MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Misspecified Bayesian Cramer-Rao Bound for Sparse Bayesian Learning

Pajovic, M.

TR2018-076 July 12, 2018

Abstract

We consider a misspecified Bayesian Cramer-Rao bound (MBCRB), justified in a scenario where the assumed data model is different from the true generative model. As an example of this scenario, we study a popular sparse Bayesian learning (SBL) algorithm where the assumed data model, different from the true model, is constructed so as to facilitate a computationally feasible inference of a sparse signal within the Bayesian framework. Formulating the SBL as a Bayesian inference with a misspecified data model, we derive a lower bound on the mean square error (MSE) corresponding to the estimated sparse signal. The simulation study validates the derived bound and shows that the SBL performance approaches the MBCRB at very high signal-to-noise ratios.

IEEE Statistical Signal Processing (SSP) Workshop

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2018
201 Broadway, Cambridge, Massachusetts 02139

MISSPECIFIED BAYESIAN CRAMÉR-RAO BOUND FOR SPARSE BAYESIAN LEARNING

Milutin Pajovic

Mitsubishi Electric Research Laboratories, Cambridge, MA, USA

ABSTRACT

We consider a misspecified Bayesian Cramér-Rao bound (MBCRB), justified in a scenario where the assumed data model is different from the true generative model. As an example of this scenario, we study a popular sparse Bayesian learning (SBL) algorithm where the assumed data model, different from the true model, is constructed so as to facilitate a computationally feasible inference of a sparse signal within the Bayesian framework. Formulating the SBL as a Bayesian inference with a misspecified data model, we derive a lower bound on the mean square error (MSE) corresponding to the estimated sparse signal. The simulation study validates the derived bound and shows that the SBL performance approaches the MBCRB at very high signal-to-noise ratios.

Index Terms— Bayesian Cramér-Rao bound, misspecified model, sparse Bayesian learning, mean square error

1. INTRODUCTION

Recovering sparse signals from limited number of noisy measurements has received considerable attention in the literature [1]. Along with the greedy [2] and convex optimization-based [3] recovery algorithms, Bayesian formulation of the sparse recovery problem has also been considered [4, Ch. 13]. Although the Bayesian sparse recovery framework is built upon an intuitively appealing generative model for the sparse signal, its main difficulty is computationally prohibitive inference. Consequently, much of the effort within the Bayesian sparse recovery framework has been geared towards formulating generative models for sparse signals which render computationally feasible inference [5]. In that regard, sparse Bayesian learning (SBL) stands out as a particularly handy and popular approach [6].

The SBL imposes on the unknown sparse signal a particular generative model with approximately sparse realizations. This potentially sacrifices signal recovery performance with the benefit of alleviating the computational burden of the inference. Essentially, the SBL prior can be seen as a misspecified model for the unknown sparse signal. In general, model mismatch is quite common in signal processing, and arises from simplifying or not fully understanding data generation mechanism, or as a result of data modeling with the aim to enable feasible or simpler computations. A theoretical treatment of the model mismatch, in particular the development of Cramér-Rao bound for misspecified data models, has relatively recently started to gain interest in signal processing community. An accessible background on this topic and an overview of important applications of the theory on misspecified bounds to signal processing models is provided in [7] and references therein.

In this paper, we derive a mean square error (MSE) bound corresponding to the SBL-based sparse signal recovery. In doing so, we view the SBL model as a misspecified Bayesian model for the sparse signal and build upon the only two existing works (to the best of our knowledge and also as claimed in [7]) on misspecified Bayesian Cramér-Rao bound (MBCRB) [8, 9]. In particular, we provide a concise development of the MBCRB from those papers, with corrections of some minor inconsistencies. Then, we apply the derived MCRB to the SBL model for sparse signal recovery. Finally, we validate the bound with numerical simulations and provide insights about the SBL performance.

