Sparse Bayesian Learning for a Bilinear Calibration Model and Mismatched CRB

Kalyana Gopala, Christo Kurisummoottil Thomas, Dirk Slock
EURECOM, Sophia-Antipolis, France, Email:{gopala,kurisumm,slock}@eurecom.fr

***Abstract*—Variational Bayesian (VB) estimation allows for approximate Bayesian inference. It determines the closest approximation in factored form of the posterior distribution by minimizing the Kullback-Leibler distance to the posterior distribution even if this last one is difficult to determine. In spite of this well motivated derivation, the performance of VB techniques is not very clear, especially compared to more classical performance bounds. In this paper we explore recently introduced mismatched Cramer-Rao bounds (mCRB) for Bayesian estimation in the context of VB estimation. We focus on the case of bilinear signal models. One particular application of these models arises in the context of internal relative reciprocity calibration of Massive antenna arrays, in which the received signals are linear in terms of an intra array channel and the relative calibration factors. We have recently shown that a VB approach allows for particularly improved estimation performance that goes beyond the classical CRB, which is now confirmed by the mCRB.**

I. Introduction

Massive MIMO (Multiple Input Multiple Output) requires CSIT (Channel state information at Tx) acquired using channel reciprocity for a TDD (Time Division Duplexing) system. However, Radio Frequency (RF) components are not reciprocal and we need to calibrate to compensate for this. This calibration is typically achieved by a simple complex scalar multiplication at each transmit antenna. Initial approaches to calibration relied on explicit channel feedback from a user equipment (UE) during the calibration phase to estimate the calibration parameters. This is typically referred to as UE aided calibration. However, what is popular today [1] is to perform the calibration across the antennas of the base station (BS) only and is referred to as internal calibration. In [2], the authors propose a generalized approach towards reciprocity calibration of which the existing estimation techniques are special cases.

Both the classical deterministic estimation theory and Bayesian framework are based on the assumption that the assumed data model and the true data model (pdf) are the same. However, in practice, either we may only have an imperfect knowledge of the true data model or due to computational complexities associated with the computation of the true posterior distributions, we prefer approximate Bayesian inference (VB). In such a misspecified estimation framework, it is important to quantify the performance of the estimator using a mismatched Cramer-Rao bounds (mCRB) [3].

A. Contributions of this paper:

- We first review the constrained CRB for the case of a bilinear system model (linear in terms of the relative calibration factors and reciprocal channel coefficients).

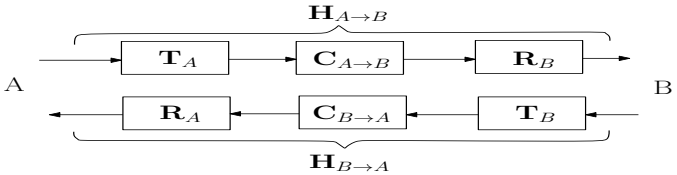

Fig. 1. Reciprocity Model

- We propose a VB (and other variants like AMAP, EC-VB) based estimation algorithm for the joint estimation of the calibration parameters, reciprocal channel coefficients and hyper-parameters (precisions of the bilinear factors).
- Simulations demonstrate that the mean square error (MSE) of the VB can be lower than that of the deterministic CRB. Motivated by this result, we derive simple and elegant expressions for the mCRB using Laplace approximation for the relative calibration factors.

Notations: The operators $tr(\cdot)$, $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$, $\|\cdot\|$ represents trace, transpose, conjugate, conjugate transpose and Frobenius norm respectively. Boldface lower-case and upper-case characters denote vectors and matrices respectively.

II. Reciprocity Calibration System Model

Consider a system as in Fig. 1, where A represents a BS and B represents a UE, each containing M_A and M_B antennas, respectively. The channel as observed in the digital domain, $\mathbf{H}_{A\to B}$ and $\mathbf{H}_{B\to A}$ can be represented by,

$$\mathbf{H}_{A\to B} = \mathbf{R}_B\mathbf{C}_{A\to B}\mathbf{T}_A, \qquad \mathbf{H}_{B\to A} = \mathbf{R}_A\mathbf{C}_{B\to A}\mathbf{T}_B, \tag{1}$$

where (diagonal) matrices $\mathbf{T}_A$, $\mathbf{R}_A$, $\mathbf{T}_B$, $\mathbf{R}_B$ model the response of the transmit and receive RF front-ends, while $\mathbf{C}_{A\to B}$ and $\mathbf{C}_{B\to A}$ model the propagation channels, respectively from A to B and from B to A. Let us consider an antenna array of M elements partitioned into G groups denoted by $A_1, A_2, \ldots, A_G$. Group A_i contains M_i antennas such that $\sum_{i=1}^{G} M_i = M$. Each group A_i transmits a sequence of L_i pilot symbols, defined by matrix $\mathbf{P}_i \in \mathbb{C}^{M_i\times L_i}$ where the rows correspond to antennas and the columns to successive channel uses. After all G groups have transmitted, the received signal for each resource block of bidirectional transmission between antenna groups i and j is given by

$$\begin{cases} \mathbf{Y}_{i\to j} = \mathbf{R}_j\mathbf{C}_{i\to j}\mathbf{T}_i\mathbf{P}_i + \mathbf{N}_{i\to j}, \\ \mathbf{Y}_{j\to i} = \mathbf{R}_i\mathbf{C}_{j\to i}\mathbf{T}_j\mathbf{P}_j + \mathbf{N}_{j\to i}\ . \end{cases} \tag{2}$$

We define $\mathbf{F}_i = \mathbf{R}_i^{-T}\mathbf{T}_i$ and $\mathbf{F}_j = \mathbf{R}_j^{-T}\mathbf{T}_j$ to be the calibration matrices for groups i and j. Also, $\mathbf{f}_i = vec(\mathbf{F}_i)$ represents

the vectorized version. This needs to be augmented with a constraint $\mathcal{C}(\hat{\mathbf{f}}, \mathbf{f}) = 0$. Typical choices for the constraint are

