On the Efficiency of Maximum-Likelihood Estimators of Misspecified Models

Mahamadou Lamine Diong, Eric Chaumette and François Vincent
University of Toulouse - ISAE-Supaero, Toulouse, France
Emails: mouhamadou.diong@isae.fr, eric.chaumette@isae.fr, francois.vincent@isae.fr

***Abstract*—The key results on maximum-likelihood (ML) estimation of misspecified models have been introduced by statisticians (P.J. Huber, H. Akaike, H. White, Q. H. Vuong) resorting to a general probabilistic formalism somewhat difficult to rephrase into the formalism widespread in the signal processing literature. In particular, Vuong proposed two misspecified Cramér-Rao bounds (CRBs) to address, respectively, the situation where the true parametric probability model is known, or not known. In this communication, derivations of the existing results on the accuracy of ML estimation of misspecified models are outlined in an easily comprehensible manner. Simple alternative derivations of these two misspecified CRBs based on the seminal work of Barankin (which underlies all the lower bounds introduced in deterministic estimation) are provided. Since two distinct CRBs exist when the true parametric probability model is known, a quasi-efficiency denomination is introduced.**

I. Introduction

Since its introduction by R.A. Fisher in deterministic estimation [1][2], the method of maximum likelihood (ML) estimation has become one of the most widespread used methods of estimation. The ongoing success of ML estimators (MLEs) originates from the fact that, under reasonably general conditions on the probabilistic observation model [1][2], the MLEs are, in the limit of large sample support, Gaussian distributed and consistent. Additionally, if the observation model is Gaussian, some additional asymptotic regions of operation yielding, for a subset of MLEs, Gaussian distributed and consistent estimates, have also been identified at finite sample support [3][4]. However, a fundamental assumption underlying the above classical results on the properties of MLEs is that the probability distribution which determines the behavior of the observations is known to lie within a specified parametric family of probability distributions (the probability model). In other words, the probability model is assumed to be "correctly specified".

Actually, in many (if not most) circumstances, a certain amount of mismatch between the true probability distribution of the observations and the probability model that we assume is present. As a consequence, it is natural to investigate what happens to the properties of MLE if the probability model is misspecified, i.e. not correctly specified. Huber [5] explored in detail the performance of MLEs under very general assumptions on misspecification, and proved consistency, normality, and derived the MLEs asymptotic covariance that is often referred to as the Huber's "sandwich" covariance in literature. However, Hubert did not explicitly discuss the information theoretic interpretation of this limit. This interpretation has been emphasized by Akaike [6] who has observed that when the true distribution is unknown, the MLE is a natural estimator for the parameters which minimizes the Kullback-Leibler information criterion (KLIC) between the true and the assumed probability model. Then White [7] provided simple conditions under which the MLE is a strongly consistent estimator for the parameter vector which minimize the KLIC. While not as general as Huber's conditions, White's conditions are however sufficiently general to have broad applicability. Lastly, Q. H. Vuong [8] proposed two misspecified Cramér-Rao bounds (CRBs) to address, respectively, the situation where the true parametric probability model is known, or not known, under a general probabilistic formalism involving regular and semi-regular parametric models.

Therefore, the purpose of this communication is twofold. Firstly, in order to foster the understanding of the works of Huber, Akaike, and White on misspecified MLEs [5][6][7], derivations of the key results are outlined in an easily comprehensible manner. Secondly, following the lead of Barankin's seminal work in deterministic estimation [9], simple alternative derivations of the two misspecified CRBs introduced by Vuong are put forward. As a by-product, the misspecified CRB proposed by Vuong [8, Theorem 4.1] when the true parametric probability model is unknown, is a least-upper CRB in the Barankin sense, which coincides with the Huber's "sandwich" covariance, and so called misspecified CRB under ML constraints in [10, (42)] or misspecified CRB for misspecified-unbiased estimators in [11, (5)]. Last, since two distinct misspecified CRBs exist when the true parametric probability model is known, a quasi-efficiency denomination is introduced.

A. Notations and assumptions

Let $\mathbf{x}_l$ be a M-dimensional complex random vector representing the outcome of a random experiment (i.e., the observation vector) whose probability density function (p.d.f.) is known to belong to a family $\mathcal{P}$. A *structure* S is a set of hypotheses, which implies a unique p.d.f. in $\mathcal{P}$ for $\mathbf{x}_l$. Such p.d.f. is indicated with $p(\mathbf{x}_l; S)$. The set of all the a priori possible structures is called a probability model [2]. We assume that the p.d.f. of the random vector $\mathbf{x}_l$ has a parametric representation, i.e., we assume that every structure S is parameterized by a

This work has been partially supported by the DGA/MRIS (2015.60.0090.00.470.75.01).