2. SPARSE BAYESIAN LEARNING

A linear underdetermined measurement model represents vector of M measurements, $\mathbf{y} \in \mathbb{C}^{M\times 1}$, as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v} \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{M\times N}$ is a (known) measurement matrix with $M < N$, $\mathbf{x} \in \mathbb{C}^{N\times 1}$ is vector representation of an unknown sparse signal and $\mathbf{v} \in \mathbb{C}^{N\times 1}$ is noise, here assumed circularly symmetric Gaussian distributed, *i.e.*, $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2 \mathbf{I}_N)$. Denoting a prior distribution (*i.e.*, the generative model) of $\mathbf{x}$ with $p(\mathbf{x})$, the posterior distribution of $\mathbf{x}$ upon observing $\mathbf{y}$ is given by

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\; p(\mathbf{x}) \tag{2}$$

where the data likelihood $p(\mathbf{y}|\mathbf{x})$ directly follows from the Gaussian noise statistics. The prior $p(\mathbf{x})$ is specified so that it reflects the sparse nature of $\mathbf{x}$. For example, each entry x_i in $\mathbf{x}$ is, according to the Bernoulli-Gaussian model [4], independently generated as the product of realizations from Bernoulli and Gaussian random variables. As such, the parameter r of the Bernoulli distribution controls the sparsity of $\mathbf{x}$ in a sense that smaller r yields smaller expected number of non-zero entries in $\mathbf{x}$. Although this model naturally generates a sparse realization of $\mathbf{x}$, the associated Bayesian inference problem (2) is a combinatorial problem with exponential complexity in N.

The SBL is a hierarchical Bayesian modeling framework developed to overcome computational difficulties arising from the Bernoulli-Gaussian and alike models $p(\mathbf{x})$. The SBL is framed around a sparsity-promoting prior $q(\mathbf{x})$, whose realizations are "softly" sparse in a sense that most entries are small in magnitude and close to zero. This is in contrast with realizations of $p(\mathbf{x})$ which are purely sparse, with the majority of entries being exactly equal to zero. While the SBL approximates the generative model of $\mathbf{x}$, it enables a computationally feasible inference of $\mathbf{x}$.

More specifically, the SBL assumes each entry in $\mathbf{x}$ is distributed according to a zero-mean Gaussian distribution whose inverse variance is a sample from a Gamma distribution. Formally,

$$q(\mathbf{x}|\boldsymbol{\alpha}) = \mathcal{CN}(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma}_\alpha) \tag{3}$$

with $\boldsymbol{\Sigma}_\alpha = \text{diag}^{-1}\{\boldsymbol{\alpha}\}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{N\times 1}$ is a precision vector. The entries in the precision vector $\alpha_i, i = 1, \ldots, N$, are independent and Gamma distributed

$$q(\alpha_i) = \text{Gamma}(\alpha_i; a, b) = \frac{b^a}{\Gamma(a)} \alpha_i^{a-1} e^{-b\alpha_i} \tag{4}$$

where a and b are, respectively, the shape and rate parameter, usually chosen such that the distribution of the precision vector is non-informative, *i.e.*, does not favor any particular realization of $\boldsymbol{\alpha}$. The intuition behind the SBL model is that the resulting prior distribution for each entry x_i, $q(x_i)$, is strongly peaked around 0 and exhibits heavy tails away from 0. In other words, it is a soft version of the hard constraint that most of x_i's are zero, as explicitly modeled with $p(\mathbf{x})$. This can be seen from a closed-form expression for $q(x_i)$ when $a = 1$, which leads to an exponentially distributed precision α_i with parameter b, such that

$$q(x_i) = \int_0^\infty q(x_i|\alpha_i) q(\alpha_i) d\alpha_i = \frac{\pi}{b} \frac{1}{(|x_i|^2 + b)^2} \tag{5}$$

We assume $a = 1$ and utilize (5) in Section 4 for analytical tractability. This is also justified because with small b, the resulting prior on α_i is non-informative, as desired. In general, the SBL also places a prior on variance σ_v^2. However, we assume it is known and, thus, the posterior distribution of $\mathbf{x}$ and $\boldsymbol{\alpha}$ is

$$q(\mathbf{x}, \boldsymbol{\alpha}|\mathbf{y}) \propto q(\mathbf{y}|\mathbf{x})\; q(\mathbf{x}|\boldsymbol{\alpha})\; q(\boldsymbol{\alpha}) \tag{6}$$

where the likelihood model $q(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$ and $q(\mathbf{x}) = \prod_{i=1}^N q(x_i)$. To estimate $\mathbf{x}$, the expectation-maximization (EM) algorithm alternates between computing a point estimate of $\boldsymbol{\alpha}$ and inferring $q(\mathbf{x}|\mathbf{y}; \boldsymbol{\alpha})$ [6]. Alternatively, the Variational Bayes (VB) approximates the posteriors on $\mathbf{x}$ and $\boldsymbol{\alpha}$ using the mean field approximation (MFA) approach [6], attaining a similar recovery performance as the EM algorithm.

