On the Performance of Joint Linear Minimum Mean Squared Error (LMMSE) Filtering and Parameter Estimation

Siouar Bensaid and Dirk Slock[†]
EURECOM, Mobile Communications Dept.
Campus SophiaTech, 450 route des Chappes
06410 Biot Sophia Antipolis, FRANCE
Email: {siouar.bensaid,dirk.slock}@eurecom.fr

***Abstract*—We consider the problem of LMMSE estimation (such as Wiener and Kalman filtering) in the presence of a number of unknown parameters in the second-order statistics, that need to be estimated also. This well-known joint filtering and parameter estimation problem has numerous applications. It is a hybrid estimation problem in which the signal to be estimated by linear filtering is random, and the unknown parameters are deterministic. As the signal is random, it can also be eliminated, allowing parameter estimation from the marginal distribution of the data. An intriguing question is then the relative performance of joint vs. marginalized parameter estimation. In this paper, we consider jointly Gaussian signal and data and we first provide contributions to Cramer-Rao bounds (CRBs). We characterize the difference between the Hybrid Fisher Information Matrix (HFIM) and the classical marginalized FIM on the one hand, and between the FIM (with CRB asymptotically attained by ML) and the popular Modified FIM (MFIM, inverse of Modified CRB) which is a loose bound. We then investigate three iterative (alternating optimization) joint estimation approaches: Alternating Maximum A Posteriori for Signal and Maximum Likelihood for parameters (AMAPML), which in spite of a better HFIM suffers from inconsistent parameter bias, Expectation-Maximization (EM) which converges to (marginalized) ML (but with AMAPML signal estimate), and Variational Bayes (VB) which yields an improved signal estimate with the parameter estimate asymptotically becoming ML.**

***Index Terms*—Joint Estimation, Maximum Likelihood (ML), Variational Bayes (VB), Expectation-Maximization (EM), Cramer-Rao Bound (CRB)**

I. Introduction

In estimation theory, the choice of an estimator depends closely on the context of the problem. When the unknown parameters are deterministic, the Maximum Likelihood estimator (ML) is often considered the best approach. It is typically consistent and asymptotically optimal (attaining the CRB)([1], [2]). For the random case, the Minimum Mean Squared Error (MMSE) is used and known (in the Gaussian case) to achieve the Bayesian CRB (BCRB) introduced by Van Trees [3]. When the MMSE estimate is intractable, it is sometimes replaced by the Maximum A posteriori (MAP) estimator. An other important estimation problem is when nuisance random parameters are affecting the estimation of the deterministic unknown parameters/signals like in synchronization problems[4]. Different scenarios have been considered. One scenario is to marginalize out the nuisance parameters which yields the previous problem of ML estimation. In some cases, the marginalization is intractable or very tedious, so we resort to joint estimation (MAP/ML, EM, VB...)([5], [6], [7]), which is also relevant when the random signals are of interest ([8], [9]). Several bounds were developped to evaluate estimator performance in this case. A well-known bound is the MCRB which was introduced for the first time in[10] for a synchronization problem. In [10], the MCRB was derived when the wanted parameter is scalar, then extended to the vectorial case in [11]. In ([10], [11]), the authors prove that the MCRB is looser than the CRB. A characterization of the difference in the scalar case is introduced indirectly in [12]. Though it is looser than the CRB, in some problems ([13], [10], [11]), the MCRB is computationally very interesting since it can be derived in closed form while this may not be possible for the CRB. Hence when they coincide, the use of the MCRB is more practical. To our knowledge, in the litterature on the MCRB, the probability density function (pdf) of the nuisance signals is always assumed to be independent of the deterministic parameters. This hypothesis is relevant in channel estimation applications where the MCRB is applied most of the time, and where the pdf of the transmitted symbols (nuisance parameters) is independent of the channel (synchronization) parameters ([11], [14]). Yet, other applications in statistical signal processing may not fulfill this condition [15]. Then, it would be of interest to find out the influence of this new condition on the MCRB/CRB relation, which we explore in this paper. Another interesting lower bound used for joint estimation is the hybrid CRB (HCRB). The need for the HCRB was expressed perhaps for first time in the work of Rockah and Schultheiss for studying passive source localization in [16]. Ten years later, Reuven and Messer generalized its formulation in [17] where they extend the Barankin bound to the case of an unknown hybrid vector. In [18], the HCRB is proved looser

[†] EURECOM's research is partially supported by its industrial members: BMW Group R&T, iABG, Monaco Telecom, Orange, SAP, SFR, ST Microelectronics, Swisscom, Symantec, and also by the EU FP7 projects WHERE2 and NEWCOM#.

than the CRB but tighter than the MCRB ([19], [14]). Many other bounds were proposed, the reader can refer to [3] for more information. In this paper, we characterize the difference between the HFIM (the deterministic parameter part) and the classical FIM in order to understand the influence of the random signals on optimal parameter estimation. In the next section we present the general framework of joint and separate estimation. In section III, we then characterize the differences (HFIM-FIM) and (FIM-MFIM). In section IV, we analy=ze the performance of the iterative algorithms AMPAML, EM and VB and introduce SOELMMSE.

