Bayesian Framework for Sparse Vector Recovery and Parameter Bounds with Application to Compressive Sensing

by

Abdulhakim Alhowaish

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved October 2019 by the
Graduate Supervisory Committee:

Christ D. Richmond, Chair
Antonia Papandreou-Suppappola
Lalitha Sankar

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Signal compressed using classical compression methods can be acquired using brute force (i.e. searching for non-zero entries in component-wise). However, sparse solutions require combinatorial searches of high computations. In this thesis, instead, two Bayesian approaches are considered to recover a sparse vector from underdetermined noisy measurements. The first is constructed using a Bernoulli-Gaussian (BG) prior distribution and is assumed to be the true generative model. The second is constructed using a Gamma-Normal (GN) prior distribution and is, therefore, a different (i.e. misspecified) model. To estimate the posterior distribution for the correctly specified scenario, an algorithm based on generalized approximated message passing (GAMP) is constructed, while an algorithm based on sparse Bayesian learning (SBL) is used for the misspecified scenario. Recovering sparse signal using Bayesian framework is one class of algorithms to solve the sparse problem. All classes of algorithms aim to get around the high computations associated with the combinatorial searches. Compressive sensing (CS) is a widely-used terminology attributed to optimize the sparse problem and its applications. Applications such as magnetic resonance imaging (MRI), image acquisition in radar imaging, and facial recognition. In CS literature, the target vector can be recovered either by optimizing an objective function using point estimation, or recovering a distribution of the sparse vector using Bayesian estimation. Although Bayesian framework provides an extra degree of freedom to assume a distribution that is directly applicable to the problem of interest, it is hard to find a theoretical guarantee of convergence. This limitation has shifted some of researches to use a non-Bayesian framework. This thesis tries to close this gab by proposing a Bayesian framework with a suggested theoretical bound for the assumed, not necessarily correct, distribution. In the simulation study, a general lower Bayesian Cramér-Rao bound (BCRB) bound is extracted along with misspecified Bayesian Cramér-Rao bound (MBCRB) for GN model. Both bounds are validated using mean square error (MSE) performances of the aforementioned algorithms. Also, a quantification of the performance in terms of gains versus losses is introduced as one main finding of this report.

"You become. It takes a long time. That is why it doesn't happen often to people who break easily, or have sharp edges, or who have to be carefully kept. Generally, by the time you are Real, most of your hair has been loved off, and your eyes drop out and you get loose in the joints and very shabby. But these things don't matter at all, because once you are Real you can't be ugly, except to people who don't understand".

The Velveteen Rabbit

ACKNOWLEDGMENTS

I owe my deepest gratitude to Prof. Christ D. Richmond, my MS thesis advisor for his unparalleled guidance and support during this journey.

I am also grateful to my MS thesis committee members: Prof.Antonia Papandreou-Suppappola, and Prof Lalitha Sankar for their invaluable comments and discussion to improve my work and suggest useful directions for Ph.D. degree completion.

I am especially grateful to King Abdulaziz City for Science and Technology (KACST) for generously providing me scholarship to obtain my MS degree.

Special thanks to my very dear friends: Motaz Alfarraj, Meshaal Alyahya, and Mohammed Aladsani for being there for me when I needed them. I would also extend my appreciation and thanks to all my friends at Arizona State University who made me feel home away from home.

I owe my deepest gratitude to my family. I cannot thank you enough for your unconditional love and support. To my father Abdulrahman who raised me to be an honorable and honest man, my mother Sharifah who always surrounded me with her prayers, my three brothers Abdulaziz, Abdulmajeed, and Abdullah, my one and only sister Azizah. Thank you for your unconditional love and support.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

TABLE OF NOTATION

Definition	Notation	Comments
Random variables	X,Y	uppercase
Parameter space	Ω	uppercase Greek
Sample space	$\mathbb{S}, \mathbb{H}$	uppercase bold
Matrices	$\mathbf{A}, \mathbf{J}$	boldface uppercase
Vectors	$\mathbf{x}, \mathbf{y}$	boldface lowercase
Realizations (data)	x,y,z	lowercase
Parameters	α, λ	lowercase Greek
Hyperparameters	a, b, c, d	lowercase
CDFs	F(x),P(x)	uppercase
PDFs	f(x),p(x),q(x)	lowercase
Prior density	$\pi(\alpha)$	may be γ or ν
norms	$\|\|\mathrm{x}\|\|_p^p$	$l_p$norm $= \big(\sum\limits_i \|\mathrm{x}_i\|^p\big)^{1/p}$
Complexity measure	$\mathcal{O}(n)$	big "Oh" notation
Identity matrix	$\mathbf{I}_N$	$N \times N$ square matrix
Marginal likelihoods	p(x;a)	semicolon breaker
Conditional likelihoods	$\mathrm{p}(\mathrm{x}\|\alpha; a, b)$	parameters and hypers
Vector operators	$\mathbf{A}^T, \mathbf{A}^H, \mathbf{A}^{-1}$	real and complex
Famous distributions	$\mathcal{N}, \mathcal{CN}, \mathcal{B}$	to be specified
Special operators	$\mathbb{E}, C$	Expectation, Cost