ISBN 978-0-9928626-7-1 © EURASIP 2017

P-dimensional vector $\boldsymbol{\psi}$, that is $p(\mathbf{x}_l;\boldsymbol{\psi}) \triangleq p(\mathbf{x}_l; S(\boldsymbol{\psi}))$, and that the model is described by a compact subspace $\mathcal{U} \subset \mathbb{R}^P$. In the following, we consider L i.i.d. observations, $\{\mathbf{x}_l\}_{l=1}^L$, for which the true parametric p.d.f. denoted $p_{\boldsymbol{\delta}}(\mathbf{x}_l) \triangleq p(\mathbf{x}_l;\boldsymbol{\delta})$, $\boldsymbol{\delta} \in \mathcal{U}_{\boldsymbol{\delta}} \subset \mathbb{R}^{P_\delta}$, and the assumed parametric p.d.f. denoted $f_{\boldsymbol{\theta}}(\mathbf{x}_l) \triangleq f(\mathbf{x}_l;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathcal{U}_{\boldsymbol{\theta}} \subset \mathbb{R}^{P_\theta}$, belong to two (generally different) families of p.d.f.'s, $\mathcal{P}_{\boldsymbol{\delta}}$ and $\mathcal{F}_{\boldsymbol{\theta}}$. Let us denote: $E_{\boldsymbol{\theta}}[\mathbf{g}(\overline{\mathbf{x}})] \triangleq E_{f_{\boldsymbol{\theta}}}[\mathbf{g}(\overline{\mathbf{x}})] = \int \mathbf{g}(\overline{\mathbf{x}}) f_{\boldsymbol{\theta}}(\overline{\mathbf{x}}) d\overline{\mathbf{x}}$, $E_{\boldsymbol{\delta}}[\mathbf{g}(\overline{\mathbf{x}})] \triangleq E_{p_{\boldsymbol{\delta}}}[\mathbf{g}(\overline{\mathbf{x}})] = \int \mathbf{g}(\overline{\mathbf{x}}) p_{\boldsymbol{\delta}}(\overline{\mathbf{x}}) d\overline{\mathbf{x}}$, where $\overline{\mathbf{x}}^T = (\mathbf{x}_1^T, \ldots, \mathbf{x}_L^T)$, $f_{\boldsymbol{\theta}}(\overline{\mathbf{x}}) = \prod_{l=1}^L f_{\boldsymbol{\theta}}(\mathbf{x}_l)$ and $p_{\boldsymbol{\delta}}(\overline{\mathbf{x}}) = \prod_{l=1}^L p_{\boldsymbol{\delta}}(\mathbf{x}_l)$. If the true model is unknown or not needed, i.e., we do not have or do not need prior information on the particular parameterization of the true distribution, we refer to $p_{\boldsymbol{\delta}}(\overline{\mathbf{x}})$ and $\mathcal{P}_{\boldsymbol{\delta}}$ only as $p(\overline{\mathbf{x}}) = \prod_{l=1}^L p(\mathbf{x}_l)$ and $\mathcal{P}$, respectively, and we denote: $E_p[\mathbf{g}(\overline{\mathbf{x}})] = \int \mathbf{g}(\overline{\mathbf{x}}) p(\overline{\mathbf{x}}) d\overline{\mathbf{x}}$.

II. ML ESTIMATION OF MISSPECIFIED MODELS

As mentioned in the introduction, several authors [5][6][7] have contributed to show that, under mild regularity conditions given in [7] (and summarized in [11, Section II.A]), the misspecified MLE (MMLE) defined as:

$$\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) = \arg\max_{\boldsymbol{\theta}} \{f_{\boldsymbol{\theta}}(\overline{\mathbf{x}})\} = \arg\max_{\boldsymbol{\theta}} \{\ln f_{\boldsymbol{\theta}}(\overline{\mathbf{x}})/L\}, \quad (1a)$$

is, in the limit of large sample support ($L \to \infty$), a strongly consistent estimator for the parameter vector $\boldsymbol{\theta}_f$[1] which minimizes the KLIC:

$$\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) \overset{a.s.}{\to} \boldsymbol{\theta}_f = \arg\min_{\boldsymbol{\theta}} \{D(p||f_{\boldsymbol{\theta}})\}, \; D(p||f_{\boldsymbol{\theta}}) = E_p[\ln(p(\mathbf{x}_l)/f_{\boldsymbol{\theta}}(\mathbf{x}_l))]. \quad (1b)$$

Indeed, as noticed in [6], since $\ln f_{\boldsymbol{\theta}}(\overline{\mathbf{x}})/L = \sum_{l=1}^L \ln f_{\boldsymbol{\theta}}(\mathbf{x}_l)/L \overset{a.s.}{\to} E_p[\ln(f_{\boldsymbol{\theta}}(\mathbf{x}_l))]$ (strong law of large numbers), $\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}})$ is in general a natural estimator of:

$$\boldsymbol{\theta}_f = \arg\max_{\boldsymbol{\theta}} \{E_p[\ln(f_{\boldsymbol{\theta}}(\mathbf{x}_l))]\} = \arg\min_{\boldsymbol{\theta}} \{E_p[\ln(p(\mathbf{x}_l)/f_{\boldsymbol{\theta}}(\mathbf{x}_l))]\},$$

