-Shalom, "Optimal simultaneous state estimation and parameter identification in linear discrete-time systems," *IEEE Trans. Automatic Control*, vol. 17, no. 3, pp. 308–319, 1972.

[11] J.S. Lim and A.V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.