1) Norm plus phase constraint (NPC):
$$\begin{aligned} &\text{norm: } \mathrm{Re}\{\mathcal{C}(\hat{\mathbf{f}}, \mathbf{f})\} = ||\hat{\mathbf{f}}||^2 - c\,,\ c = ||\mathbf{f}||^2, \\ &\text{phase: } \mathrm{Im}\{\mathcal{C}(\hat{\mathbf{f}}, \mathbf{f})\} = \mathrm{Im}\{\hat{\mathbf{f}}^H \mathbf{f}\} = 0. \end{aligned}$$
2) Linear constraint: $\mathcal{C}(\hat{\mathbf{f}}, \mathbf{f}) = \hat{\mathbf{f}}^H \mathbf{g} - c = 0$. (3)

If we choose the vector $\mathbf{g} = \mathbf{f}$ and $c = ||\mathbf{f}||^2$, then the $\mathrm{Im}\{.\}$ part of (3) corresponds to (3). The most popular linear constraint is the First Coefficient Constraint (FCC), which is (3) with $\mathbf{g} = \mathbf{e}_1$, $c = 1$. From (2), we have

$$\mathbf{Y}_{i\to j} = \underbrace{\mathbf{R}_j \mathbf{C}_{i\to j} \mathbf{R}_i^T}_{\mathcal{H}_{i\to j}} \mathbf{F}_i \mathbf{P}_i + \mathbf{N}_{i\to j}. \tag{4}$$

We define $\mathcal{H}_{i\to j} = \mathbf{R}_j \mathbf{C}_{i\to j} \mathbf{R}_i^T$ to be an auxiliary internal channel (not corresponding to any physically measurable quantity) that appears as a nuisance parameter in the estimation of the calibration parameters. Note that the auxiliary channel $\mathcal{H}_{i\to j}$ inherits the reciprocity from the channel $\mathbf{C}_{i\to j}$: $\mathcal{H}_{i\to j} = \mathcal{H}_{j\to i}^T$. Upon applying the vectorization operator for each bidirectional transmission between groups i and j, we have

$$\mathrm{vec}(\mathbf{Y}_{i\to j}) = (\mathbf{P}_i^T * \mathcal{H}_{i\to j})\, \mathbf{f}_i + \mathrm{vec}(\mathbf{N}_{i\to j}). \tag{5}$$

In the reverse direction, using $\mathcal{H}_{i\to j} = \mathcal{H}_{j\to i}^T$, we have

$$\mathrm{vec}(\mathbf{Y}_{j\to i}^T) = (\mathcal{H}_{i\to j}^T * \mathbf{P}_j^T)\mathbf{f}_j + \mathrm{vec}(\mathbf{N}_{j\to i})^T. \tag{6}$$

Alternatively, (5) and (6) may also be written as

$$\begin{aligned} \mathrm{vec}(\mathbf{Y}_{i\to j}) &= \left[(\mathbf{F}_i \mathbf{P}_i)^T \otimes \mathbf{I}\right] \mathrm{vec}(\mathcal{H}_{i\to j}) + \mathrm{vec}(\mathbf{N}_{i\to j}) \\ \mathrm{vec}(\mathbf{Y}_{j\to i}^T) &= \left[\mathbf{I} \otimes (\mathbf{P}_j^T \mathbf{F}_j)\right] \mathrm{vec}(\mathcal{H}_{i\to j}) + \mathrm{vec}(\mathbf{N}_{j\to i}). \end{aligned} \tag{7}$$

Stacking these observations into a vector $\mathbf{y} = \left[\mathrm{vec}(\mathbf{Y}_{1\to 2})^T\, \mathrm{vec}(\mathbf{Y}_{2\to 1}^T)^T\, \mathrm{vec}(\mathbf{Y}_{1\to 3})^T \dots\right]^T$, the above two alternative formulations can be summarized into

$$\mathbf{y} = \mathcal{H}(\mathbf{h}, \mathbf{P})\mathbf{f} + \mathbf{n} = \mathcal{F}(\mathbf{f}, \mathbf{P})\mathbf{h} + \mathbf{n}, \tag{8}$$

where $\mathbf{h} = \left[\mathrm{vec}(\mathcal{H}_{1\to 2})^T\, \mathrm{vec}(\mathcal{H}_{1\to 3})^T\, \mathrm{vec}(\mathcal{H}_{2\to 3})^T \dots\right]^T$, and $\mathbf{n}$ is the corresponding noise vector. The expressions for the composite matrices $\mathcal{H}$ and $\mathcal{F}$ are the same as given in [4, equation (18)]. The scenario is now identical to that encountered in some blind channel estimation scenarios and hence we can take advantage of some existing tools [5], [6], which we exploit next.

A. Cramér-Rao bound

Treating $\mathbf{h}$ and $\mathbf{f}$ as deterministic unknown parameters, and assuming that the receiver noise $\mathbf{n}$ is distributed as $\mathcal{CN}(0, \sigma^2\mathbf{I})$, the Fisher Information Matrix (FIM) $\mathbf{J}$ for jointly estimating $\mathbf{f}$ and $\mathbf{h}$ can immediately be obtained from (8) as

$$\mathbf{J} = \frac{1}{\sigma^2} \begin{bmatrix} \mathcal{H}^H \\ \mathcal{F}^H \end{bmatrix} \begin{bmatrix} \mathcal{H} & \mathcal{F} \end{bmatrix}. \tag{9}$$

The computation of the CRB requires $\mathbf{J}$ to be non-singular. However, for the problem at hand, $\mathbf{J}$ is inherently singular. In fact, the calibration factors (and the auxiliary channel) can only be estimated up to a complex scale factor since the received data (8) involves the product of the channel and the calibration factors, $\mathcal{H}\mathbf{f} = \mathcal{F}\mathbf{h}$. As a result the FIM has the following null space [7], [8]

$$\mathbf{J} \begin{bmatrix} \mathbf{f}^T & -\mathbf{h}^T \end{bmatrix}^T = \frac{1}{\sigma^2} \begin{bmatrix} \mathcal{H} & \mathcal{F} \end{bmatrix}^H (\mathcal{H}\mathbf{f} - \mathcal{F}\mathbf{h}) = \mathbf{0}. \tag{10}$$