II. Jointly Gaussian Framework

Let $\boldsymbol{y}$ denote the $N \times 1$ measurement signal on the basis of which we want to estimate the $M \times 1$ random process $\boldsymbol{x}$. However, an $L \times 1$ vector of parameters θ intervenes in the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{x}$ and we have a likelihood function of the form $f(\boldsymbol{y},\boldsymbol{x}|\theta)$ which means MAP for $\boldsymbol{x}$ and ML for θ

$$\begin{aligned} \ln\, f(\boldsymbol{y},\boldsymbol{x}|\theta) &= \ln\, f(\boldsymbol{y}|\boldsymbol{x},\theta) + \ln\, f(\boldsymbol{x}|\theta) \\ &= \ln\, f(\boldsymbol{y}|\theta) + \ln\, f(\boldsymbol{x}|\boldsymbol{y},\theta) \end{aligned} \tag{1}$$

where $\ln f(\boldsymbol{y}|\theta)$ corresponds to the separate loglikelihood for the parameters with elimination of the random $\mathbf{x}$, and $\ln f(\boldsymbol{x}|\boldsymbol{y},\theta)$ is the posterior distribution for $\boldsymbol{x}$. In this paper we consider real quantities and zero means. In the jointly Gaussian zero-mean setting, the whole estimation problem is characterized by the joint covariance matrix $\boldsymbol{R}(\theta)$

$$\boldsymbol{R}(\theta) = \left[\begin{array}{cc} \boldsymbol{R}_{\boldsymbol{xx}}(\theta) & \boldsymbol{R}_{\boldsymbol{xy}}(\theta) \\ \boldsymbol{R}_{\boldsymbol{yx}}(\theta) & \boldsymbol{R}_{\boldsymbol{yy}}(\theta) \end{array}\right] . \tag{2}$$

A. Separate (/Marginalized/ML) Parameter Estimation

The random vector $\boldsymbol{x}$ can be integrated out, leading to marginalized ML estimate for θ, $\widehat{\theta}_{ML} = \arg\max_\theta \ln f(\boldsymbol{y}|\theta)$ with loglikelihood

$$\ln f(\boldsymbol{y}|\theta) = -\frac{1}{2}\ln\det \boldsymbol{R}_{\boldsymbol{yy}}(\theta) - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{R}_{\boldsymbol{yy}}^{-1}(\theta)\boldsymbol{y} \tag{3}$$

This leads to the marginalized Fisher Information Matrix $\boldsymbol{J}(\theta)$

$$= -\,\mathrm{E}_{\boldsymbol{y}|\theta}\frac{\partial^2 \ln f(\boldsymbol{y}|\theta)}{\partial\theta\partial\theta^T} = \frac{1}{2}\boldsymbol{R}_{\boldsymbol{yy},\theta}^T(\boldsymbol{R}_{\boldsymbol{yy}}^{-1} \otimes \boldsymbol{R}_{\boldsymbol{yy}}^{-1})\boldsymbol{R}_{\boldsymbol{yy},\theta} \tag{4}$$

where $\boldsymbol{R}_\theta = \frac{\partial \mathrm{vec}\{\boldsymbol{R}\}}{\partial\theta^T}$, $\mathrm{vec}\{\boldsymbol{R}\}$ is a column vector obtained by stacking the consecutive columns of $\boldsymbol{R}$, and $\boldsymbol{A}\otimes\boldsymbol{B} = [\boldsymbol{A}_{ij}\boldsymbol{B}]$ denotes the Kronecker product of matrices $\boldsymbol{A}$, $\boldsymbol{B}$. Since the ML estimate $\widehat{\theta}_{ML}$ is consistent, its estimation error $\widetilde{\theta}_{ML}$ reaches the Cramer-Rao lower bound (CRB)

$$\boldsymbol{R}^M_{\widetilde{\theta}\widetilde{\theta}}(\theta) = \mathrm{CRB}^M = \boldsymbol{J}^{-1}\ . \tag{5}$$

The CRB (so the FIM) is a function of θ but for notational convenience, we shall not mention it explicitly.

B. Joint (MAPML) Signal (MAP) and Param. (ML) Estimation

In particular, we get for the posterior distribution

$$f(\boldsymbol{x}|\boldsymbol{y},\theta) = \mathcal{N}(\widehat{\boldsymbol{x}}(\theta), \boldsymbol{P}(\theta)) \tag{6}$$

where $\widehat{\boldsymbol{x}}(\theta) = \boldsymbol{F}(\theta)\,\boldsymbol{y}$ is the LMMSE estimate, and

$$\begin{aligned} &\widehat{\boldsymbol{x}}(\theta) = \boldsymbol{F}(\theta)\,\boldsymbol{y}\ ,\ \ \boldsymbol{F}(\theta) = \boldsymbol{R}_{\boldsymbol{xy}}(\theta)\ \boldsymbol{R}_{\boldsymbol{yy}}^{-1}(\theta) \\ &\boldsymbol{P}(\theta) = \boldsymbol{R}_{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}}(\theta) = \boldsymbol{R}_{\boldsymbol{xx}}(\theta) - \boldsymbol{R}_{\boldsymbol{xy}}(\theta)\ \boldsymbol{R}_{\boldsymbol{yy}}^{-1}(\theta)\boldsymbol{R}_{\boldsymbol{yx}}(\theta). \end{aligned} \tag{7}$$

Note that

$$\begin{aligned} \arg\max_{\boldsymbol{x}} \ln\, f(\boldsymbol{x}|\boldsymbol{y},\theta) &= \widehat{\boldsymbol{x}}(\theta) \\ \max_{\boldsymbol{x}} \ln\, f(\boldsymbol{x}|\boldsymbol{y},\theta) &= -\tfrac{1}{2}\ln\det \boldsymbol{P}(\theta)\,. \end{aligned} \tag{8}$$

Hence due to this separability, as noted in [7],

$$\max_{\boldsymbol{x}} \ln\, f(\boldsymbol{y},\boldsymbol{x}|\theta) = \ln\, f(\boldsymbol{y}|\theta) - \frac{1}{2}\ln\det \boldsymbol{P}(\theta) \tag{9}$$

which is the compressed joint likelihood which remains to be optimized w.r.t. θ. The performance for the estimation of θ in the joint estimation problem is governed by the so-called Hybrid CRB (HCRB), so called because of the mix of random and deterministic parameters. If we denote by $\boldsymbol{w} = \left[\theta^T\,\boldsymbol{x}^T\right]^T$ the hybrid vector, the Hybrid Fisher Information Matrix (HFIM) is defined

$$\widetilde{\boldsymbol{J}} = \mathrm{E}\left\{\frac{\partial \ln f\,(\boldsymbol{y},\boldsymbol{x}|\theta)}{\partial \boldsymbol{w}}\,\frac{\partial \ln f\,(\boldsymbol{y},\boldsymbol{x}|\theta)}{\partial \boldsymbol{w}^T}\right\} = \begin{bmatrix} \widetilde{\boldsymbol{J}}_\theta & \widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}} \\ \widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T & \widetilde{\boldsymbol{J}}_{\boldsymbol{x}} \end{bmatrix}$$