which can be proved to be strongly consistent under mild regularity conditions given in [7]. Therefore the gradient of the ML objective function (1a) can be well approximated via a first order Taylor series expansion about $\boldsymbol{\theta}_f$:

$$\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) \overset{a.s.}{\to} \boldsymbol{\theta}_f - \left[\frac{\partial^2 \ln f(\overline{\mathbf{x}};\boldsymbol{\theta}_f)}{L\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1} \frac{\partial \ln f(\overline{\mathbf{x}};\boldsymbol{\theta}_f)}{L\partial\boldsymbol{\theta}}, \quad (2a)$$

which, in the limit of large sample support, yields:

$$\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) \overset{a.s.}{\to} \boldsymbol{\theta}_f - \mathbf{W}(\boldsymbol{\theta}_f)^{-1} \frac{\partial \ln f(\overline{\mathbf{x}};\boldsymbol{\theta}_f)}{L\partial\boldsymbol{\theta}}, \quad \mathbf{W}(\boldsymbol{\theta}) = E_p\left[\frac{\partial^2 \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right], \quad (2b)$$

[1]The value $\boldsymbol{\theta}_f$ is called the pseudo-true parameter of $\boldsymbol{\theta}$ for the model $f_{\boldsymbol{\theta}}(\mathbf{x}_l)$ when $p(\mathbf{x}_l)$ is the true p.d.f..

following from similar argument given by Cramér [1, pp. 500-503][7]. Hence $\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}})$ is asymptotically normal [7, Th. 3.2]:

$$\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) \overset{A}{\sim} \mathcal{N}(\mathbf{m}_{\widehat{\boldsymbol{\theta}}}, \mathbf{C}_{\widehat{\boldsymbol{\theta}}}), \quad (3a)$$

$$\mathbf{m}_{\widehat{\boldsymbol{\theta}}} \to \boldsymbol{\theta}_f - \mathbf{W}(\boldsymbol{\theta}_f)^{-1} E_p\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}}\right], \quad (3b)$$

$$\mathbf{C}_{\widehat{\boldsymbol{\theta}}} \to \mathbf{C}(\boldsymbol{\theta}_f) = \mathbf{W}(\boldsymbol{\theta}_f)^{-1} \mathbf{C}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_f) \mathbf{W}(\boldsymbol{\theta}_f)^{-1}, \quad (3c)$$

where $\mathbf{C}_{\widehat{\boldsymbol{\theta}}} = E_p\left[\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \mathbf{m}_{\widehat{\boldsymbol{\theta}}}\right)\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \mathbf{m}_{\widehat{\boldsymbol{\theta}}}\right)^T\right]$, $\mathbf{m}_{\widehat{\boldsymbol{\theta}}} = E_p\left[\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}})\right]$, $\boldsymbol{\zeta} \triangleq \boldsymbol{\zeta}(\overline{\mathbf{x}};\boldsymbol{\theta}) = \frac{\partial \ln f(\overline{\mathbf{x}};\boldsymbol{\theta})}{L\partial\boldsymbol{\theta}}$ and:

$$L\mathbf{C}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}) = E_p\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right] - E_p\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] E_p\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right]. \quad (3d)$$

In the particular case of the MMLE (1a-1b) and under the regularity conditions summarized in [11, Section II.A], $\boldsymbol{\theta}_f$ is an interior point of $\mathcal{U}_{\boldsymbol{\theta}}$, i.e. a local minimum of divergence $D(p||f_{\boldsymbol{\theta}})$ which satisfies:

$$\boldsymbol{\theta}_f = \arg\min_{\boldsymbol{\theta}} \{D(p||f_{\boldsymbol{\theta}})\} = \arg\max_{\boldsymbol{\theta}} \{E_p[\ln(f_{\boldsymbol{\theta}}(\mathbf{x}_l))]\} = \arg_{\boldsymbol{\theta}} \left\{E_p\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] = \mathbf{0}\right\}. \quad (4a)$$

Then:

$$E_p\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}}\right] = \mathbf{0}, \; \mathbf{m}_{\widehat{\boldsymbol{\theta}}} = E_p\left[\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}})\right] = \boldsymbol{\theta}_f, \quad \mathbf{C}_{\widehat{\boldsymbol{\theta}}} = E_p\left[\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \boldsymbol{\theta}_f\right)\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \boldsymbol{\theta}_f\right)^T\right], \quad (4b)$$

and the asymptotic covariance matrix (3c) can be further simplified and reduces to the Huber's "sandwich" covariance:

$$\mathbf{C}_H(\boldsymbol{\theta}_f) = \frac{\mathbf{W}(\boldsymbol{\theta}_f)^{-1}}{L} \times E_p\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}} \frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}^T}\right] \mathbf{W}(\boldsymbol{\theta}_f)^{-1}. \quad (4c)$$