To determine the CRB when the FIM is singular, constraints have to be added to regularize the estimation problem. As the calibration parameters are complex, one complex constraint corresponds to two real constraints. Another issue is that we are mainly interested in the CRB for $\mathbf{f}$, the parameters of interest, in the presence of the nuisance parameters $\mathbf{h}$. Hence we are only interested in the $(1,1)$ block of the inverse of the 2×2 block matrix $\mathbf{J}$ in (9). Incorporating the effect of the constraint (3) on $\mathbf{f}$, we can derive from [8] the following constrained CRB for $\mathbf{f}$

$$\mathrm{CRB}_{\mathbf{f}} = \sigma^2 \mathcal{V}_{\mathbf{f}} \left(\mathcal{V}_{\mathbf{f}}^H \mathcal{H}^H \mathcal{P}_{\mathcal{F}}^{\perp} \mathcal{H} \mathcal{V}_{\mathbf{f}}\right)^{-1} \mathcal{V}_{\mathbf{f}}^H, \tag{11}$$

where $\mathcal{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^H\mathbf{X})^{\dagger}\mathbf{X}^H$ and $\mathcal{P}_{\mathbf{X}}^{\perp} = \mathbf{I} - \mathcal{P}_{\mathbf{X}}$ are the projection operators on resp. the column space of matrix $\mathbf{X}$ and its orthogonal complement, and $\dagger$ corresponds to the Moore-Penrose pseudo inverse. Note that in some group calibration scenarios, $\mathcal{F}^H\mathcal{F}$ can be singular (i.e, $\mathbf{h}$ could be not identifiable even if $\mathbf{f}$ is identifiable or even known). The $M \times (M-1)$ matrix $\mathcal{V}_{\mathbf{f}}$ is such that its column space spans the orthogonal complement of that of $\frac{\partial \mathcal{C}(f)}{\partial \mathbf{f}^*}$, i.e., $\mathcal{P}_{\mathcal{V}_{\mathbf{f}}} = \mathcal{P}_{\frac{\partial \mathcal{C}}{\partial \mathbf{f}^*}}^{\perp}$.

It is shown in [7], [8], [9] that a choice of constraints such that their linearized version $\frac{\partial \mathcal{C}}{\partial \mathbf{f}^*}$ fills up the null space of the FIM results in the lowest CRB, while not adding information in subspaces where the data provides information. One such choice is the set (3) (NPC). Another choice is (3) with $\mathbf{g} = \mathbf{f}$. With such constraints, $\frac{\partial \mathcal{C}}{\partial \mathbf{f}^*} \sim \mathbf{f}$ which spans the null space of $\mathcal{H}^H \mathcal{P}_{\mathcal{F}}^{\perp} \mathcal{H}$. The CRB then corresponds to the pseudo inverse of the FIM and (11) becomes $\mathrm{CRB}_{\mathbf{f}} = \sigma^2 \left(\mathcal{H}^H \mathcal{P}_{\mathcal{F}}^{\perp} \mathcal{H}\right)^{\dagger}$. If the FCC constraint is used instead (i.e., (3) with $\mathbf{g} = \mathbf{e}_1$, $c = 1$), where $\mathbf{e}_1$ is an all zero vector with only the first coefficient one, the corresponding CRB is (11) where $\mathcal{V}_{\mathbf{f}}$ corresponds now to an identity matrix without the first column (and hence its column space is the orthogonal complement of $\mathbf{e}_1$).

B. Variational Bayes (VB) Estimation

In VB, a Bayesian estimate is obtained by computing an approximation to the posterior distribution of the parameters $\mathbf{h}, \mathbf{f}$ with priors $\mathbf{f} \sim \mathcal{CN}(0, \alpha^{-1}\mathbf{I}_{\mathbf{M}})$, $\mathbf{h} \sim \mathcal{CN}(0, \beta^{-1}\mathbf{I}_{\mathbf{N_h}})$ and α, β are assumed to have themselves a uniform prior. N_h is the number of elements in $\mathbf{h}$. This approximation, called the variational distribution, is chosen to minimize the Kullback-Leibler distance between the true posterior distribution $p(\mathbf{h}, \mathbf{f}, \alpha, \beta \mid \mathbf{y})$ and a factored variational distribution $q(\mathbf{h}, \mathbf{f}, \alpha, \beta \mid \mathbf{y}) = q_{\mathbf{h}}(\mathbf{h})\, q_{\mathbf{f}}(\mathbf{f})\, q_{\alpha}(\alpha)\, q_{\beta}(\beta)$. The factors can be obtained in an alternating fashion as [10],

$$\ln(q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)) = < \ln p(\mathbf{y}, \mathbf{h}, \mathbf{f}, \alpha, \beta) >_{k\neq i} + c_i, \tag{12}$$

where $\boldsymbol{\theta}_i$ refers to the i^{th} block of $\boldsymbol{\theta} = [\mathbf{h}, \mathbf{f}, \alpha, \beta]$ and $<>_{k\neq i}$ represents the expectation operator over the distributions $q_{\boldsymbol{\theta}_k}$ for all $k \neq i$. c_i is a normalizing constant. Further considering the constraints on $\mathbf{f}$ ($\mathbf{f}_{\perp}$ represents the component of $\mathbf{f}$ in the null space of the constraint) and applying VB (12),

$\mathbf{f} = \mathbf{f}' + \mathcal{V}_f \mathbf{f}_\perp, \mathbf{f}' = \mathbf{g}\frac{c}{\|\mathbf{g}\|^2}, \mathbf{f}^H\mathbf{g} = c > 0, \mathcal{V}_f^H\mathcal{V}_f = \mathbf{I},$
$\ln q_{\mathbf{f}}(\mathbf{f}) = \frac{1}{\sigma^2}(\mathbf{f}'^H + \mathbf{f}_\perp^H\mathcal{V}_f^H) < \mathcal{H}^H > \mathbf{y} + \frac{1}{\sigma^2}\mathbf{y}^H < \mathcal{H} >$
$(\mathbf{f}' + \mathcal{V}_f\mathbf{f}_\perp) - \frac{1}{\sigma^2}(\mathbf{f}'^H + \mathbf{f}_\perp^H\mathcal{V}_f^H) < \mathcal{H}^H\mathcal{H} > (\mathbf{f}' + \mathcal{V}_f\mathbf{f}_\perp)$
$- < \alpha > \|\mathbf{f}_\perp\|^2 + c_f,$
$\ln q_{\mathbf{h}}(\mathbf{h}) = \frac{\mathbf{h}^H < \mathcal{F}^H > \mathbf{y} + \mathbf{y}^H < \mathcal{F} > \mathbf{h} - \mathbf{h}^H < \mathcal{F}^H\mathcal{F} > \mathbf{h}}{\sigma^2} - < \beta > \mathbf{h}^H\mathbf{h}.$ (13)