The HCRB is computed as the (1,1) block of $\widetilde{\boldsymbol{J}}^{-1}(\theta)$, which results in $\left(\widetilde{\boldsymbol{J}}_\theta - \widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\widetilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T\right)^{-1}$. In [18], the authors prove that the CRB is tighter than the HCRB. They also provide a necessary and sufficient condition on the joint pdf to reach HCRB=CRB. Note however that in terms of performance, the term $-\frac{1}{2}\ln\det \boldsymbol{P}(\theta)$ in (9) is in fact misleading for the estimation of θ and leads to a bias that leads to inconsistency (if M grows with N) [7]. Hence the HCRB is certainly not reached by MAPML.

III. Characterizing the Differences (HFIM-FIM) and (FIM-MFIM)

A. Difference between HFIM and FIM

From the relation between the joint and marginalized distributions in the second line of (1), we compute the Hessian relative to $\boldsymbol{w}$ and after applying the joint expectation, we get

$$\begin{bmatrix} \widetilde{\boldsymbol{J}}_\theta & \widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}} \\ \widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T & \widetilde{\boldsymbol{J}}_{\boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{J} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{G}}_\theta & \tilde{\mathbf{G}}_{\theta,\boldsymbol{x}} \\ \tilde{\mathbf{G}}_{\theta,\boldsymbol{x}}^T & \tilde{\mathbf{G}}_{\boldsymbol{x}} \end{bmatrix} \tag{10}$$

where $\tilde{\mathbf{G}} = -\,\mathrm{E}_{\boldsymbol{x},\boldsymbol{y}|\theta}\frac{\partial^2 \ln f(\boldsymbol{x}|\boldsymbol{y},\theta)}{\partial\boldsymbol{w}\,\partial\boldsymbol{w}^T}$. From (10), we deduce the relation between HCRB^{-1} and CRB^{-1} :

$$\begin{aligned} HCRB^{-1} &= \widetilde{\boldsymbol{J}}_\theta - \widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\widetilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\widetilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T \\ &= \boldsymbol{J} + \tilde{\mathbf{G}}_\theta - \tilde{\mathbf{G}}_{\theta,\boldsymbol{x}}\tilde{\mathbf{G}}_{\boldsymbol{x}}^{-1}\tilde{\mathbf{G}}_{\theta,\boldsymbol{x}}^T \\ &= CRB^{-1} + \tilde{\mathbf{G}}_\theta - \tilde{\mathbf{G}}_{\theta,\boldsymbol{x}}\tilde{\mathbf{G}}_{\boldsymbol{x}}^{-1}\tilde{\mathbf{G}}_{\theta,\boldsymbol{x}}^T \end{aligned} \tag{11}$$

This expression is valid for any distribution and shows that the hybrid (joint) inverse CRB (information in the presence of nuisance parameters) for θ equals the inverse marginal/separate CRB plus an inverse CRB that would correspond to joint estimation from the posterior density $f(\boldsymbol{x}|\boldsymbol{y},\theta)$. In the zero mean Gaussian case, ($\tilde{\mathbf{G}}_{\theta,\boldsymbol{x}} = 0$) and the full FIM is

$$\widetilde{\boldsymbol{J}} \ = \left[\begin{array}{cc} \frac{1}{2}\boldsymbol{R}_\theta^T(\boldsymbol{R}^{-1}\otimes\boldsymbol{R}^{-1})\boldsymbol{R}_\theta & 0 \\ 0 & P^{-1}(\theta) \end{array}\right] .$$

Then the HCRB^{-1} is reduced to $\widetilde{\boldsymbol{J}}_\theta$ and

$$\begin{aligned}\widetilde{\boldsymbol{J}}_\theta &= \tfrac{1}{2}\boldsymbol{R}_\theta^T(\boldsymbol{R}^{-1}\otimes\boldsymbol{R}^{-1})\boldsymbol{R}_\theta = CRB^{-1}+\tilde{\mathbf{G}}_\theta\\ &= \tfrac{1}{2}\boldsymbol{R}_{\boldsymbol{yy},\theta}^T(\boldsymbol{R}_{\boldsymbol{yy}}^{-1}\otimes\boldsymbol{R}_{\boldsymbol{yy}}^{-1})\boldsymbol{R}_{\boldsymbol{yy},\theta}\\ &+\tfrac{1}{2}P_\theta^T(\theta)\left(P^{-1}(\theta)\otimes P^{-1}(\theta)\right)P_\theta(\theta)\\ &+F_\theta^T(\theta)\left(P^{-1}(\theta)\otimes\boldsymbol{R}_{\boldsymbol{yy}}(\theta)\right)F_\theta(\theta).\end{aligned}$$

where the last two terms correspond to the difference in inverse CRB, and correspond to the information for θ that can be extracted from the covariance and the mean of the Gaussian posterior $f(\boldsymbol{x}|\boldsymbol{y},\theta)$.