Last, since any covariance matrix satisfies the covariance inequality [2], that is $\forall \boldsymbol{\eta}(\overline{\mathbf{x}})$:

$$\mathbf{C}_{\widehat{\boldsymbol{\theta}}} \geq E_p\left[\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \boldsymbol{\theta}_f\right)\boldsymbol{\eta}(\overline{\mathbf{x}})^T\right] E_p\left[\boldsymbol{\eta}(\overline{\mathbf{x}})\boldsymbol{\eta}(\overline{\mathbf{x}})^T\right]^{-1} \times E_p\left[\boldsymbol{\eta}(\overline{\mathbf{x}})\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \boldsymbol{\theta}_f\right)^T\right], \quad (5a)$$

consequently, according to (3c), almost surely $\forall \boldsymbol{\eta}(\overline{\mathbf{x}})$:

$$\mathbf{C}_H(\boldsymbol{\theta}_f) \geq E_p\left[\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \boldsymbol{\theta}_f\right)\boldsymbol{\eta}(\overline{\mathbf{x}})^T\right] \times E_p\left[\boldsymbol{\eta}(\overline{\mathbf{x}})\boldsymbol{\eta}(\overline{\mathbf{x}})^T\right]^{-1} E_p\left[\boldsymbol{\eta}(\overline{\mathbf{x}})\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \boldsymbol{\theta}_f\right)^T\right], \quad (5b)$$

which is referred to as the Huber's "sandwich" (covariance) inequality in the literature on misspecified models.

ISBN 978-0-9928626-7-1 © EURASIP 2017

III. CRB FOR MISSPECIFIED MODELS IN THE BARANKIN SENSE

When the true parametric model $p(\overline{\mathbf{x}};\boldsymbol{\delta})$ is known and assumed to be correctly specified, all the lower bounds introduced in deterministic estimation [12] have been derived from the seminal work of Barankin. In [9] Barankin established the general form of the greatest lower bound on any sth absolute central moment ($s > 1$) of a uniformly unbiased estimator with respect to $p(\overline{\mathbf{x}};\boldsymbol{\delta})$, generalizing the earlier works of Cramér, Rao and Battacharayya on locally unbiased estimators. Barankin showed [9, Section 6], among other things, that the definition of the CRB can be generalized to any absolute moment as the limiting form of the Hammersley-Chapman-Robbins bound (HaChRB). The general results introduced by Barankin require not only the knowledge of the parameterization of the true distribution $p(\overline{\mathbf{x}};\boldsymbol{\delta})$ but also the formulation of a uniform unbiasedness constraint of the form:

$$E_{\boldsymbol{\delta}}\left[\widehat{\mathbf{g}}(\overline{\mathbf{x}})\right] = \mathbf{g}(\boldsymbol{\delta}),\ \forall \boldsymbol{\delta} \in \mathcal{U}_{\boldsymbol{\delta}}, \tag{6a}$$

if one wants to derive lower bounds on the MSE of an unbiased estimate $\widehat{\mathbf{g}}(\overline{\mathbf{x}})$ of the vector $\mathbf{g}(\boldsymbol{\delta})$ of functions of $\boldsymbol{\delta}$. Since the MMLE $\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}})$ (1a-1b) is in the limit of large sample support an unbiased estimate of $\boldsymbol{\theta}_f$ (4b), and since there exists an implicit relationship between $\boldsymbol{\theta}_f$ and $\boldsymbol{\delta}$ (4a):

$$\boldsymbol{\theta}_f(\boldsymbol{\delta}) = \underset{\boldsymbol{\theta}}{\arg}\left\{E_{\boldsymbol{\delta}}\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}\right\}, \tag{6b}$$

the form of (6a) of interest is therefore:

$$E_{\boldsymbol{\delta}}\left[\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}})\right] = \boldsymbol{\theta}_f(\boldsymbol{\delta}),\ \forall \boldsymbol{\delta} \in \mathcal{U}_{\boldsymbol{\delta}}. \tag{6c}$$

Then, according to [9] and in the particular case of the MSE (absolute moment of order 2), the misspecified CRB (MCRB) for unbiased estimates of the pseudo-true parameter $\boldsymbol{\theta}_f$ is defined, for any selected value $\boldsymbol{\delta}^0$, as the limiting form of the HaChRB obtained from the covariance inequality (5a) where [13]:

$$\boldsymbol{\eta}(\overline{\mathbf{x}})^T = \left(1, \frac{p\left(\overline{\mathbf{x}};\boldsymbol{\delta}^0 + \mathbf{u}_1 d\delta\right)}{p\left(\overline{\mathbf{x}};\boldsymbol{\delta}^0\right)}, \ldots, \frac{p\left(\overline{\mathbf{x}};\boldsymbol{\delta}^0 + \mathbf{u}_{P_\delta} d\delta\right)}{p\left(\overline{\mathbf{x}};\boldsymbol{\delta}^0\right)}\right),$$

$$E_{\boldsymbol{\delta}^0}\left[\left(\widehat{\boldsymbol{\theta}}(\overline{\mathbf{x}}) - \boldsymbol{\theta}_f^0\right)\boldsymbol{\eta}(\overline{\mathbf{x}})^T\right] = \left[\mathbf{0}\ \ \boldsymbol{\theta}_f\left(\boldsymbol{\delta}^0 + \mathbf{u}_1 d\delta\right) - \boldsymbol{\theta}_f^0\ \ \ldots\ \ \boldsymbol{\theta}_f\left(\boldsymbol{\delta}^0 + \mathbf{u}_{P_\delta} d\delta\right) - \boldsymbol{\theta}_f^0\right],$$