Here, N_y refers to the number of elements in $\mathbf{y}$ and c is a constant. Here c_p, c_f represents the normalization constants for the respective pdfs. We shall assume here that the noise variance σ^2 is known (or estimated in a separate training procedure). It is now straightforward to see that proceeding as in (12), α, β would have a Gamma distribution and a complex normal distribution for $\mathbf{f} \sim \mathcal{CN}(\hat{\mathbf{f}}, \mathbf{C}_{\tilde{f}\tilde{f}})$ and $\mathbf{h} \sim \mathcal{CN}(\hat{\mathbf{h}}, \mathbf{C}_{\tilde{h}\tilde{h}})$. The detailed expressions are summarized in Algorithm 1. When $G = M$, $\mathbf{C}_{\tilde{f}\tilde{f}}$ and $\mathbf{C}_{\tilde{h}\tilde{h}}$ are diagonal and $< \mathcal{F}^H(\tilde{\mathbf{f}})\mathcal{F}(\tilde{\mathbf{f}}) >$, $< \mathcal{H}^H(\tilde{\mathbf{h}})\mathcal{H}(\tilde{\mathbf{h}}) >$ can be computed easily (diagonal). However, when $G < M$, these matrices are block diagonal.

Algorithm 1 VB Estimation of calibration parameters

1: **Initialization:** Initialize $\hat{\mathbf{f}}$ using existing calibration methods. Use $\hat{\mathbf{f}}$ to determine $\hat{\mathbf{h}}, <\alpha>, <\beta>$, with $\mathbf{g} = \mathbf{e}_1$.
2: **repeat**
3: $< \mathcal{H}^H\mathcal{H} > = \mathcal{H}^H(\hat{\mathbf{h}})\mathcal{H}(\hat{\mathbf{h}}) + < \mathcal{H}^H(\tilde{\mathbf{h}})\mathcal{H}(\tilde{\mathbf{h}}) >$.
4: $\hat{\mathbf{f}}_\perp = (\mathcal{V}_f^H(< \mathcal{H}^H\mathcal{H} > + \sigma^2 < \alpha > \mathbf{I})\mathcal{V}_f)^{-1}\mathcal{V}_f^H(< \mathcal{H}^H > \mathbf{y} - < \mathcal{H}^H\mathcal{H} > \mathbf{f}')$
5: $\mathbf{C}_{\tilde{f}\tilde{f}} = \mathcal{V}_f(\mathcal{V}_f^H(\frac{1}{\sigma^2} < \mathcal{H}^H\mathcal{H} > + < \alpha > \mathbf{I})\mathcal{V}_f)^{-1}\mathcal{V}_f^H$
6: $< \mathcal{F}^H\mathcal{F} > = \mathcal{F}^H(\hat{\mathbf{f}})\mathcal{F}(\hat{\mathbf{f}}) + < \mathcal{F}^H(\tilde{\mathbf{f}})\mathcal{F}(\tilde{\mathbf{f}}) >$
7: $\hat{\mathbf{h}} = (< \mathcal{F}^H\mathcal{F} > + \sigma^2 < \beta > \mathbf{I})^{-1}\mathcal{F}^H\mathbf{y}, \mathbf{C}_{\tilde{h}\tilde{h}} = (\frac{1}{\sigma^2} < \mathcal{F}^H\mathcal{F} > + < \beta > \mathbf{I})^{-1}, < \alpha > = \frac{M}{<\|\mathbf{f}_\perp\|^2>}, < \|\mathbf{f}_\perp\|^2 > = \hat{\mathbf{f}}_\perp^H\hat{\mathbf{f}}_\perp + tr\{\mathbf{C}_{\tilde{f}_\perp\tilde{f}_\perp}\}$.
8: $< \beta > = \frac{N_h + 1}{<\|\mathbf{h}\|^2>}, < \|\mathbf{h}\|^2 > = \hat{\mathbf{h}}^H\hat{\mathbf{h}} + tr\{\mathbf{C}_{\tilde{h}\tilde{h}}\}$.
9: **until** convergence.

An approximate version of Algorithm 1, EC-VB (Expectation Consistent [11] VB) [4] where the error covariance matrix are approximated to be multiple of identity is also considered in the simulations. Note here that by forcing the matrices $\mathbf{C}_{\tilde{f}\tilde{f}}$, $\mathbf{C}_{\tilde{h}\tilde{h}}$ to zero and α, β to zero, this algorithm reduces to the Alternating Maximum Likelihood (AML) algorithm [5], [6] which iteratively maximizes the likelihood by alternating between the desired parameters $\mathbf{f}$ and the nuisance parameters $\mathbf{h}$ for the formulation (8). The penalized ML method used in [14] uses quadratic regularization terms for both $\mathbf{f}$ and $\mathbf{h}$ which can be interpreted as Gaussian priors and which may improve estimation in ill-conditioned cases. In our case, we arrive at a similar solution from the VB perspective and more importantly, the regularization terms are optimally tuned.

III. Mismatched CRB's

As can be seen in Fig. 2, VB allows to attain lower MSE than the CRB (for deterministic parameters). One possibility to evaluate the performance is to consider the Bayesian CRB. However, VB is an approximate Bayesian estimation technique. Also, a Bayesian CRB is valid only if the (Gaussian) priors for $\mathbf{f}$ and $\mathbf{h}$ are the correct priors. However, the interest of the VB technique is that it will converge to the most appropriate priors even if in fact the parameters $\mathbf{f}$ and $\mathbf{h}$ are deterministic! This requires Mismatched CRBs. In this paper, we explore the Bayesian mCRB exposed in [12], [13].