B. The Difference between FIM and MFIM

In [12], Moeneclaey computed $(MFIM - FIM)$ for the case $f(\boldsymbol{x}|\theta) = f(\boldsymbol{x})$, θ scalar and white Gaussian observation noise. The extension of his result to vectorial θ and general observation noise covariance (but independent of θ) R_{vv} is straightforward and can be written as follows

$$\boldsymbol{J} = \boldsymbol{J}_M - E_{\boldsymbol{y}|\theta} Cov_{\boldsymbol{x}|\boldsymbol{y},\theta}\left\{\frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{x},\theta\right)}{\partial\theta}\right\} \qquad (12)$$

where $\boldsymbol{J}_M = -\mathrm{E}\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\theta)}{\partial\theta\partial\theta^T}$ denotes the MFIM. Here we shall extend this to the case of $f(\boldsymbol{x}|\theta)$. We claim the following result

$$\boldsymbol{J} = \boldsymbol{J}_M - E\left\{-\frac{\partial^2 \ln f\left(\boldsymbol{x}|\boldsymbol{y},\theta\right)}{\partial\theta\partial\theta^T}\right\} + E\left\{-\frac{\partial^2 \ln f\left(\boldsymbol{x}|\theta\right)}{\partial\theta\partial\theta^T}\right\} \qquad (13)$$

Proof. The Hessian of the two lines in (1) relative to θ results in

$$\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\theta)}{\partial\theta\partial\theta^T} + \frac{\partial^2 \ln f\left(\boldsymbol{x}|\theta\right)}{\partial\theta\partial\theta^T} = \frac{\partial^2 \ln f(\boldsymbol{y}|\theta)}{\partial\theta\partial\theta^T} + \frac{\partial^2 \ln f\left(\boldsymbol{x}|\boldsymbol{y},\theta\right)}{\partial\theta\partial\theta^T}$$

Applying the $E\{.\}$ operator over all random variables and changing the terms in the right side results in the claimed result. □

Notice that Moeneclaey's result is a special case of (13). In fact, when $f(\boldsymbol{x}|\theta) = f(\boldsymbol{x})$, the last term in (13) vanishes and the second term can be proved easily equal to the covariance term in (12) when we notice that $E_{\boldsymbol{x}|\boldsymbol{y},\theta}\left\{\frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\theta)}{\partial\theta}\right\}$ is simply equal to $\frac{\partial \ln f(\boldsymbol{y}|\theta)}{\partial\theta}$. In terms of interpretation, this term may be interpreted as the difference in information between $\boldsymbol{x}$ being deterministic or random. The second term is new and corresponds to the information on θ in the prior distribution $f(\boldsymbol{x}|\theta)$.

C. Performance-CRB Comparison

Let $\widehat{\theta}^M$ refer to the ML estimate and $\widehat{\theta}^J$ to $\widehat{\theta}$ in the joint MAPML estimation with $\boldsymbol{x}$. Asymptotically (in the amount of data $\boldsymbol{y}$), we get

$$C^J_{\widetilde{\theta}\widetilde{\theta}} \overset{(i)}{\geq} C^M_{\widetilde{\theta}\widetilde{\theta}} \overset{(ii)}{=} \mathrm{CRB}^M_\theta \overset{(iii)}{\geq} \mathrm{CRB}^J_\theta \qquad (14)$$

where (ii) is due to $\widehat{\theta}^M$ being consistent, (i) is due to the bias in and hence inconsistency of $\widehat{\theta}^J$ (which normally leads to performance degradation), and (iii) was analyzed above, though the operational meaning of CRB^J_θ =HCRB is not clear yet.

IV. Iterative ML Algorithms

A first iterative algorithm is AMAPML (Alternating MAPML) in which the MAPML loglikelihood (1) gets maximized by alternating maximization w.r.t. $\boldsymbol{x}$ and θ.

A. EM Algorithm

The Expectation-Maximization algorithm was introduced to iterate towards the ML estimate while having reduced complexity iterations [5]. At iteration $i+1$ we get for $\widehat{\theta}$

$$\widehat{\theta}^{i+1} = \arg\max_\theta \; \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i} \ln f(\boldsymbol{x},\boldsymbol{y}|\theta) \qquad (15)$$

which is usually spelled out in 2 steps:

$$\begin{aligned}&\text{E step: } \ln q^{i+1}(\theta|\boldsymbol{y}) \doteq \textstyle\int f(\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i)\ln f(\boldsymbol{x},\boldsymbol{y}|\theta)d\boldsymbol{x}\\ &\text{M step: } \widehat{\theta}^{i+1} = \arg\max_\theta \ln q^{i+1}(\theta|\boldsymbol{y})\end{aligned} \qquad (16)$$

where $\doteq$ will denote equality up to ”constants” (which in this instance could be a function of $\boldsymbol{y}$). Depending on the application, some simplifications may occur. For instance if the (e.g. AR) parameters of interest only appear in $f(\boldsymbol{x}|\theta)$, then we can use $\ln f(\boldsymbol{x},\boldsymbol{y}|\theta) = \ln f(\boldsymbol{x}|\theta) + \ln f(\boldsymbol{y}|\boldsymbol{x},\theta) = \ln f(\boldsymbol{x}|\theta) + \ln f(\boldsymbol{y}|\boldsymbol{x})$. Then, apart from additive constants, we get

$$\begin{aligned}&-2\ln q^{i+1}(\theta|\boldsymbol{y}) \doteq \mathrm{tr}\{\widehat{\boldsymbol{R}}^i_{\boldsymbol{xx}}\boldsymbol{R}^{-1}_{\boldsymbol{xx}} - \boldsymbol{I}\} - \ln\det(\widehat{\boldsymbol{R}}^i_{\boldsymbol{xx}}\boldsymbol{R}^{-1}_{\boldsymbol{xx}})\\ &\text{where } \; \widehat{\boldsymbol{R}}^i_{\boldsymbol{xx}} = \widehat{\boldsymbol{x}}(\widehat{\theta}^i)\widehat{\boldsymbol{x}}^T(\widehat{\theta}^i) + \boldsymbol{P}(\widehat{\theta}^i)\end{aligned} \qquad (17)$$

which is the Itakura-Saito distance (ISD) between $\boldsymbol{R}_{\boldsymbol{xx}}(\theta)$ and $\widehat{\boldsymbol{R}}^i_{\boldsymbol{xx}}$. In the general case, (16) leads to ISD minimization between the joint covariance matrix $\boldsymbol{R}(\theta)$ from (2) and $\widehat{\boldsymbol{R}}^i = \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i}\,\boldsymbol{w}\boldsymbol{w}^T$ with $\boldsymbol{w}^T = [\boldsymbol{x}^T\boldsymbol{y}^T]$. Now, using the block UDL factorization

$$\boldsymbol{R} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{F}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{P} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}_{\boldsymbol{yy}}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{F}^T & \boldsymbol{I}\end{bmatrix} \qquad (18)$$

