Generalized Barankin-Type Lower Bounds for Misspecified Models

Mouhamadou Lamine Diong, Eric Chaumette and François Vincent

Université de Toulouse/ISAE-Supaéro

Abstract

When the assumed probability distribution of the observations differs from the true distribution, the model is said to be misspecified. The key results on maximum-likelihood estimation of misspecified models have been introduced in the limit of large sample support and depend on a parameters vector solution of a computationally expensive non-linear optimization problem. As a possible strategy to circumvent these limitations, we extend the approach lately proposed by *Fritsche et al*. It is shown that the lower bound derived in *Fritsche et al* is a representative of a family of lower bounds deriving from a misspecified unbiasedness constraint leading to generalized Barankin-type lower bounds. For future use, we derive the standard representative of the "Small Errors" and "Large Errors" bounds, namely the generalized CRB and the generalized McAulay-Seidman bound.

Background on MLEs under misspecification

- Maximum likelihood estimators (MLEs) are, under reasonably general conditions on the probabilistic observation model, in the limit of large sample support, Gaussian distributed and consistent, *if the probability distribution function (p.d.f.) which determines the behavior of the observations is assumed to be "correctly specified"*.
- ctually, in many (if not most) circumstances, a certain amount of mismatch between the true p.d.f. of the observations denoted $p(\boldsymbol{x}_t)$, $\{\boldsymbol{x}_t\}_{t=1}^{T}$ i.i.d., and the probability model $f_{\boldsymbol{\theta}}(\boldsymbol{x}_t) \triangleq f(\boldsymbol{x}_t|\boldsymbol{\theta})$ that we assume is present.
- As a consequence, it is natural to investigate what happens to the properties of MLEs if the probability model is misspecified, i.e. not correctly specified [Huber, Akaike, White, Vuong].
- Under mild regularity conditions, the misspecified MLE (MMLE) defined as:

$$\widehat{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}) = \arg\max_{\boldsymbol{\theta}} \left\{ f_{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}_1)\dots f_{\boldsymbol{\theta}}(\boldsymbol{x}_T) \right\}, \quad \overline{\boldsymbol{x}}^T = \left(\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_T^T\right), \quad (1)$$

is, in the limit of large sample support ($T \to \infty$), a strongly consistent estimator for the parameters vector which minimizes the KLIC:

$$\widehat{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}) \overset{a.s.}{\to} \boldsymbol{\theta}_f = \arg\min_{\boldsymbol{\theta}} \left\{ E_p\left[\ln\left(p(\boldsymbol{x}_t)\right) - \ln\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_t)\right)\right] \right\}, \quad (2)$$

$p(\overline{\boldsymbol{x}}) = p(\boldsymbol{x}_1)\dots p(\boldsymbol{x}_T)$, $E_p[\boldsymbol{g}(\overline{\boldsymbol{x}})] = \int \boldsymbol{g}(\overline{\boldsymbol{x}})\, p(\overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}}$.
- Moreover $\widehat{\boldsymbol{\theta}}(\overline{\boldsymbol{x}})$ is asymptotically normal: $\widehat{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}) \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\theta}_f, \boldsymbol{C}_{\widehat{\boldsymbol{\theta}}}\right)$, $\boldsymbol{C}_{\widehat{\boldsymbol{\theta}}} \overset{a.s.}{\to} \boldsymbol{C}_{HS}(\boldsymbol{\theta}_f)$, where the asymptotic covariance matrix $\boldsymbol{C}_{HS}(\boldsymbol{\theta}_f)$, the so-called **Huber's "sandwich covariance"**, is given by:

$$T\boldsymbol{C}_{HS}(\boldsymbol{\theta}_f) = E_p\left[\frac{\partial^2 \ln f(\boldsymbol{x}_t|\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1} E_p\left[\frac{\partial \ln f(\boldsymbol{x}_t|\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}}\frac{\partial \ln f(\boldsymbol{x}_t|\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}^T}\right] \times E_p\left[\frac{\partial^2 \ln f(\boldsymbol{x}_t|\boldsymbol{\theta}_f)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}. \quad (3)$$