$\boldsymbol{\theta}_f^0 = \boldsymbol{\theta}_f\left(\boldsymbol{\delta}^0\right)$, $\mathbf{u}_p$ is the pth column of the identity matrix $\mathbf{I}_{P_\delta}$ and $d\delta \to 0$, leading to $\mathbf{C}_{\widehat{\boldsymbol{\theta}}} \geq \mathbf{MCRB}_{\boldsymbol{\delta}^0}$ where:

$$\mathbf{MCRB}_{\boldsymbol{\delta}^0} = \frac{\partial \boldsymbol{\theta}_f\left(\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T} \times E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln p\left(\overline{\mathbf{x}};\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}} \frac{\partial \ln p\left(\overline{\mathbf{x}};\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right]^{-1} \frac{\partial \boldsymbol{\theta}_f\left(\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}. \tag{7}$$

$\partial \boldsymbol{\theta}_f(\boldsymbol{\delta})/\partial \boldsymbol{\delta}^T$ can be easily obtained using the following implicit function theorem [14, Theorem 9.28]. Let $\mathbf{h}(\boldsymbol{\theta},\boldsymbol{\delta}) = \left(h_1(\boldsymbol{\theta},\boldsymbol{\delta}), \cdots, h_{P_\theta}(\boldsymbol{\theta},\boldsymbol{\delta})\right)^T$ be a function of $\mathbb{R}^{P_\theta} \times \mathbb{R}^{P_\delta} \to \mathbb{R}^{P_\theta}$. Let us assume the following: A1) $h_p(\boldsymbol{\theta},\boldsymbol{\delta})$ for $p = 1,\ldots,P_\theta$ are differentiable functions on a neighborhood of the point $\left(\boldsymbol{\theta}^0,\boldsymbol{\delta}^0\right)$ in $\mathbb{R}^{P_\theta} \times \mathbb{R}^{P_\delta}$, A2) $\mathbf{h}\left(\boldsymbol{\theta}^0,\boldsymbol{\delta}^0\right) = \mathbf{0}$, A3) the $P_\theta \times P_\theta$ Jacobian matrix of $\mathbf{h}(\boldsymbol{\theta},\boldsymbol{\delta})$ with respect to $\boldsymbol{\theta}$ is nonsingular at $\left(\boldsymbol{\theta}^0,\boldsymbol{\delta}^0\right)$. Then, there is a neighborhood Δ of the point $\boldsymbol{\delta}^0$ in $\mathbb{R}^{P_\delta}$, there is a neighborhood Θ of the point $\boldsymbol{\theta}^0$ in $\mathbb{R}^{P_\theta}$, and there is a unique mapping $\boldsymbol{\varphi}: \Delta \to \Theta$ such that $\boldsymbol{\theta}^0 = \boldsymbol{\varphi}\left(\boldsymbol{\delta}^0\right)$ and $\mathbf{h}\left(\boldsymbol{\varphi}(\boldsymbol{\delta}),\boldsymbol{\delta}\right) = \mathbf{0}$ for all $\boldsymbol{\delta}$ in Δ. Furthermore, $\boldsymbol{\varphi}(\boldsymbol{\delta})$ is differentiable at $\boldsymbol{\delta}^0$ and satisfies:

$$\boldsymbol{\varphi}(\boldsymbol{\delta}) - \boldsymbol{\varphi}\left(\boldsymbol{\delta}^0\right) = -\left(\frac{\partial \mathbf{h}\left(\boldsymbol{\theta}^0,\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\theta}^T}\right)^{-1} \frac{\partial \mathbf{h}\left(\boldsymbol{\theta}^0,\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\left(\boldsymbol{\delta} - \boldsymbol{\delta}^0\right) + o\left(\left\|\boldsymbol{\delta} - \boldsymbol{\delta}^0\right\|\right).$$

In the case addressed (6b): $\mathbf{h}(\boldsymbol{\theta},\boldsymbol{\delta}) = E_{\boldsymbol{\delta}}\left[\frac{\partial \ln f(\mathbf{x}_l;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$ and $\boldsymbol{\varphi}(\boldsymbol{\delta}) = \boldsymbol{\theta}_f(\boldsymbol{\delta})$. Then:

$$\frac{\partial \boldsymbol{\theta}_f\left(\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T} = -\mathbf{W}\left(\boldsymbol{\theta}_f^0\right)^{-1} \times E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}} \frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right],$$

and (7) becomes:

$$\mathbf{MCRB}_{\boldsymbol{\delta}^0} = \mathbf{W}\left(\boldsymbol{\theta}_f^0\right)^{-1} E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}} \frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right] \times \frac{1}{L} E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}} \frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right]^{-1} \times E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}} \frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}^T}\right] \mathbf{W}\left(\boldsymbol{\theta}_f^0\right)^{-1}. \tag{8}$$