and considering that both tr$\{.\}$ and $\ln\det(.)$ allow cyclic commutation of the factors in their argument, we get

$$\begin{aligned}&\boldsymbol{R}^{-1}(\theta)\;\widehat{\boldsymbol{R}}^i\\ &= \begin{bmatrix}\boldsymbol{P}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}^{-1}_{\boldsymbol{yy}}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & -\boldsymbol{F}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ -\boldsymbol{F}^T & \boldsymbol{I}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{P}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}^{-1}_{\boldsymbol{yy}}\end{bmatrix}\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i}\begin{bmatrix}\widetilde{\boldsymbol{x}}\\ \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\widetilde{\boldsymbol{x}}\\ \boldsymbol{y}\end{bmatrix}^T = \boldsymbol{D}^{-1}(\theta)\,\widehat{\boldsymbol{D}}^i\,,\\ &\text{where } \; \boldsymbol{D}(\theta) = \begin{bmatrix}\boldsymbol{P}(\theta) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}_{\boldsymbol{yy}}(\theta)\end{bmatrix},\\ &\widehat{\boldsymbol{D}}^i = \begin{bmatrix}\boldsymbol{P}(\widehat{\theta}^i) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}\widehat{\boldsymbol{x}}(\widehat{\theta}^i) - \widehat{\boldsymbol{x}}(\theta)\\ \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{x}}(\widehat{\theta}^i) - \widehat{\boldsymbol{x}}(\theta)\\ \boldsymbol{y}\end{bmatrix}^T\end{aligned} \qquad (19)$$

At convergence we get (with $\theta = \widehat{\theta}^\infty$)

$$\boldsymbol{D}^{-1}(\theta)\,\widehat{\boldsymbol{D}}^\infty = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}^{-1}_{\boldsymbol{yy}}(\theta)\,\boldsymbol{y}\boldsymbol{y}^T\end{bmatrix}\,. \qquad (20)$$

Hence this convergence to the ISD between $\boldsymbol{R}_{\boldsymbol{yy}}(\theta)$ and $\boldsymbol{y}\boldsymbol{y}^T$ and hence to the ML loglikelihood. Actually, $\ln q^i(\theta|\boldsymbol{y})$ does not measure exactly the ISD between $\boldsymbol{R}(\theta)$ and $\widehat{\boldsymbol{R}}^i$, but

$$-2\ln q^{i+1}(\theta|\boldsymbol{y}) \doteq \ln\det(\boldsymbol{R}(\theta)) + \mathrm{tr}\{\boldsymbol{R}^{-1}(\theta)\;\widehat{\boldsymbol{R}}^i\} \qquad (21)$$

which also converges to the ML loglikelihood. The difference between $q^{i+1}(\theta|\boldsymbol{y})$ and the ISD is due to the fact that both $\boldsymbol{R}(\theta)$ and $\widehat{\boldsymbol{R}}^i$ depend on θ, leading to the difference term $\ln\det(\widehat{\boldsymbol{R}}^i)$.

B. Variational Bayes (VB) Approach

Even though $\ln f(\boldsymbol{x}|\boldsymbol{y},\theta)$ is quadratic and $\ln f(\theta|\boldsymbol{y})$ is asymptotically quadratic, the joint $\ln f(\boldsymbol{x},\theta|\boldsymbol{y})$ contains products of both quadratic terms and hence is not Gaussian. Variational Bayes is an approach to approximate the true joint posterior pdf by a product form

$$f(\boldsymbol{x},\theta|\boldsymbol{y}) \approx q(\boldsymbol{x}|\boldsymbol{y})\; q(\theta|\boldsymbol{y})\;. \tag{22}$$

The factors in the product are obtained by minimizing the Kullback-Leibler distance between the two sides of (22), leading to implicit equations that can be iterated:

$$\begin{array}{rcl} \ln\, q^{i+1}(\theta|\boldsymbol{y}) & \doteq & \int q^i(\boldsymbol{x}|\boldsymbol{y})\; \ln\, f(\boldsymbol{x},\theta,\boldsymbol{y})\, d\boldsymbol{x} \\ \ln\, q^{i+1}(\boldsymbol{x}|\boldsymbol{y}) & \doteq & \int q^{i+1}(\theta|\boldsymbol{y})\; \ln\, f(\boldsymbol{x},\theta,\boldsymbol{y})\, d\theta \end{array} \tag{23}$$

which can be solved iteratively. Apart from approximating the true posterior pdf by a factored form, one can furthermore require the factors to be of a certain parametric form. In the case considered here however, $q(\boldsymbol{x}|\boldsymbol{y})$ is automatically Gaussian, whereas we shall force $q(\theta|\boldsymbol{y})$ to be Gaussian. This is done by taking the mean and covariance of the RHS in (23). Note that asymptotically, $q(\theta|\boldsymbol{y})$ becomes Gaussian automatically. Also note that $q(\boldsymbol{x}|\boldsymbol{y})$ and $q(\theta|\boldsymbol{y})$ are not the marginals of $f(\boldsymbol{x},\theta|\boldsymbol{y})$. They are factors of which the product attempts to approximate the joint pdf as well as possible. The equalities in (23) should be interpreted as up to additive ”constants” (possibly functions of $\boldsymbol{y}$). Hence $f(\boldsymbol{x},\theta|\boldsymbol{y})$ is equivalent to $f(\boldsymbol{x},\theta,\boldsymbol{y})$ in (23). Finally, VB is an approach that normally applies to the fully Bayesian case in which both $\boldsymbol{x}$ and θ are considered random. However, we shall consider the prior $f(\theta)$ to be uniform so that $f(\boldsymbol{x},\boldsymbol{y},\theta)$ becomes equivalent to $f(\boldsymbol{x},\boldsymbol{y}|\theta)$.