A covariance matrix is the tightest LB on itself since it satisfies the covariance inequality. Thus $\forall\boldsymbol{\eta}(\overline{\boldsymbol{x}})$:

$$\boldsymbol{C}_{HS}(\boldsymbol{\theta}_f) \geq E_p\left[\left(\widehat{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}) - \boldsymbol{\theta}_f\right)\boldsymbol{\eta}(\overline{\boldsymbol{x}})^T\right] E_p\left[\boldsymbol{\eta}(\overline{\boldsymbol{x}})\boldsymbol{\eta}(\overline{\boldsymbol{x}})^T\right]^{-1} \times E_p\left[\boldsymbol{\eta}(\overline{\boldsymbol{x}})\left(\widehat{\boldsymbol{\theta}}(\overline{\boldsymbol{x}}) - \boldsymbol{\theta}_f\right)^T\right], \quad (4)$$

also called the **Huber's "sandwich" (covariance) inequality**. Note that $\boldsymbol{C}_{HS}(\boldsymbol{\theta}_f)$ (3) is obtained for $\boldsymbol{\eta}(\overline{\boldsymbol{x}}) = \frac{\partial \ln f(\overline{\boldsymbol{x}}|\boldsymbol{\theta}_f)}{T\partial\boldsymbol{\theta}}$ [Richmond - Horowitz].
- However, any lower bound deriving from (4), including (3), depends on $\boldsymbol{\theta}_f \Rightarrow$ its numerical evaluation requires to solve (2) for each value of $\boldsymbol{\theta}$, a procedure suffering from a large computational cost when the dimension of $\boldsymbol{\theta}$ increases $\Rightarrow$ a possible strategy to circumvent these limitations is the alternative approach proposed in [Fritsche et al]

Barankin-Type Lower Bounds for Correctly Specified Models

- We focus on the estimation of a single unknown real deterministic parameter θ, and denote $E_\theta[\mathbf{g}(\overline{\mathbf{x}})] \triangleq E_{f_\theta}[\mathbf{g}(\overline{\mathbf{x}})] = \int \mathbf{g}(\overline{\mathbf{x}})\, f_\theta(\overline{\mathbf{x}})\, d\overline{\mathbf{x}}$.
- 1) The MSE of an estimator $\widehat{\theta^0}$ of θ^0, $\widehat{\theta^0} \triangleq \widehat{\theta^0}(\overline{\mathbf{x}})$, is a norm:

$$MSE_{\theta^0}\left[\widehat{\theta^0}\right] = \left\|\widehat{\theta^0}(\overline{\mathbf{x}}) - \theta^0\right\|_{\theta^0}^2, \quad \langle u(\overline{\mathbf{x}}) \mid v(\overline{\mathbf{x}})\rangle_\theta = E_\theta[u(\overline{\mathbf{x}})\, v(\overline{\mathbf{x}})] \quad (5)$$

2) Uniform unbiasedness, if Θ denotes the parameter space, can be recasted as:

$$\forall\theta \in \Theta : E_\theta\left[\widehat{\theta^0}(\overline{\mathbf{x}})\right] = \theta \Leftrightarrow \left\langle \widehat{\theta^0}(\overline{\mathbf{x}}) - \theta^0 \mid \upsilon_{\theta^0}(\overline{\mathbf{x}};\theta)\right\rangle_{\theta^0} = \theta - \theta^0 \quad (6)$$

where $\upsilon_{\theta^0}(\overline{\mathbf{x}};\theta) = \frac{f_\theta(\overline{\mathbf{x}})}{f_{\theta^0}(\overline{\mathbf{x}})}$ denotes the likelihood ratio (LR).
- $\Rightarrow$ the MVUE is the solution of a norm minimization under linear constraints:

$$\min\left\{\left\|\widehat{\theta^0}(\overline{\mathbf{x}}) - \theta^0\right\|_{\theta^0}^2\right\} \text{ under } \forall\theta \in \Theta : \left\langle \widehat{\theta^0}(\overline{\mathbf{x}}) - \theta^0 \mid \upsilon_{\theta^0}(\overline{\mathbf{x}};\theta)\right\rangle_{\theta^0} = \theta - \theta^0. \quad (7)$$

- All "computable" LBs for *correctly specified models* derive from sets of discrete or integral linear transform of (6) and are obtained from lemma: The problem of the minimization of $\|\mathbf{u}\|^2$ under the K linear constraints $\langle \mathbf{u} \mid \mathbf{c}_k\rangle = v_k$, $k \in [1, K]$, then has the solution:

$$\min\left\{\|\mathbf{u}\|^2\right\} = \mathbf{v}^T\mathbf{R}^{-1}\mathbf{v}, \quad \mathbf{R}_{n,k} = \langle \mathbf{c}_k \mid \mathbf{c}_n\rangle. \quad (8)$$