Under a mismatched distribution model, it is important to define the convergence point $\overline{\boldsymbol{\theta}}$ (also called as pseudo true parameter) which is used to evaluate the effectiveness of the estimator, since no true parameter vector may exists under the assumed distribution q. The VB convergence point (of complete $\boldsymbol{\theta}$) is the MAP of $E_p(\sum_i \ln(q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)))$ (assuming large amount of data), so $\ln$ of product of q's $=$ sum of $\ln$ of q's and converges to it's expected value according to actual pdf p (law of large numbers). Similar to [12] (misspecified CRBs) which considers deterministic case, we do it also for random $\boldsymbol{\theta}$, but not neglecting priors in the asymptotic regime (considering some fictitious asymptotic regime in which prior information scales similarly as information in data, so that both continue to count, but get a Gaussian concentration around convergence point).

A. mCRB Bilinear Model

CRB corresponds to Laplace approximation of MAP or VB. Laplace approximation [10] refers to the evaluation of marginal likelihood or free energy using Laplace's method. This is equivalent to a Gaussian approximation of the posterior q around a maximum a posteriori (MAP) estimate, motivated by the fact that in the asymptotic limit (large amount of data or high SNR), the posterior approaches a Gaussian around the MAP point [12]. Gradients of $\ln q$ can be taken from the recursions for $\ln q$ (12), so it's the gradients of $\ln p$ as usual, except with averaging over $q_{\bar{i}}$ for gradient and Hessian as we will show here. But the final error covariance matrix of Laplace approximation (2^{nd} order Taylor) is expectation with p. Let $\widehat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$ based on the approximate posterior q and the assumed prior. Let $\zeta = \widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}$, where the estimator mean is evaluated at the point $\overline{\boldsymbol{\theta}}$. First, we need to find $\overline{\boldsymbol{\theta}}$. This corresponds to the peak of the posterior pdf in an asymptotic scenario of large amount of data or high SNR, computation of which is derived in III-B. Throughout the paper, the vector $\boldsymbol{\theta}_i$ represents a subset of $\boldsymbol{\theta}$ and θ_i represents a scalar parameter in $\boldsymbol{\theta}$. In this section $\boldsymbol{\theta}$ (a column vector) contains the parameters $\mathbf{h}, \mathbf{f}$ and $\boldsymbol{\psi}$ the precision parameters, α, β and $\boldsymbol{\theta}^0$ denotes the true value of $\boldsymbol{\theta}$. $\overline{\boldsymbol{\theta}}$ (or $\overline{\boldsymbol{\psi}}$) can be evaluated as,

$$\begin{aligned}\overline{\boldsymbol{\theta}}_i &= \arg\max_{\boldsymbol{\theta}_i} \mathrm{E}_{p(\mathbf{y},\boldsymbol{\theta}^0)} \ln q(\boldsymbol{\theta}_i) \\ &= \arg\max_{\boldsymbol{\theta}_i} \mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta}^0)} \ln < p(\mathbf{y},\boldsymbol{\theta}) >_{\bar{i}}.\end{aligned} \quad (14)$$

Even though the parameters are modeled as random for estimation, but we assume that in reality they are deterministic. So the expectation over $p(\boldsymbol{\theta})$ disappears in (14). Also, we define $\widetilde{\overline{\boldsymbol{\theta}}} = \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}$, $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = \zeta + \widetilde{\overline{\boldsymbol{\theta}}}$. For any choice of score function η using a matrix generalization of the Cauchy Schwartz inequality [3], [13], the error correlation matrix can be written as,

$$\mathbf{mCRB} = \mathbf{R}_{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}} = \mathrm{E}_p \widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^H \geq \mathbf{R}_{\zeta\eta}\mathbf{R}_{\eta\eta}^{-1}\mathbf{R}_{\eta\zeta} + \widetilde{\overline{\boldsymbol{\theta}}}\widetilde{\overline{\boldsymbol{\theta}}}^H, \quad (15)$$

where $\mathbf{R}_{\zeta\eta} = \mathrm{E}(\zeta\eta^H)$ and $\mathbf{R}_{\zeta\zeta} = \mathrm{E}(\zeta\zeta^H)$.

The score function can be written as,

$$\begin{aligned}\eta &= \tfrac{\partial}{\partial \boldsymbol{\theta}^*} \ln q(\boldsymbol{\theta}) \mid_{\overline{\boldsymbol{\theta}}} - \mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta}^0)} \tfrac{\partial}{\partial \boldsymbol{\theta}^*} \ln q(\boldsymbol{\theta}) \mid_{\overline{\boldsymbol{\theta}}} \\ &= \tfrac{\partial}{\partial \boldsymbol{\theta}^*} \ln q(\boldsymbol{\theta}) \mid_{\overline{\boldsymbol{\theta}}} - \mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta}^0)} (\tfrac{\partial}{\partial \boldsymbol{\theta}^*} \ln q(\boldsymbol{\theta}) \mid_{\overline{\boldsymbol{\theta}}}).\end{aligned} \quad (16)$$

The choice of the score function is motivated by the requirements for the tightness of the CRB detailed in [3] that it should be zero mean and depends on the sufficient statistic for estimating $\boldsymbol{\theta}$. So the score function here is the score function for the deterministic CRB minus its possibly non zero-mean under the true model $p(\mathbf{y}, \boldsymbol{\theta}^0)$. Also, the particular choice score function (16) results in $\mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta}^o)}(\frac{\partial}{\partial \boldsymbol{\theta}^*} \ln q(\boldsymbol{\theta}) \mid_{\overline{\boldsymbol{\theta}}}) = 0$, due to the Laplace approximation of $\boldsymbol{\theta}$ around the asymptotic estimate $\overline{\boldsymbol{\theta}}$. Further under concentration conditions (data asymptotics or SNR asymptotics, or perhaps prior asymptotics (becoming very precise)), we can do a 2^{nd} order Taylor series of misspecified posterior. The Taylor series expansion of the data likelihood around $\overline{\boldsymbol{\theta}}$ is given by,

$$\log q(\mathbf{y}, \overline{\boldsymbol{\theta}} + \Delta\theta) = \log q(\mathbf{y}, \overline{\boldsymbol{\theta}}) + \Delta\boldsymbol{\theta}^H \frac{\partial \log q(\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \mid_{\overline{\boldsymbol{\theta}}} + \Delta\boldsymbol{\theta}^H \frac{\partial^2 \log q(\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} \mid_{\overline{\boldsymbol{\theta}}} \Delta\boldsymbol{\theta} + o(\|\Delta\boldsymbol{\theta}\|^2). \tag{17}$$