The EM algorithm can be viewed as a limiting case of the VB approach, in which θ is treated as deterministic and hence can be viewed as random with prior $f(\theta') = \delta(\theta'-\theta)$ (where θ is the unknown true value). As a result also the posterior becomes of the form $q(\theta|\boldsymbol{y}) = \delta(\theta - \widehat{\theta})$ and hence is characterized solely by the point estimate $\widehat{\theta}$. Under some regularity conditions, the EM estimate is known to converge to the (separate) ML estimate and hence has the same performance. EM can be viewed as a case of VB in which $q^i(\theta|\boldsymbol{y})$ is forced to be of the form $\delta(\theta - \widehat{\theta}^i)$ in the M step. The best approximation is obviously obtained for $\widehat{\theta}^i = \arg\max_\theta q^i(\theta|\boldsymbol{y})$ where $q^i(\theta|\boldsymbol{y})$ is obtained from the first equation in (23). If now furthermore $q^i(\theta|\boldsymbol{y}) = \delta(\theta - \widehat{\theta}^i)$, then we get from the second equation in (23) $\ln\, q^i(\boldsymbol{x}|\boldsymbol{y}) \doteq \int\, \delta(\theta - \widehat{\theta}^i)\, \ln\, f(\boldsymbol{x},\boldsymbol{y}|\theta)\, d\theta = \ln\, f(\boldsymbol{x},\boldsymbol{y}|\widehat{\theta}^i) \doteq \ln\, f(\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i)$ since $f(\boldsymbol{y}|\widehat{\theta}^i)$ does not depend on $\boldsymbol{x}$. Hence $q^i(\boldsymbol{x}|\boldsymbol{y}) = f(\boldsymbol{x}|\widehat{\theta}^i,\boldsymbol{y})$. This finally leads to $\widehat{\theta}^{i+1} = \arg\max_\theta\, \int f(\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i)\, \ln f(\boldsymbol{x},\boldsymbol{y}|\theta)\, d\boldsymbol{x}$ for EM.

Now consider the actual VB updates (23) with Gaussian $q^i(\theta|\boldsymbol{y}) = \mathcal{N}(\widehat{\theta}^i, \boldsymbol{C}^i_\theta)$. Motivated by asymptotics we shall determine the Gaussian approximation by a 2nd order Taylor series expansion

$$\begin{aligned} &\ln q^{i+1}(\theta|\boldsymbol{y}) \doteq g^i(\theta) \doteq -\tfrac{1}{2}(\theta - \widehat{\theta}^{i+1})^T(\boldsymbol{C}^{i+1}_\theta)^{-1}(\theta - \widehat{\theta}^{i+1}) \\ &\doteq g^i(\widehat{\theta}^i) + (\theta - \widehat{\theta}^i)^T \tfrac{\partial g^i(\widehat{\theta}^i)}{\partial\theta} + \tfrac{1}{2}(\theta - \widehat{\theta}^i)^T \tfrac{\partial^2 g^i(\widehat{\theta}^i)}{\partial\theta\,\partial\theta^T}\,(\theta - \widehat{\theta}^i) \end{aligned} \tag{24}$$

Equating the last two lines yields

$$\widehat{\theta}^{i+1} = \widehat{\theta}^i + \boldsymbol{C}^{i+1}_\theta\, \frac{\partial \ln g^i(\widehat{\theta}^i)}{\partial\theta}\;,\; \boldsymbol{C}^{i+1}_\theta = \left(-\frac{\partial^2 \ln g^i(\widehat{\theta}^i)}{\partial\theta\,\partial\theta^T}\right)^{-1}. \tag{25}$$

This converges to a point $\widehat{\theta}^V$ for which $\frac{\partial \ln g^i(\widehat{\theta}^V)}{\partial\theta} = 0$ and for which $f^V(\theta|\boldsymbol{y}) = \mathcal{N}(\widehat{\theta}^V, \boldsymbol{C}^V_\theta)$. Now, we have for $g^i(\theta)$ in (24), from (23):

$$\begin{aligned} &g^i(\theta) = \;\mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})} \ln f(\boldsymbol{x},\boldsymbol{y}|\theta) \\ &= \ln f(\boldsymbol{y}|\theta) + \;\mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})} \ln f(\boldsymbol{x}|\boldsymbol{y},\theta) \\ &\doteq -\tfrac{1}{2}\ln\det(\boldsymbol{R}(\theta)) - \tfrac{1}{2}\;\mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})} \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}^T \boldsymbol{R}^{-1}(\theta) \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \\ &= -\tfrac{1}{2}[\ln\det(\boldsymbol{R}(\theta)) + \mathrm{tr}\{\boldsymbol{R}^{-1}(\theta)\; \widehat{\boldsymbol{R}}^i\}] \end{aligned} \tag{26}$$

where now $\widehat{\boldsymbol{R}}^i = \mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})}\, \boldsymbol{w}\boldsymbol{w}^T$ with $\boldsymbol{w}^T = [\boldsymbol{x}^T\boldsymbol{y}^T]$. Hence, the computation of $g^i(\theta)$ in VB is identical to that in EM except that $f(\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i)$ in EM is replaced by $q^i(\boldsymbol{x}|\boldsymbol{y})$ in VB. However, asymptotically, for the second-order expansion in (24), $q(\boldsymbol{x}|\boldsymbol{y})$ can equivalently be replaced by $f(\boldsymbol{x}|\boldsymbol{y},\widehat{\theta}^i)$. Hence, asymptotically there is no information for θ in $\mathrm{E}_{q(\boldsymbol{x}|\boldsymbol{y})} \ln f(\boldsymbol{x}|\boldsymbol{y},\theta)$ and $f^V(\theta|\boldsymbol{y}) = f^E(\theta|\boldsymbol{y}) = f^M(\theta|\boldsymbol{y}) = \mathcal{N}(\widehat{\theta}^M, \mathrm{CRB}^M)$. If VB for $\widehat{\theta}$ is asymptotically equivalent to ML, nevertheless

- this establishes that asymptotically one can not do better than ML (and CRB^M!),
- the convergence behavior of the VB iterations may be more interesting,
- non-asymptotically, the VB performance may be better than ML.