Generalized Barankin-Type Lower Bounds for Misspecified Models

- If $f_\theta(\overline{\mathbf{x}})$ is the true p.d.f., the MLE $\widehat{\theta_{ML}^0}$ of θ^0 is, in the limit of large sample support, uniformly unbiased with respect to $f_\theta(\overline{\mathbf{x}})$:

$$\forall\theta \in \Theta : E_\theta\left[\widehat{\theta_{ML}^0}(\overline{\mathbf{x}})\right] = \theta. \quad (9)$$

- If $f_\theta(\overline{\mathbf{x}})$ is **not** the true p.d.f. of the observations, then (9) is no longer the uniform unbiasedness constraint (6) but a given linear constraint:

$$\int \widehat{\theta_{ML}^0}(\overline{\mathbf{x}})\, f_\theta(\overline{\mathbf{x}})\, d\overline{\mathbf{x}} = \theta, \quad \int f_\theta(\overline{\mathbf{x}})\, d\overline{\mathbf{x}} = 1. \quad (10)$$

As $f_\theta(\overline{\mathbf{x}})$ is a p.d.f., it makes sense to regard (9-10) as **a misspecification of the uniform unbiasedness property**.
- Then, any estimator $\widehat{\theta^0}$ verifying (9) satisfies,

$$\forall\theta \in \Theta : E_p\left[\left(\widehat{\theta^0}(\overline{\mathbf{x}}) - \theta^0\right)\omega_p(\overline{\mathbf{x}};\theta)\right] = \theta - \theta^0, \quad \omega_p(\overline{\mathbf{x}};\theta) = \frac{f_\theta(\overline{\mathbf{x}})}{p(\overline{\mathbf{x}})}, \quad (11)$$

and (7) becomes:

$$\min\left\{MSE_p\left[\widehat{\theta^0}\right] = \left\|\widehat{\theta^0}(\overline{\mathbf{x}}) - \theta^0\right\|_p^2\right\} \text{ under}$$
$$\forall\theta \in \Theta : \left\langle \widehat{\theta^0}(\overline{\mathbf{x}}) - \theta^0 \mid \omega_p(\overline{\mathbf{x}};\theta)\right\rangle_p = \theta - \theta^0, \quad (12)$$

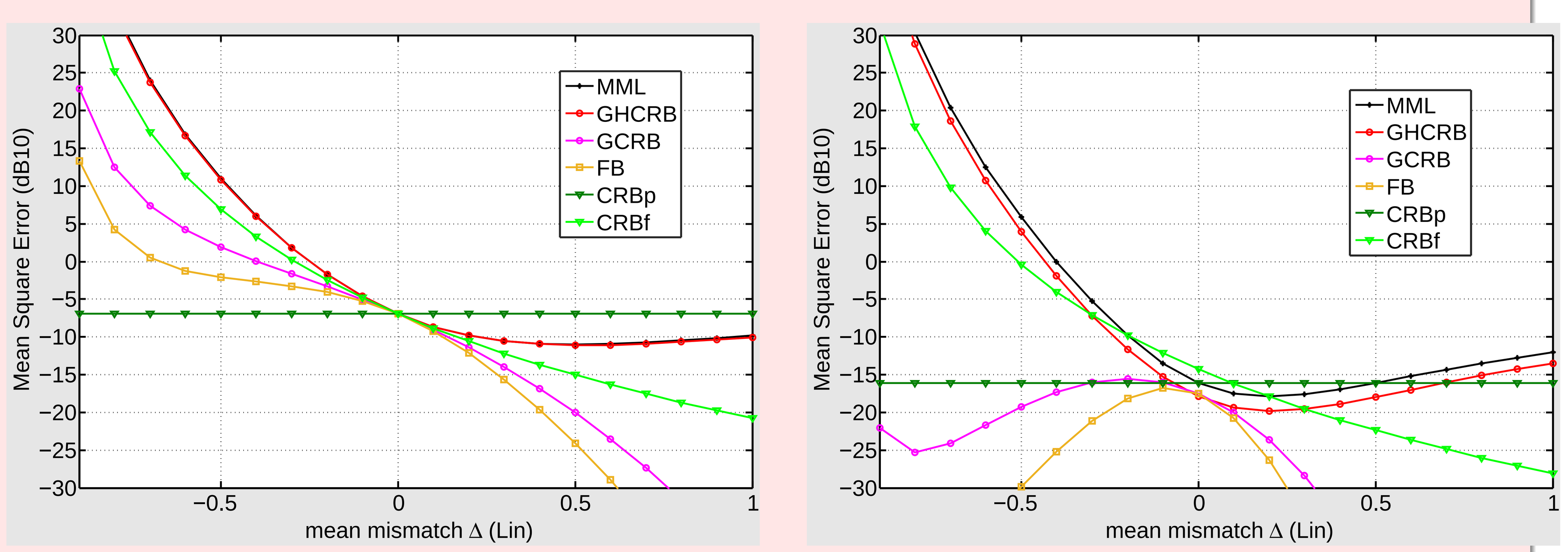
where $\langle u(\overline{\mathbf{x}}) \mid v(\overline{\mathbf{x}})\rangle_p = E_p[u(\overline{\mathbf{x}})\, v(\overline{\mathbf{x}})] \Rightarrow$ **generalization of the BB under unbiasedness misspecification** and an unknown true parametric p.d.f. $p(\overline{\mathbf{x}})$.