Further neglecting the higher order terms and equating the derivative w.r.t $\Delta\boldsymbol{\theta}^*$ to be zero yields an approximation of the error term ζ as,

$$\zeta = -(\frac{\partial^2 \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} \mid_{\overline{\boldsymbol{\theta}}})^{-1} \frac{\partial \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \mid_{\overline{\boldsymbol{\theta}}}. \tag{18}$$

Note that we can replace the Hessian and $\frac{\partial \log q(\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}$ in (18) by $\mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta})}(\frac{\partial^2 \log q(\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T})$ and $\mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta})}(\frac{\partial \log q(\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*})$ respectively in the asymptotic limit. Taking the derivative of the data log-likelihood gives,

$$\frac{\partial \log q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} = -\frac{1}{\sigma^2} \begin{bmatrix} \mathcal{V}_f^H < \mathcal{H}^H \mathcal{H} > \mathbf{f} - \mathcal{V}_f^H < \mathcal{H} > \mathbf{y} + < \alpha > \mathbf{f}_\perp \\ < \mathcal{F}^H \mathcal{F} > \mathbf{h} - < \mathcal{F}^H > \mathbf{y} + < \beta > \mathbf{h} \end{bmatrix}^0,$$
$$\mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta})} \frac{\partial^2 \log q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} = -\mathcal{V}^H \mathbf{Q} \mathcal{V}, \mathbf{Q} = \frac{1}{\sigma^2} \mathrm{blkdiag}\,(0, \mathcal{V}_f^H < \mathcal{H}^H \mathcal{H} > \mathcal{V}_f + < \alpha > \mathbf{I}, < \mathcal{F}^H \mathcal{F} > + < \beta > \mathbf{I}). \tag{19}$$

where $\mathrm{blkdiag}\,(\cdot)$ represents the block diagonal matrix formed by the respective matrix elements in the block. The evaluation of $\mathbf{Q}$ at the asymptotic limit, $\overline{\boldsymbol{\theta}}$, be denoted as $\overline{\mathbf{Q}}$. Let $\mathrm{E}(\frac{\partial \log q(\mathbf{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}) \mid_{\overline{\boldsymbol{\theta}}} = f(\overline{\boldsymbol{\theta}})$. The error term ζ can then be expressed as, $\zeta = \mathcal{V}(\mathcal{V}^H \overline{\mathbf{Q}} \mathcal{V})^{-1} \mathcal{V}^H f(\mathbf{y}, \boldsymbol{\theta})$. Note that $\mathcal{V} = [\mathbf{0}\ \mathbf{I}]$. The cross correlation matrix between ζ and η becomes,

$$\mathbf{R}_{\zeta\eta} = -\mathcal{V}(\mathcal{V}^H \overline{\mathbf{Q}} \mathcal{V})^{-1} \mathcal{V}^H f(\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}, \boldsymbol{\theta})^H = -(\mathcal{V}(\mathcal{V}^H \overline{\mathbf{Q}} \mathcal{V})^{-1} \mathcal{V}^H f(\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}, \boldsymbol{\theta})^H). \tag{20}$$

Here $f(\overline{\boldsymbol{\theta}}) f(\overline{\boldsymbol{\theta}})^H = \mathbf{J}_q$. Finally substituting (20) in (15), we obtain (define $\mathbf{MFIM}$ to be the corresponding mismatched FIM),

$$\mathbf{mCRB} = \mathcal{V}(\mathcal{V}^H \overline{\mathbf{Q}} \mathcal{V})^{-1} \mathcal{V}^H \mathbf{J}_q \mathcal{V}(\mathcal{V}^H \overline{\mathbf{Q}} \mathcal{V})^{-1} \mathcal{V}^H + \widetilde{\overline{\boldsymbol{\theta}}} \widetilde{\overline{\boldsymbol{\theta}}}^H. \tag{21}$$

Further we derive the mCRB for VB ($\mathbf{mCRB}_{VB}$) with the posteriors of $\mathbf{h}, \mathbf{f}$ being factorized.

Lemma 1. *If the parameter partitioning in VB is such that the different parameter blocks are decoupled at the level of Fisher Information Matrix, then VB is not suboptimal in terms of (mismatched) Cramer-Rao Bound. If a finer partitioning granularity is used (such as up to scalar level as in mean field), then VB becomes quite suboptimal.*

So in the too fine partitioning case, the VB partitioning is applied to the MFIM, taking a too fine blockdiagonal part, and since that partitioning is finer than the blockdiagonal MFIM structure, then the inverse of the too fine blockdiagonal part of the FIM does not give the correct CRB. So $\mathbf{mCRB}_{VB} = (\mathrm{blockdiag}(\mathbf{MFIM}))^{-1} \neq \mathbf{mCRB}$.

$$\mathbf{mCRB}_{VB} = \mathcal{V}_f(\mathcal{V}_f^H(\mathbf{A}_{\mathbf{f},\mathbf{f}})^{-1}\mathcal{V}_f)^{-1}\mathcal{V}_f^H + \widetilde{\overline{\boldsymbol{\theta}}} \widetilde{\overline{\boldsymbol{\theta}}}^H$$
$$\mathbf{A} = \mathcal{V}(\mathcal{V}^H \overline{\mathbf{Q}} \mathcal{V})^{-1} \mathcal{V}^H \mathbf{J}_q \mathcal{V}(\mathcal{V}^H \overline{\mathbf{Q}} \mathcal{V})^{-1} \mathcal{V}^H, \tag{22}$$

$\mathbf{A}$ evaluated at $\overline{\boldsymbol{\theta}}$, $\mathbf{A}_{\mathbf{ff}} = (\mathbf{f}, \mathbf{f})$ block of $\mathbf{A}$ (here it is the product of block diagonal of 3 factors), mCRB above for given $\boldsymbol{\theta}^o$. Some remarks which follow from our mCRB analysis are stated below.