Clearly the Barankin approach provides a simpler alternative derivation of (8) than the one earlier proposed by Vuong [8] (Theorem 3.1), under a general probabilistic formalism involving regular and semi-regular parametric models, and somewhat difficult to follow. Moreover, by the covariance inequality:

$$E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}} \frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}^T}\right] \geq E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}} \frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right] \times E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}} \frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right]^{-1} \times E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}} \frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}^T}\right] \tag{9a}$$

with equality iff [2]:

$$\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}} = E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}{\partial \boldsymbol{\theta}}\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right]\times$$
$$E_{\boldsymbol{\delta}^0}\left[\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}}\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}^T}\right]^{-1}\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{\partial \boldsymbol{\delta}}. \quad (9b)$$

Therefore, when $p\left(\overline{\mathbf{x}};\boldsymbol{\delta}\right)$ is known, one can assert that in most cases:

$$\mathbf{C}_{\widehat{\boldsymbol{\theta}}} \rightarrow \mathbf{C}_H\left(\boldsymbol{\theta}_f^0\right) > \mathbf{MCRB}_{\boldsymbol{\delta}^0}, \quad (10)$$

in other words, in most cases, $\widehat{\boldsymbol{\theta}}$ is no longer an efficient estimator of $\boldsymbol{\theta}_f$ (in comparison with the correctly specified case [1]). Furthermore, if $p\left(\overline{\mathbf{x}};\ \right)$ and $f\left(\overline{\mathbf{x}};\ \right)$ share the same parameterization, i.e. $p\left(\overline{\mathbf{x}};\boldsymbol{\theta}\right)$ and $f\left(\overline{\mathbf{x}};\boldsymbol{\theta}\right)$, then the MSE of $\widehat{\boldsymbol{\theta}}$ is:

$$\mathbf{MSE}_{\widehat{\boldsymbol{\theta}}}\left(\boldsymbol{\theta}^0\right) = E_{p_{\boldsymbol{\theta}^0}}\left[\left(\widehat{\boldsymbol{\theta}}\left(\overline{\mathbf{x}}\right)-\boldsymbol{\theta}^0\right)\left(\widehat{\boldsymbol{\theta}}\left(\overline{\mathbf{x}}\right)-\boldsymbol{\theta}^0\right)^T\right], \quad (11a)$$

and, in the limit of large sample support, is given by:

$$\mathbf{MSE}_{\widehat{\boldsymbol{\theta}}}\left(\boldsymbol{\theta}^0\right) \rightarrow \mathbf{C}_H\left(\boldsymbol{\theta}_f^0\right) + \left(\boldsymbol{\theta}_f^0-\boldsymbol{\theta}^0\right)\left(\boldsymbol{\theta}_f^0-\boldsymbol{\theta}^0\right)^T. \quad (11b)$$

Therefore, when the true parametric model is known, one can assert that in most cases, the MMLE is not a consistent estimator of $\boldsymbol{\theta}$, and whenever it is consistent, it is not an efficient estimator of $\boldsymbol{\theta}$, which contrasts with the behavior of MLEs, since, if the MLEs are consistent then they are also asymptotically efficient [3].

IV. Least-upper CRB for misspecified models in the Barankin sense

If the true model is unknown, i.e., we do not have prior information on the particular parameterization of the true distribution $p\left(\mathbf{x}_l\right)$, the formulation of uniform unbiasedness (6c) is no longer possible. However, the Barankin approach can still be used by building on Vuong's work where it is shown [8, Theorem 4.1] that, under mild regularity conditions summarized in [11, Section II.A], the following surrogate parametric model:

$$\widetilde{p}_{\boldsymbol{\theta}}\left(\mathbf{x}_l\right) \triangleq \widetilde{p}\left(\mathbf{x}_l;\boldsymbol{\theta}\right) = \frac{p\left(\mathbf{x}_l\right)}{c\left(\boldsymbol{\theta}\right)}\left(1+\exp\left(1-\frac{f\left(\mathbf{x}_l;\boldsymbol{\theta}\right)}{f\left(\mathbf{x}_l;\boldsymbol{\theta}_f\right)}\right)\right), \quad (12)$$