Application to Linear Gaussian Models

- The following true linear Gaussian model is considered: $\mathbf{x} = \mathbf{d}_p\theta + \mathbf{n}$, $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_p)$, where $\mathbf{d}_p \in \mathbb{R}^M$ and $\mathbf{C}_p \in \mathbb{R}^{M\times M}$ are supposed to be known. For any selected value θ^0 of the parameter θ, the true p.d.f. of the observation is then $p(\mathbf{x}) \triangleq p_{\theta^0}(\mathbf{x}) \triangleq p_{\mathcal{N}}(\mathbf{x}|\mathbf{d}_p\theta^0, \mathbf{C}_p)$.
- Even if the linear structure and the noise p.d.f. are known, generally $\mathbf{d}_p$ and $\mathbf{C}_p$ are not accurately known and are replaced by assumed values $\mathbf{d}_f$ and $\mathbf{C}_f$, leading to the following assumed p.d.f. $f_\theta(\mathbf{x}) \triangleq p_{\mathcal{N}}(\mathbf{x}|\mathbf{d}_f\theta, \mathbf{C}_f)$.
- We compare using two examples, the MSE of MMLE:

$$\widehat{\theta^0}(\mathbf{x}) = \mathbf{w}_f^T\mathbf{x}, \quad \mathbf{w}_f = \mathbf{C}_f^{-1}\mathbf{d}_f \,/\, \mathbf{d}_f^T\mathbf{C}_f^{-1}\mathbf{d}_f, \quad (13)$$

the Huber's MSE prediction computed from the "sandwich covariance" (3), the *generalized* Hammersley-Chapman-Robbins bound (GHCRB), the *generalized* Cramer-Rao bound (GCRB), the *generalized* bound derived by *Fritsche et al* (referred to as FB) and the CRB associated to the true and assumed p.d.f.s.



$(a) M = 2, C_f = C_p$ $(b) M = 5, C_f = 1.2 \times C_p$.

Figure: MSE vs. Δ of (a) Example 1 and (b) Example 2.

We assume that $\mathbf{d}_p = \mathbf{1}_M$, $\mathbf{d}_f = (1+\Delta)\,\mathbf{d}_p$ where Δ is varied in the interval $[-1, 1]$, $\theta^0 = 1$, and $\mathbf{C}_p = \mathbf{I}_M$ (unit noise power).
- In the first example, $M = 2$, $\mathbf{C}_f = \mathbf{C}_p$, that is the true noise power is accurately known, whereas in the second example, $M = 5$, and the true noise power is known up to a scalar factor, which is assumed to be 1.2: $\mathbf{C}_f = 1.2 \times \mathbf{C}_p$.
- Fig. 1b) exemplify the fact that the standard CRBs no longer provide a lower bound on estimation performance whatever the misspecification considered.
- It appears that the *generalized* "Small Errors" bounds (FB and GCRB) are unlikely to be informative in a large domain of misspecification values $(\Delta, \mathbf{C}_f)$, since they become overly optimistic as soon as the misspecification on $(\Delta, \mathbf{C}_f)$ increases. Fortunately, the behavior of the GHCRB suggests that the use of *generalized* "Large Errors" bounds (*generalized* Barankin-Type LBs) will allow to increase the domain of misspecification values $(\Delta, \mathbf{C}_f)$ where such LBs remain tight enough to be relevant.