So we get for $\widehat{\theta}$

- AMAPML: $\widehat{\theta}$ from $\widehat{\boldsymbol{x}}$ only (as if $\widehat{\boldsymbol{x}} = \boldsymbol{x}$), $\widehat{\boldsymbol{x}}$ from $\widehat{\theta}$ only (as if $\widehat{\theta} = \theta$). AMAPML converges to joint MAPML.
- EM: $\widehat{\theta}$ from $\widehat{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{x}}$, $\widehat{\boldsymbol{x}}$ from $\widehat{\theta}$.
  EM converges to the marginalized ML approach.
- VB: $\widehat{\theta}$ from $\widehat{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{x}}$, $\widehat{\boldsymbol{x}}$ from $\widehat{\theta}$ and $\widetilde{\theta}$. Asymptotically same performance as ML and EM (hence efficient).

Note: all these iterative algorithms only require one iteration to converge if initialized with a consistent $\widehat{\theta}$.

We get for the VB update of $q(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\widehat{\boldsymbol{x}}, \boldsymbol{C}_{\boldsymbol{x}})$, from (23):

$$\begin{aligned} &\ln\, q^i(\boldsymbol{x}|\boldsymbol{y}) \doteq \int q^i(\theta|\boldsymbol{y})\; \ln\, f(\boldsymbol{x},\theta,\boldsymbol{y})\, d\theta \\ &\doteq \int q^i(\theta|\boldsymbol{y})\; \ln\, f(\boldsymbol{x}|\boldsymbol{y},\theta)\, d\theta \\ &\doteq -\tfrac{1}{2}\;\mathrm{E}_{q^i(\theta|\boldsymbol{y})}(\boldsymbol{x} - \boldsymbol{F}\boldsymbol{y})^T\boldsymbol{P}^{-1}(\boldsymbol{x} - \boldsymbol{F}\boldsymbol{y}) \\ &\doteq \;\mathrm{E}_{q^i(\theta|\boldsymbol{y})}\{\boldsymbol{y}^T\boldsymbol{F}^T\boldsymbol{P}^{-1}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{P}^{-1}\boldsymbol{x}\} \\ &\doteq -\tfrac{1}{2}\;(\boldsymbol{x} - \widehat{\boldsymbol{x}}^i)^T(\boldsymbol{C}^i_{\boldsymbol{x}})^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{x}}^i) \end{aligned} \tag{27}$$

where the Gaussian pdf comes out automatically. So we get

$$\widehat{\boldsymbol{x}}^i = \boldsymbol{C}_{\boldsymbol{x}}^i \left(\mathrm{E}_{q^i(\theta|\boldsymbol{y})}\, \boldsymbol{P}^{-1}(\theta)\boldsymbol{F}(\theta)\right) \boldsymbol{y},\, \boldsymbol{C}_{\boldsymbol{x}}^i = \left(\mathrm{E}_{q^i(\theta|\boldsymbol{y})}\boldsymbol{P}^{-1}(\theta)\right)^{-1} \tag{28}$$

which can be computed (asymptotically) by second-order expansions in θ of $\boldsymbol{P}^{-1}(\theta)$ and $\boldsymbol{P}^{-1}(\theta)\,\boldsymbol{F}(\theta)$.

Asymptotically, when $q(\theta|\boldsymbol{y})$ becomes $f(\theta|\boldsymbol{y})$, $\widehat{\boldsymbol{x}}^V$ attains a CRB corresponding to the following FIM

$$\begin{aligned} \boldsymbol{J}_{\boldsymbol{xx}}^V &= -\,\mathrm{E}_{\boldsymbol{x},\theta|\boldsymbol{y}} \frac{\partial^2 f(\boldsymbol{x},\boldsymbol{y},\theta)}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} \\ &= -\,\mathrm{E}_{\boldsymbol{x},\theta|\boldsymbol{y}} \frac{\partial^2 f(\boldsymbol{x}|\boldsymbol{y},\theta)}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} = \,\mathrm{E}_{\theta|\boldsymbol{y}} \boldsymbol{P}^{-1}(\theta) \end{aligned} \tag{29}$$

So, asymptotically, $\boldsymbol{C}_{\boldsymbol{x}}^V = \left(\boldsymbol{J}_{\boldsymbol{xx}}^V\right)^{-1}$. Note that the VB update of $q(\boldsymbol{x}|\boldsymbol{y})$ is non-iterative in $\boldsymbol{x}$, due to the quadratic nature of $\ln f(\boldsymbol{x}|\boldsymbol{y},\theta)$: $q(\boldsymbol{x}|\boldsymbol{y})$ is just a function of $q(\theta|\boldsymbol{y})$ (whereas $q^{i+1}(\theta|\boldsymbol{y})$ depends on both $q^i(\boldsymbol{x}|\boldsymbol{y})$ and $q^i(\theta|\boldsymbol{y})$). Hence the VB update for $\boldsymbol{x}$ is an extension of LMMSE estimation, accounting for the model parameter θ inaccuracies. There has been some work in recent years to account for the channel estimation error in LMMSE receiver design [20].

C. Second-Order Extended LMMSE (SOELMMSE)

One instance of LMMSE estimation is the Kalman Filter (KF). In the literature, variations on the KF theme have been derived to handle the joint filtering and parameter estimation problem, such as e.g. the widely used EM-KF algorithm ([21], [22], [23]). Another well-known variation is the Extended KF (EKF) algorithm, which can handle general nonlinear state space models. In this case, the state is extended with the unknown parameters, rendering the new state update equation nonlinear. A third derivation is the truncated Second-Order EKF (SOEKF) introduced by [24], [25] in which nonlinearities are expanded up to second order, third and higher order statistics being neglected. A corrected derivation of this filter is presented in [26]. In ([25], [27]), the Gaussian SOEKF is derived in which fourth-order terms in the Taylor series expansions are retained and approximated by assuming that the underlying joint probability distribution is Gaussian. Hence, various variations exist on the SOEKF theme.