While a number of reported simulation studies have shown that the SBL succeeds to correctly capture the sparse support of $\mathbf{x}$, to the best of our knowledge, there are no theoretical results in this domain. Using some recent results that characterize bounds for misspecified models, we derive a MSE bound corresponding to the estimation of $\mathbf{x}$ under the misspecified model $q(\mathbf{x})$. We emphasize that the presented analysis is derived for a general model $p(\mathbf{x})$.

3. MISSPECIFIED BAYESIAN BOUND

This part summarizes important results on the misspecified Bayesian Cramér-Rao bound (MBCRM) from [8] and [9]. Aside from providing a concise summary of the MBCRB derivation, we slightly expand on some derivation steps outlined in those references and correct minor inconsistencies.

A common approach in the derivation of bounds is to assume that data and parameters are real-valued. Since all quantities in the considered problem are complex-valued, we present the background on the MBCRB assuming complex space. The derivation of the Cramér-Rao bound involves taking derivatives of a real-valued function (logarithm of the data likelihood) with respect to a complex-valued $\mathbf{x}$. This is done by treating the real-valued function as a function of $\mathbf{x}$ and $\mathbf{x}^*$ (complex conjugate of $\mathbf{x}$) and taking the derivative with respect to $\mathbf{x}^*$, assuming $\mathbf{x}$ is an independent variable [10].

We define the parameter vector $\boldsymbol{\theta}$ to comprise of the unknown $\mathbf{x}$ and its complex-conjugate $\mathbf{x}^*$, that is

$$\boldsymbol{\theta} = \begin{bmatrix} \mathbf{x}^T & \mathbf{x}^H \end{bmatrix}^T \tag{7}$$

Given that $\boldsymbol{\theta}$ does not contain additional information aside from $\mathbf{x}$, note that replacing $\mathbf{x}$ with $\boldsymbol{\theta}$ makes no change in the corresponding probability distribution. For example, $p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}, \mathbf{x})$ and we interchangeably use $\boldsymbol{\theta}$ and $\mathbf{x}$ in the expressions for probability distributions.

Denoting with $\hat{\boldsymbol{\theta}}_q(\mathbf{y})$ an estimate of $\boldsymbol{\theta}$, obtained from the measurement $\mathbf{y}$ under the assumed generative model $q(\mathbf{y}, \mathbf{x})$, the estimation error is defined as

$$\boldsymbol{\epsilon}(\mathbf{y}, \boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}_q(\mathbf{y}) - \boldsymbol{\theta}, \tag{8}$$

where $\boldsymbol{\theta}$ is true value of the unknown parameter. The estimator's conditional mean is evaluated by taking the expectation of $\hat{\boldsymbol{\theta}}_q(\mathbf{y})$ with respect to $p(\mathbf{y}|\boldsymbol{\theta})$,

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})} \left[\hat{\boldsymbol{\theta}}_q(\mathbf{y})\right] \tag{9}$$

The correlation matrix of the estimation error $\boldsymbol{\epsilon}$ can be expressed as

$$\mathbb{E}_p \left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H\right] = \mathbb{E}_p \left[\boldsymbol{\xi}\boldsymbol{\xi}^H\right] + \mathbb{E}_{p(\boldsymbol{\theta})} \left[\mathbf{b}\mathbf{b}^H\right] \tag{10}$$

where, to keep notation uncluttered, $p \triangleq p(\mathbf{y}, \boldsymbol{\theta})$, and the error term $\boldsymbol{\xi}$ and bias $\mathbf{b}$ are, respectively, defined as

$$\boldsymbol{\xi}(\mathbf{y}, \boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}_q(\mathbf{y}) - \boldsymbol{\mu}(\boldsymbol{\theta}) \tag{11}$$

$$\mathbf{b}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta} \tag{12}$$

In the following, we lower bound the correlation matrix of the error term $\boldsymbol{\xi}$. This is done by using the matrix generalization of the Cauchy-Schwarz inequality so that [8, 9]