- mCRB in this paper are along the lines of [3] and it is applicable to all estimators with same bias and cross-correlation matrix.
- This mCRB, is mismatched because we introduce an artificial prior. Asymptotically (i.e. at high SNR), the MSE of either alternating MAP (AMAP) or VB or EC-VB should match this mCRB.
- Asymptotically, the suboptimality of VB is not in its mean, it's only in the approximation of the error covariance, which should underestimate the actual error covariance: $[(\mathbf{J}_q)^{-1}]_{1,1} > ((\mathbf{J}_q)_{1,1})^{-1}$.
- Our view point of first working for given $\boldsymbol{\theta}$ is compatible with the view that actually the $\boldsymbol{\theta}$ may be be deterministic (prior for $\boldsymbol{\theta} = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^0)$, dirac delta function at true value) and the idea of doing Bayesian or VB is just to create a bias so that the biased estimator would reach lower MSE, in particular below the CRB. James-Stein estimator [15] was the first instance of this. In case of James-Stein, they are able to show that the deterministic MSE is lowered by adding the prior (with optimized/estimated variance hyperparameter). Then VB (with estimated = optimized hyperparameters) is a way of making sure that this bias is useful, optimizes MSE in some sense, within the class of estimators determined by the stucture of the prior chosen. In other words, these Bayesian estimators provide a way to introduce a useful bias (shrinkage) that allows to lower MSE (from the point of view of deterministic parameters, with a single true value).

B. Computation of $\overline{\boldsymbol{\theta}}$:

Starting from (14), the resulting (deterministic) $\overline{\boldsymbol{\theta}}$ is obtained by running alternating MAP (initialized by the true $\boldsymbol{\theta}^o$). Or one can also run the VB, by putting $\mathbf{n} = 0$ in $\mathbf{y}$, and considering $\widetilde{\mathbf{h}} = 0, \widetilde{\mathbf{f}} = 0$,hence also $C_{\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}} = 0, C_{\widetilde{\mathbf{f}}\widetilde{\mathbf{f}}} = 0$. So, the VB converges to $\overline{\boldsymbol{\theta}}$. For computing $\overline{\mathbf{f}}$, substituting for $\mathbf{y} = \mathcal{H}^0 \mathbf{f}^0 + \mathbf{n}$ in (14) (similarly for the computation of $\overline{\mathbf{h}}$, need to consider the alternative representation of $\mathbf{y}$ (8)),

$$\mathrm{E}_{p(\mathbf{y}/\boldsymbol{\theta}) \ln < p(\mathbf{y},\boldsymbol{\theta},\boldsymbol{\psi}) >_{\bar{i}}} = -N_y \ln \sigma^2 - \frac{1}{\sigma^2}(< \left\| \mathcal{H}^0 \mathbf{f}^0 - \mathcal{H}\mathbf{f} \right\|^2 > + \sigma^{2,0} N_y) + (M-1) < \ln \alpha > - < \alpha > \|\mathbf{f}_\perp\|^2 + N_h < \ln \beta > - < \beta >< \|\mathbf{h}\|^2 > + c. \tag{23}$$

The derivative of (23) w.r.t $\mathbf{f}, \alpha, \beta, \mathbf{h}$ leads to Algorithm 2. Note that the Algorithm 2 applies to any partitioning of the variables in the approximate posterior q, where for VB (12), there will only be one iteration with the initial values for $\overline{\mathcal{H}}^H \overline{\mathcal{H}} = < \mathcal{H}^H \mathcal{H} >$ or $\overline{\mathcal{H}} = < \mathcal{H} >$ (by the converged values of VB).

Algorithm 2 Computation of Asymptotic Estimates, $\overline{\boldsymbol{\theta}}$

1: **Initialization:** Initialize $\overline{\mathbf{f}}$ using existing calibration methods ($\overline{\mathbf{f}} = \mathbf{f}' + \mathcal{V}_f \overline{\mathbf{f}}_\perp$).
2: **repeat**
3: $\quad \overline{\mathbf{f}}_\perp = (\mathcal{V}_f^H \overline{\mathcal{H}}^H \overline{\mathcal{H}} \mathcal{V}_f + \sigma^2 \overline{\alpha} \mathbf{I})^{-1} (\mathcal{V}_f^H \overline{\mathcal{H}}^H \mathcal{H}^0 \mathbf{f}^0 - \mathcal{V}_f^H \overline{\mathcal{H}}^H \overline{\mathcal{H}} \mathbf{f}')$.
4: $\quad \overline{\mathbf{h}} = (\overline{\mathcal{F}}^H \overline{\mathcal{F}} + \sigma^2 \overline{\beta} \mathbf{I})^{-1} (\overline{\mathcal{F}}^H \mathcal{F}^0 \mathbf{h}^0)$.
5: $\quad \overline{\alpha} = \frac{M}{<\|\overline{\mathbf{f}}_\perp\|^2>}$, $\overline{\beta} = \frac{N_h+1}{<\|\overline{\mathbf{h}}\|^2>}$, $\sigma^2 = \sigma^{2,0} + \frac{1}{N_y} \left\| \mathcal{H}^0 \mathbf{f}^0 - \overline{\mathcal{H}} \overline{\mathbf{f}} \right\|^2$.
6: **until** convergence.

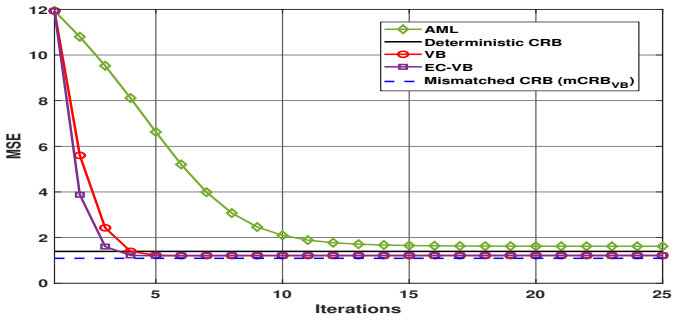

Fig. 2. Convergence of the various iterative schemes for $M = G = 16$.