where $c\left(\boldsymbol{\theta}\right)$ is a normalizing constant, is a locally least favorable true parametric model in the MSE sense. Indeed, the minimization of the KLIC $D\left(\widetilde{p}_{\boldsymbol{\theta}}||f_{\boldsymbol{\theta}}\right)$ (1b) at the vicinity of $\boldsymbol{\theta}_f$ yields a locally unbiased estimator of $\boldsymbol{\theta}_f$, allowing for the derivation of the MCRB (8) associated with $\widetilde{p}\left(\mathbf{x}_l;\boldsymbol{\theta}\right)$, which satisfies (9b) at $\boldsymbol{\theta}_f$ since [8, (A.62)]: $\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f\right)/\partial\boldsymbol{\theta} = \alpha\partial \ln \widetilde{p}\left(\mathbf{x}_l;\boldsymbol{\theta}_f\right)/\partial\boldsymbol{\theta}$, $\alpha = -2$. Thus the MCRB associated with $\widetilde{p}\left(\mathbf{x}_l;\boldsymbol{\theta}\right)$ coincides at $\boldsymbol{\theta}_f$ with the Huber's "sandwich" covariance $\mathbf{C}_H\left(\boldsymbol{\theta}_f\right)$ (4c). Therefore, by reference to (5b), the Huber's "sandwich" covariance appears to be the least-upper MCRB (LUMCRB) for locally unbiased estimates of the pseudo-true parameter $\boldsymbol{\theta}_f$:

$$\mathbf{LUMCRB}\left(\boldsymbol{\theta}_f\right) = \mathbf{C}_H\left(\boldsymbol{\theta}_f\right), \quad (13)$$

both in Vuong and Barankin senses. Another noteworthy point stressed in [10, Section VII.C] is that, since in the limit of large sample support the MMLE satisfies [10, (57)]:

$$E_p\left[\left(\widehat{\boldsymbol{\theta}}\left(\overline{\mathbf{x}}\right)-\boldsymbol{\theta}_f\right)\frac{\partial \ln f\left(\overline{\mathbf{x}};\boldsymbol{\theta}_f\right)}{L\partial\boldsymbol{\theta}^T}\right] = \frac{-\mathbf{W}\left(\boldsymbol{\theta}_f\right)^{-1}}{L}$$
$$\times E_p\left[\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f\right)}{\partial\boldsymbol{\theta}}\frac{\partial \ln f\left(\mathbf{x}_l;\boldsymbol{\theta}_f\right)}{\partial\boldsymbol{\theta}^T}\right],$$

$\mathbf{C}_H\left(\boldsymbol{\theta}_f\right)$ (4c) is also obtained from (5b) where the score function $\boldsymbol{\eta}\left(\overline{\mathbf{x}}\right)$ is defined as $\boldsymbol{\eta}\left(\overline{\mathbf{x}}\right) \triangleq \partial \ln f\left(\overline{\mathbf{x}};\boldsymbol{\theta}_f\right)/\partial\boldsymbol{\theta}$. Finally, the so called MCRB under ML constraints [10, (42)] and the so called MCRB for misspecified-unbiased estimators [8, Theorem 4.1][11, (5)] appear to be, in the Barankin sense, the LUMCRB (13).

A. Quasi-efficiency

Interestingly enough, if the true parametric model $p\left(\mathbf{x}_l;\boldsymbol{\theta}\right)$ is known, the parametric model (12) can still be defined as:

$$\widetilde{p}_{\boldsymbol{\theta}}\left(\mathbf{x}_l\right) = \frac{p\left(\mathbf{x}_l;\boldsymbol{\delta}^0\right)}{c\left(\boldsymbol{\theta}\right)}\left(1+\exp\left(1-\frac{f\left(\mathbf{x}_l;\boldsymbol{\theta}\right)}{f\left(\mathbf{x}_l;\boldsymbol{\theta}_f^0\right)}\right)\right),$$

where $\boldsymbol{\theta}_f^0 = \boldsymbol{\theta}_f\left(\boldsymbol{\delta}^0\right)$, and all the results mentioned above still hold as well. Then the covariance matrix of a locally unbiased estimator of $\boldsymbol{\theta}_f$ may be either equal to the MCRB or to the LUMCRB. In the Barankin sense, the former case defines an efficient estimator. Hence the need to introduce a new denomination for the latter case. We propose to call such an estimator a quasi-efficient estimator. Finally, one can assert that in most cases, the MMLE is not a consistent estimator of $\boldsymbol{\theta}$, and whenever it is consistent, it is only a quasi-efficient estimator of $\boldsymbol{\theta}$. As expected, both the MCRB and the LUMCRB reduce to the usual CRB when the model is known to be correctly specified, i.e. $f\left(\mathbf{x}_l;\boldsymbol{\theta}\right) \triangleq p\left(\mathbf{x}_l;\boldsymbol{\theta}\right)$, since then $\boldsymbol{\theta}_f = \boldsymbol{\theta}$ and $\widetilde{p}\left(\mathbf{x}_l;\boldsymbol{\theta}\right) = p\left(\mathbf{x}_l;\boldsymbol{\theta}\right)$, leading to:

$$\mathbf{MCRB}_{\boldsymbol{\theta}} = \mathbf{LUMCRB}\left(\boldsymbol{\theta}\right) =$$
$$\frac{1}{L}E_{\boldsymbol{\theta}}\left[\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\frac{\partial \ln p\left(\mathbf{x}_l;\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}^T}\right]^{-1} = \mathbf{CRB}_{\boldsymbol{\theta}},$$

and, in the limit of large sample support, any quasi-efficient estimator becomes an efficient estimator. And last but not least, since the MMLEs asymptotic covariance matrix $\mathbf{C}_H\left(\boldsymbol{\theta}_f\right)$ is available (4c), the derivation of additional lower bounds via the Huber's sandwich inequality (5b) may seem questionable. It is probably the reason why misspecified lower bounds have received little consideration in the literature [8][15], except very recently [10][11].