$$\mathbb{E}_p\left[\boldsymbol{\xi}\boldsymbol{\xi}^H\right] \succeq \mathbb{E}_p\left[\boldsymbol{\xi}\boldsymbol{\eta}^H\right]\mathbb{E}_p\left[\boldsymbol{\eta}\boldsymbol{\eta}^H\right]^{-1}\mathbb{E}_p\left[\boldsymbol{\eta}\boldsymbol{\xi}^H\right] \quad (13)$$

where $\mathbf{X} \succeq \mathbf{Y}$ means $\mathbf{X}-\mathbf{Y}$ is a positive semi-definite matrix and $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathbf{y}, \boldsymbol{\theta})$ is the score function which controls the tightness of the bound [7]. A detailed derivation of (13) is presented in [11] for the case of deterministic parameter $\boldsymbol{\theta}$ when all expectations in (13) are taken with respect to $p(\mathbf{y}; \boldsymbol{\theta})$. A generalization to the Bayesian setting is done by taking the expectations with respect to $p(\mathbf{y}, \boldsymbol{\theta})$.

As a general suggestion, the score function is selected so that the lower bound (13) is as tight as possible. However, there is no definite rule as to how to specify $\boldsymbol{\eta}$. The studies on misspecified bounds commonly choose

$$\boldsymbol{\eta}(\mathbf{y}, \boldsymbol{\theta}) = \frac{\partial \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} - \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})}\left[\frac{\partial \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right] \quad (14)$$

The motivation for this choice for $\boldsymbol{\eta}$ stems from its relationship to the classical CRB [11]. Namely, in the case of a perfectly specified generative model (that is, $p(\mathbf{y}, \boldsymbol{\theta}) \equiv q(\mathbf{y}, \boldsymbol{\theta})$) and assuming the second term of (14) is zero (which is a well-known regularity condition for deterministic CRB), the deterministic CRB is obtained from (13), where the score function is the first term of (14). Consequently, the score function $\boldsymbol{\eta}$ for the misspecified Bayesian setting is given as the score function which yields the deterministic CRB, minus its (possibly) non-zero mean under the true model $p(\mathbf{y}|\boldsymbol{\theta})$ [11].

Having specified $\boldsymbol{\eta}$, we turn our attention to $\boldsymbol{\xi}$. In general, we may constrain the cross-correlation matrix $\mathbb{E}_p\left[\boldsymbol{\xi}\boldsymbol{\eta}^H\right]$ and obtain a bound valid for all estimators satisfying such a constraint, as indicated in [9]. Instead, we directly approximate the error term $\boldsymbol{\xi}$ using the approach from [8]. In that regard, it is assumed that $\hat{\boldsymbol{\theta}}_q(\mathbf{y})$ is the MAP estimate, obtained from maximizing the posterior $q(\boldsymbol{\theta}|\mathbf{y}) \propto q(\mathbf{y}, \boldsymbol{\theta})$. In addition, the MAP estimate is assumed to be in the vicinity of $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\boldsymbol{\theta})$. The Taylor series expansion of the data log-likelihood function around $\boldsymbol{\mu}$ is given by

$$\log q(\mathbf{y}, \boldsymbol{\mu} + \Delta\boldsymbol{\theta}) = \log q(\mathbf{y}, \boldsymbol{\mu}) + \left.\frac{\partial \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right|_{\boldsymbol{\mu}} \Delta\boldsymbol{\theta} + \frac{1}{2}\left.\frac{\partial^2 \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T}\right|_{\boldsymbol{\mu}} \Delta\boldsymbol{\theta}^H\boldsymbol{\theta} + O(\|\Delta\boldsymbol{\theta}\|^2) \quad (15)$$

Neglecting the higher-order terms in the Taylor series expansion (15), equating to zero the first derivative of the resulting expression with respect to $\Delta\boldsymbol{\theta}$, and solving for $\Delta\boldsymbol{\theta}$ yields an approximation for the error term

$$\boldsymbol{\xi}(\mathbf{y}, \boldsymbol{\theta}) \approx -\left(\left.\frac{\partial^2 \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T}\right|_{\boldsymbol{\mu}}\right)^{-1} \left.\frac{\partial \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right|_{\boldsymbol{\mu}} \quad (16)$$

Note that $\boldsymbol{\xi}$ is still function of $\boldsymbol{\theta}$ because the mean $\boldsymbol{\mu}$ is, in general, a function of $\boldsymbol{\theta}$.