Inspired by this state of the art, one possible extension of LMMSE to account for parameter estimation performance would be to optimize a linear estimator $\boldsymbol{F}$ in $\widehat{\boldsymbol{x}} = \boldsymbol{F}\,\boldsymbol{y}$ on the basis of an extended MSE criterion:

$$\mathrm{MSE} = \,\mathrm{E}_{\theta|\boldsymbol{y}}\,\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\theta}||\boldsymbol{x} - \boldsymbol{F}\,\boldsymbol{y}||^2 \tag{30}$$

leading to

$$\boldsymbol{F} = \left(\,\mathrm{E}_{\theta|\boldsymbol{y}}\,\boldsymbol{R}_{\boldsymbol{xy}}(\theta)\right)\left(\,\mathrm{E}_{\theta|\boldsymbol{y}}\,\boldsymbol{R}_{\boldsymbol{yy}}(\theta)\right)^{-1} \tag{31}$$

where $\boldsymbol{R}_{\boldsymbol{xy}}(\theta) = \,\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\theta}\boldsymbol{x}\,\boldsymbol{y}^T$ etc., which would be applicable also in the case of non-Gaussian $f(\boldsymbol{x}|\boldsymbol{y},\theta)$ and $f(\theta|\boldsymbol{y})$. Another possible approach would be to consider a jointly Gaussian $q(\boldsymbol{x},\theta|\boldsymbol{y})$.

References

[1] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. Wiley-Interscience, 2001.

[2] A. Wald, “Note on the Consistency of the Maximum Likelihood Estimate,” *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 595–601, Dec. 1949.

[3] H. Van Trees and K. Bell, *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*. Wiley-IEEE Press, 2007.

[4] W. Lindsey, *Synchronization Systems in Communication and Control*. Pearson Education Ltd, 1972.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum Likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[6] M. Beal, “Variational Algorithms for Approximate Bayesian Inference,” PHD, University of Cambridge, UK, 2003.

[7] A. Yeredor, “The Joint MAP-ML Criterion and its Relation to ML and to Extended Least-Squares,” *IEEE Trans. Signal Processing*, Dec. 2000.

[8] A. Guarnieri and S. Tebaldini, “Hybrid cramer-rao bounds for crustal displacement field estimators in sar interferometry,” *Signal Processing Letters, IEEE*, vol. 14, no. 12, pp. 1012–1015, 2007.

[9] P. Tichavsky and K. Wong, “Quasi-fluid-mechanics-based quasi-bayesian crame acute;r-rao bounds for deformed towed-array direction finding,” *Signal Processing, IEEE Transactions on*, vol. 52, no. 1, pp. 36–47, 2004.

[10] A. D’Andrea, U. Mengali, and R. Reggiannini, “The Modified Cramer-Rao Bound and its Application to Synchronization Problems,” *IEEE Trans. Communications*, Feb/Mar/Apr 1994.

[11] F. Gini, R. Reggiannini, and U. Mengali, “The Modified Cramer-Rao Bound in Vector Parameter Estimation,” *IEEE Trans. Communications*, Jan. 1998.

[12] M. Moeneclaey, “On the True and the Modified Cramer-Rao Bounds for the Estimation of a Scalar Parameter in the Presence of Nuisance Parameters,” *IEEE Trans. Communications*, Nov. 1998.

[13] S. Narasimhan and J. L. Krolik, “Fundamental limits on acoustic source range estimation performance in uncertain ocean channels,” *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 215–226, 1995.

[14] F. Gini and R. Reggiannini, “On the use of Cramer-Rao-like Bounds in the Presence of Random Nuisance Parameters,” *IEEE Trans. Communications*, Dec. 2000.

[15] S. Bay, B. Geller, A. Renaux, J. P. Barbot, and J. Brossier, “On the hybrid cramr rao bound and its application to dynamical phase estimation,” *Signal Processing Letters, IEEE*, vol. 15, pp. 453–456, 2008.

[16] Y. Rockah and P. Schultheiss, “Array shape calibration using sources in unknown locations–part i: Far-field sources,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 3, pp. 286–299, Mar .1987.

[17] I. Reuven and H. Messer, “A Barankin-type Lower Bound on the Estimation Error of a Hybrid Parameter Vector,” *IEEE Trans. Information Theory*, Mar. 1997.

[18] Y. Noam and H. Messer, “Notes on the Tightness of the Hybrid Cramer Rao Lower Bound,” *IEEE Trans. Signal Processing*, june 2009.

[19] E. M.-W. B. Z. Bobrovsky and M. Zakai, “Some classes of global cramrrao bounds,” *The Annals of Statistics*, vol. 15, no. 4, pp. 1421 –1438, Dec. 1987.

[20] P. Piantanida, S. Sadough, and P. Duhamel, “On the Outage Capacity of a Practical Decoder Accounting for Channel Estimation Inaccuracies,” *IEEE Trans. Communications*, May 2009.

[21] C. Couvreur and Y. Bresler, “Decomposition of a mixture of Gaussian AR processes,” *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 3, pp. 1605–1608, 1995.

[22] W. Gao, T. S., and J. Lehnert, “Diversity combining for DS/SS systems with time-varying, correlated fading branches,” *Communications, IEEE Transactions on*, vol. 51, no. 2, pp. 284–295, Feb 2003.

[23] M. Feder and E. Weinstein, “Parameter estimation of superimposed signals using the EM algorithm ,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 4, pp. 477–489, Apr 1988.

[24] R. D. Bass, V. D. Norum, and L. Swartz, “Optimal multichannel nonlinear filtering,” *J. Mufh. Anal. Appl.*, vol. 16, pp 152 – 164, 1966.

[25] A. H. Jazwinski, *Stochastic processes and filtering theory*, 1970.

[26] R. Henriksen, “The truncated second-order nonlinear filter revisited,” *Automatic Control, IEEE Transactions on*, vol. 27, no. 1, pp. 247 – 251, feb 1982.

[27] M. Athans, R. Wishner, and A. Bertolini, “Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements,” *Automatic Control, IEEE Transactions on*, vol. 13, no. 5, pp. 504 – 514, oct 1968.