V. An illustrative example

We revisit the problem of the estimation of the variance of Gaussian data in the presence of misspecified mean value proposed in [11, Section III]. Let us assume to have a set of L i.i.d. scalar observations $\overline{\mathbf{x}} = \left(x_1,\ldots,x_L\right)^T$, distributed according to a Gaussian p.d.f. with a known mean value m_x and an unknown variance $\sigma_x^2 \triangleq \theta$, i.e. $p\left(\overline{\mathbf{x}};\theta\right) =$

ISBN 978-0-9928626-7-1 © EURASIP 2017

$p_{\mathcal{N}}(\overline{\mathbf{x}}; m_x \mathbf{1}_L, \theta \mathbf{I}_L)$. Suppose now that the assumed Gaussian p.d.f. is $f(\overline{\mathbf{x}}; \theta) = p_{\mathcal{N}}(\overline{\mathbf{x}}; m\mathbf{1}_L, \theta \mathbf{I}_L)$, so we misspecify the mean value. Then, (1a-1b) become [11, Section III]:

$$\widehat{\theta}(\overline{\mathbf{x}}) = \frac{1}{L}\sum_{l=1}^{L}(x_l - m)^2, \quad \theta_f = \theta + (m_x - m)^2,$$

where $E_p\left[\widehat{\theta}(\overline{\mathbf{x}})\right] = \theta_f$, and the LUMCRB is given by [11, (22)]:

$$LUMCRB(\theta_f) = \frac{2\theta^2}{L} + \frac{4\theta(m_x - m)^2}{L} + (m_x - m)^4. \tag{14}$$

According to (7), since $\partial\theta_f(\theta)/\partial\theta = 1$, the MCRB is simply:

$$MCRB_\theta = \frac{1}{E_\theta\left[\left(\frac{\partial \ln p(\overline{\mathbf{x}};\theta)}{\partial\theta}\right)^2\right]} = CRB_\theta = \frac{2\theta^2}{L}, \tag{15}$$

which exemplify that the MMLE is only a quasi-efficient estimator of θ, as predicted in most cases when the true parametric model is known (10).

REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1946

[2] E. L. Lehmann and G. Casella, *Theory of Point Estimation (2nd ed.)*. Springer, 1998

[3] S. Haykin, J. Litva, and T. J. Shepherd, *Radar Array Processing,* Chapter 4, Springer-Verlag, 1993

[4] A. Renaux, P. Forster, E. Chaumette, and P. Larzabal, "On the high snr cml estimator full statistical characterization", IEEE Trans. on SP, 54(12): 4840-4843, 2006

[5] P. J. Huber, "The behavior of maximum likelihood estimates under non-standard conditions," in Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability, 1967.

[6] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle", in Proceeding of IEEE ISIT, 1973

[7] H. White, "Maximum likelihood estimation of misspecified models," Econometrica, vol. 50, pp. 1-25, Jan. 1982.

[8] Q. H. Vuong, "Cramér-Rao bounds for misspecified models,", working paper 652, Div. of he Humanities and Social Sci., Caltech, Pasadena, USA, 1986

[9] E.W. Barankin, "Locally best unbiased estimates", Ann. Math. Stat., 20(4): 477-501, 1949.

[10] C.D. Richmond and L.L. Horowitz, "Parameter Bounds on Estimation Accuracy Under Model Misspecification", IEEE Trans. on SP, 63(9): 2263-2278, 2015

[11] S. Fortunati, F. Gini, and M. S. Greco, "The misspecified CRB and its application to the scatter matrix estimation in complex elliptically symmetric distributions," in IEEE Trans. Signal Process, 64(9): 2387-2399, 2016

[12] K. Todros and J. Tabrikian, "General Classes of Performance Lower Bounds for Parameter Estimation-Part I: Non-Bayesian Bounds for Unbiased Estimators", IEEE Trans. on IT, 56(10): 5064-5082, 2010.

[13] R. McAulay, E.M. Hofstetter, " Barankin Bounds on parameter estimation", IEEE Trans. on IT, 17(6): 669-676, 1971

[14] W. Rudin, *Principles of Mathematical Analysis*. New York: Mc-Graw-Hill, 1976.

[15] T. B. Fomby and R. C. Hill, *Maximum-Likelihood Estimation of Misspecified Models: Twenty Years Later*. Oxford, U.K.: Elsevier, Ltd., 2003

ISBN 978-0-9928626-7-1 © EURASIP 2017