Substituting (14) and (16) into (13) yields lower bound for the correlation matrix of the error term $\boldsymbol{\xi}$. The sum of this bound and the term related to the bias $\mathbf{b}$, yields the lower bound on the correlation matrix of the estimation error, $\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H\right]$. The derived bound applies to all MAP estimators, of mean $\boldsymbol{\mu}$, obtained from the observed data $\mathbf{y}$ under the assumed model $q(\mathbf{y}, \mathbf{x})$. The requirement on the estimator's mean limits the applicability of the derived bound. Derivation of a bound not subject to this limitation is viewed as a next important development in this area [7].

4. MSE BOUND FOR SBL MODEL

We derive in this part the MBCRB corresponding to the SBL model $q(\mathbf{y}, \boldsymbol{\theta})$. We start with evaluating the score function $\boldsymbol{\eta}$. Taking the first derivative of the data log-likelihood yields the expression for the first term in (14),

$$\mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) \triangleq \frac{\partial \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} = -\frac{1}{\sigma_v^2}\begin{bmatrix} \mathbf{B}\mathbf{x} - \mathbf{A}^H\mathbf{y} \\ (\mathbf{B}\mathbf{x} - \mathbf{A}^H\mathbf{y})^* \end{bmatrix} \quad (17)$$

where $\mathbf{B} = \mathbf{A}^H\mathbf{A}$. Noting that $\mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})}\left[\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})\right] = 0$, the score function is given by

$$\boldsymbol{\eta}(\mathbf{y}, \boldsymbol{\theta}) = \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) \quad (18)$$

The correlation matrix of $\boldsymbol{\eta}$, needed for the bound in (13), is after algebraic computations obtained as

$$\mathbb{E}_p\left[\boldsymbol{\eta}(\mathbf{y}, \boldsymbol{\theta})\boldsymbol{\eta}^H(\mathbf{y}, \boldsymbol{\theta})\right] = \frac{1}{\sigma_v^2}\begin{bmatrix} \mathbf{B} & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times N} & \mathbf{B}^T \end{bmatrix} \triangleq \mathbf{F} \quad (19)$$

To evaluate the error term $\boldsymbol{\xi}$ using (16), we assume $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. That is, the bound derived here holds for all estimators whose mean with respect to $p(\mathbf{y}|\boldsymbol{\theta})$ is equal to the true value of the unknown parameter $\boldsymbol{\theta}$. Hence, the term with the first derivative in (16) is given by

$$\begin{aligned} \frac{\partial \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} &= \frac{\partial \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} + \frac{\partial \log q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \\ &= \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{g}(\boldsymbol{\theta}), \end{aligned} \quad (20)$$

where $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ is given by (17) and the k-th entry in $\mathbf{g}(\boldsymbol{\theta})$ is using (5) computed as

$$[\mathbf{g}(\boldsymbol{\theta})]_k = -2 \times \begin{cases} \frac{x_k}{|x_k|^2 + b}, & k = 1, \ldots, N \\ \frac{x_{k-N}^*}{|x_{k-N}|^2 + b}, & k = N+1, \ldots, 2N \end{cases} \quad (21)$$

The term with the second derivative in (16) is given by

$$\frac{\partial^2 \log q(\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} = \frac{\partial^2 \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} + \frac{\partial^2 \log q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} \quad (22)$$

The second derivative of the data log-likelihood in (22) is computed by taking the first derivative of $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ from (17). After some algebraic manipulation this evaluates to

$$\frac{\partial^2 \log q(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} = -\frac{1}{\sigma_v^2} \begin{bmatrix} \mathbf{B} & \mathbf{0}_{N\times N} \\ \mathbf{0}_{N\times N} & \mathbf{B}^T \end{bmatrix} = -\mathbf{F} \quad (23)$$