IV. Simulations

In this section, we assess numerically the performance of various calibration algorithms and also compare them against their CRBs. The Tx and Rx calibration parameters for the BS antennas are assumed to have random phases uniformly distributed over $[-\pi, \pi]$ and amplitudes uniformly distributed in the range $[1-\delta, 1+\delta]$. SNR is defined as the ratio of the average received signal power across channel realizations at an antenna and the noise power at that antenna. In Fig. 2, it is clear that VB MSE can go lower than the determinsitic CRB and close to the mCRB. In Fig. 3, we compare the MSE performance of various VB variants with $mCRB_{VB}$ and deterministic CRB. It shows the performance improvement of VB w.r.t deterministic CRB or AML at all SNR and also the accurate behaviour of our derived mCRB expressions. We consider transmit schemes that transmit from one antenna at a time ($G = M$) and compare their MSE performance with the CRB. The MSE with FCC for Argos, Rogalin [16] and the VB method in Algorithm 1 is plotted. The curves are generated over one realization of an i.i.d. Rayleigh channel and known first coefficient constraint is used. These curves are compared with the CRB derived in II-A for the FCC case and it can be seen that the AML curve overlaps with the CRB at higher SNRs. Also plotted is the CRB as given in [14] assuming the internal propagation channel is fully known (the mean is known and the variance is negligible) and a (small) underestimation of the MSE can be observed as expected.

V. Conclusions

In this paper, we came up with a simple and elegant derivation of the mCRB for a general calibration framework that includes as subsets all existing calibration techniques. For the case of groups involving a single antenna, the conventional CRB derivation assuming first coefficient known has also been provided. An optimal estimation algorithm based on VB is also introduced along with its variants. We further derived mismatched CRB to validate the performance improvement over deterministic CRB. All these techniques have been compared via simulations in terms of both MSE performance and speed of convergence.

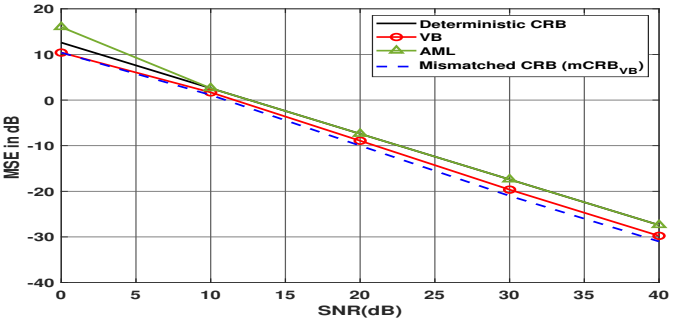

Fig. 3. Comparison of single antenna transmit schemes with the CRB ($G = M = 16, L_i = 1, \forall i, \delta = 0.5$).

References

[1] C. Shepard, N. Yu, H.and Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proc. ACM Intern. Conf. Mobile Computing and Netw. (Mobicom)*, Istanbul, Turkey, Aug. 2012.

[2] X. Jiang, A. Decurninge, K. Gopala, F. Kaltenberger, M. Guillaud, D. Slock, and L. Deneire, "A Framework for Over-the-air Reciprocity Calibration for TDD Massive MIMO Systems," *IEEE Trans. Wireless Commun.*, Sept. 2018.

[3] C. D. Richmond, and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," *IEEE Trans. on Sig. Process.*, May. 2015.

[4] K. Gopala and D. Slock, "Optimal algorithms and CRB for reciprocity calibration in massive MIMO," in *ICASSP*, Apr., 2018.

[5] E. de Carvalho and D. Slock, "Semi–Blind Methods for FIR Multichannel Estimation," in *SPAWC*, Apr., 2000.

[6] E. de Carvalho, S.M. Omar, and D. Slock, "Performance and complexity analysis of blind FIR channel identification algorithms based on deterministic maximum likelihood in SIMO systems," *Cir., Sys., and Sig. Process.*, vol. 34, no. 4, Aug. 2012.

[7] E. de Carvalho and D. Slock, "Blind and semi-blind FIR multichannel estimation: (Global) identifiability conditions," *IEEE Trans. Sig. Proc.*, Apr. 2004.

[8] E. de Carvalho and D. Slock, "Cramér-Rao bounds for blind multichannel estimation," arXiv:1710.01605 [cs.IT], 2017.

[9] E. de Carvalho, J.M. Cioffi, and D. Slock, "Cramér-Rao bounds for blind multichannel estimation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, CA, USA, Nov. 2000.

[10] V. Smdl, and A. Quinn,, " The variational Bayes method in signal processing.," in *New York: Springer-Verlag,* , 2005.

[11] M. Opper and O. Winther, "Expectation Consistent Approximate Inference," *J. Mach. Learn. Res.*, vol. 6, Dec. 2005.

[12] S. Fortunati, F. Gini, M.S. Greco, and C.D. Richmond, "Performance Bounds for Parameter Estimation under Misspecified Models [Fundamental findings and applications]," *IEEE Sig. Proc. Mag.*, Nov. 2017.

[13] J.M. Kantor, C.D. Richmond, B. Correll Jr., D.W. Bliss, "Prior Mismatch in Bayesian Direction of Arrival Estimation for Sparse Arrays," in *Proc. RadarCon*, 2015.

[14] Joao Vieira, Fredrik Rusek, Ove Edfors, Steffen Malkowsky, Liang Liu, and Fredrik Tufvesson, "Reciprocity Calibration for Massive MIMO: Proposal, Modeling and Validation," *IEEE Trans. Wireless Commun.*, May 2017.

[15] W. James, and C. M. Stein, "Estimation with Quadratic Loss ," *Breakthroughs in Statistics, Springer, New York, NY*, 1992.

[16] R. Rogalin, et. al., "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *IEEE Trans. Wire. Commun.*, vol. 13, no. 4, Apr. 2014.