The second derivative of $q(\boldsymbol{\theta})$ is using (21) evaluated as

$$\frac{\partial^2 \log q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \boldsymbol{\theta}^T} = \begin{bmatrix} -\boldsymbol{\Sigma}(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\Sigma}(\mathbf{x}) \end{bmatrix} \triangleq -\mathbf{G}(\boldsymbol{\theta}) \quad (24)$$

where $\boldsymbol{\Sigma}(\mathbf{x})$ is the diagonal matrix with the k-th diagonal entry given by

$$[\boldsymbol{\Sigma}(\mathbf{x})]_k = \frac{2b}{(|x_k|^2 + b)^2},\ k = 1, \dots, N \quad (25)$$

The error term $\boldsymbol{\xi}$ is then using (20), (22), (23) and (24) succinctly expressed as

$$\boldsymbol{\xi}(\mathbf{y}, \boldsymbol{\theta}) = -(\mathbf{F} + \mathbf{G}(\boldsymbol{\theta}))^{-1}(\mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{g}(\boldsymbol{\theta})) \quad (26)$$

The cross-correlation matrix between the error term $\boldsymbol{\xi}$ and the score function $\boldsymbol{\eta}$ is using (26) and (18), evaluated as

$$\begin{aligned} &\mathbb{E}_p\left[\boldsymbol{\xi}(\mathbf{y}, \boldsymbol{\theta})\boldsymbol{\eta}^H(\mathbf{y}, \boldsymbol{\theta})\right] = \\ &-\mathbb{E}_p\left[(\mathbf{F} + \mathbf{G}(\boldsymbol{\theta}))^{-1}(\mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{g}(\boldsymbol{\theta}))\mathbf{f}^H(\mathbf{y}, \boldsymbol{\theta})\right] = \\ &-\mathbb{E}_{p(\boldsymbol{\theta})}\left\{(\mathbf{F} + \mathbf{G}(\boldsymbol{\theta}))^{-1}\mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})}\left[\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})\mathbf{f}^H(\mathbf{y}, \boldsymbol{\theta})\right]\right\} \quad (27) \end{aligned}$$

where we utilize $\mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})}[\mathbf{f}(\boldsymbol{\theta})] = 0$. The derivation of (19) reveals that

$$\mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})}\left[\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})\mathbf{f}^H(\mathbf{y}, \boldsymbol{\theta})\right] = \mathbf{F} \quad (28)$$

Thus,

$$\mathbb{E}_p\left[\boldsymbol{\xi}(\mathbf{y}, \boldsymbol{\theta})\boldsymbol{\eta}^H(\mathbf{y}, \boldsymbol{\theta})\right] = -\mathbb{E}_{p(\boldsymbol{\theta})}\left[(\mathbf{F} + \mathbf{G}(\boldsymbol{\theta}))^{-1}\right]\mathbf{F} \quad (29)$$

Substituting (29) and (19) into (13) yields

$$\mathbb{E}_p\left[\boldsymbol{\xi}\boldsymbol{\xi}^H\right] \succeq \mathbb{E}_{p(\boldsymbol{\theta})}\left[(\mathbf{F} + \mathbf{G}(\boldsymbol{\theta}))^{-1}\right]\ \mathbf{F}\ \mathbb{E}_{p(\boldsymbol{\theta})}\left[(\mathbf{F} + \mathbf{G}(\boldsymbol{\theta}))^{-1}\right] \quad (30)$$

Since we limited the bound derivation for the estimators satisfying $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, the bias term $\mathbf{b} = \mathbf{0}$. Hence, the lower bound from (30) also lower bounds the correlation matrix of the estimation error, $\mathbf{E}_p\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^H\right]$.

The lower bound on the norm of the estimation error is obtained by taking the trace of the lower bound matrix in (30). According to the parameter vector definition (7), the norm of the error in estimating $\mathbf{x}$ is lower bounded by taking the trace of the upper-left $N \times N$ submatrix of the lower bound matrix in (30). Using (19) and (24), we finally obtain the MBCRB for the error in estimating $\mathbf{x}$

$$\mathbb{E}_p\left[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right] \geq \sigma_v^2\ \mathrm{tr}\left\{\mathbf{HBH}\right\} \quad (31)$$

where tr{ } denotes the trace operator and

$$\mathbf{H} = \mathbb{E}_{p(\mathbf{x})}\left[\left(\mathbf{B} + \sigma_v^2\boldsymbol{\Sigma}(\mathbf{x})\right)^{-1}\right] \quad (32)$$

Above, $\mathbf{H}$ is, in general, evaluated using Monte-Carlo simulations for a given true generative model $p(\mathbf{x})$.

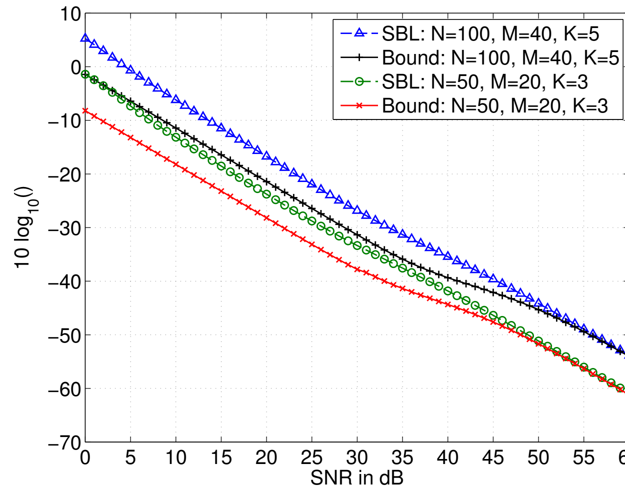

Fig. 1. Simulated MSE of the SBL vs. corresponding bound.

5. NUMERICAL STUDY

To validate the derived bound, we simulate the MSE performance of the SBL algorithm when used to recover N-dimensional sparse vector with K non-zero entries from M observations. We set $b = 10^{-5}$ in the SBL and employ the EM algorithm for inference, whose MAP estimate is the final estimate for $\mathbf{x}$. The measurement matrix $\mathbf{A}$ is pre-fixed by sampling its entries from $\mathcal{CN}(0, 1)$ and normalizing columns to the unit norm. Although the bound in (31) is valid for a general $p(\mathbf{x})$, the simulation study is done for a fixed sparse signal $\mathbf{x}_0$ so that $p(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_0)$, which admits a direct (without simulations) computation of the derived bound. More specifically, we (randomly) generate $\mathbf{x}_0$ and keep it fixed over 1,000 Monte-Carlo runs, randomized over AWGN realizations for a given SNR, defined as SNR $= \|\mathbf{x}_0\|^2/M\sigma_v^2$. The comparison between the SBL and the corresponding bound for two different sets of parameters N, M and K in shown in Fig. 1. As can be seen, the SBL performance is within few dBs with respect to our bound and approaches it at large SNR values.

6. CONCLUSIONS AND FUTURE WORK

We present a concise derivation of the misspecified Bayesian Cramér-Rao bound and its application to the sparse Bayesian learning (SBL) framework. The numerical tests show the SBL performance is within few dB's from the bound and approaches it for large SNR values. As for future work, the derived bound can be generalized to the case of unknown noise variance. Also, the development of possibly tighter bounds at moderate SNR values is justified. In addition, alternatives to approximating the error term in the derivation of the MBCRB, which restricts its validity to a particular class of estimators, are needed. Finally, a more thorough bound validation is required.

7. REFERENCES

[1] Alfred M. Bruckstein, David L. Donoho, and Michael Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

[2] Joel Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[3] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[4] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[5] Raj Tejas Suryaprakash, Milutin Pajovic, Kyeong Jin Kim, and Philip Orlik, "Millimeter wave communications channel estimation via Bayesian group sparse recovery," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016.

[6] Michael E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Sep 2001.

[7] Stefano Fortunati, Fulvio Gini, Maria S. Greco, and Christ D. Richmond, "Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 142 – 157, Nov. 2017.

[8] Joshua M. Kantor, Christ D. Richmond, Bill Correll, and Daniel W. Bliss, "Prior mismatch in Bayesian direction of arrival for sparse arrays," *Proc. IEEE Radar Conference (RadarCon)*, pp. 811–816, May 2015.

[9] Christ D. Richmond and Prabahan Basu, "Bayesian framework and radar: On misspecified bounds and radar-communication cooperation," *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–4, June 2016.

[10] D. H. Brandwood, "A complex gradiaent operator and its application in adaptive array theory," *IEE Proceedings F - Communications, Radar and Signal Processing*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[11] Christ D. Richmond and Larry L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2263–2278, May 2015.