Chapter 1

INTRODUCTION

1.1 Prologue

A sparse vector recovery problem seeks an approximate solution to an unknown vector $\mathbf{x}$

$$\text{Find sparse vector} \quad \mathbf{x} \quad \text{such that} \quad \mathbf{Ax} \approx \mathbf{y} \tag{1.1}$$

where $\mathbf{y}$ is the observed vector, and $\mathbf{A}$ is a real or complex matrix. $\mathbf{A}$ columns have a unit Euclidean norm: $\sum_{i=1}^{M} ||a_{ij}||_2 = 1$. $\mathbf{A}$ is often referred to as a *dictionary*, and its columns called *subdictionaries.* The target vector $\mathbf{x}$ is characterized by how many non zero elements it has. L-sparse vector $\mathbf{x}$ has at maximum L non-zero elements. (i.e. $||\mathbf{x}||_0 \leq L$). The counting operator $||\ .\ ||_0$ (sometimes called l-zero norm 'l_0', or pseudo-norm) returns the number of nonzero components of the argument. Observed vector $\mathbf{y}$ is interpreted based on the nature of the target vector $\mathbf{x}$. If $\mathbf{x}$ represents Dirac delta functions, then $\mathbf{y}$ is seen as sampled vector of the dictionary $\mathbf{A}$. Similarly, if $\mathbf{x}$ represents an indicator function of pixels, then $\mathbf{y}$ is an image acquisition of the dictionary. If $\mathbf{x}$ represents sunisoidal signal, then the observed vector $\mathbf{y}$ represents Fourier coefficients [1]. Figure 1.1 shows a visual representation of the aforementioned applications.

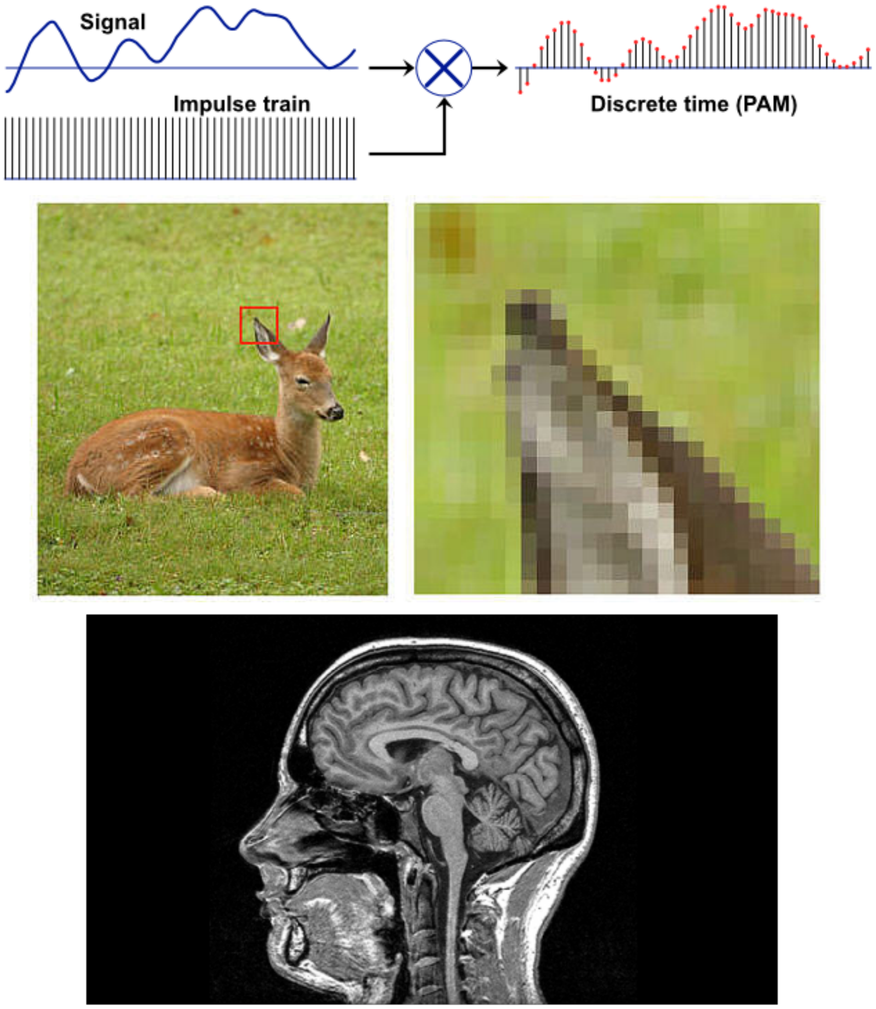

Figure 1.1: CS Representations Of The Observed Vector $\mathbf{y}$

In CS domain, signals tend to have redundant data (i.e compressible), rather than sparse. This redundancy can be reduced without loss of information. This reduction enables us to approximately recover signals sampled below the Nyquist rate. As in [2], it is often more difficult to identify approximate representations of compressible signals than of sparse signals. The intuitive solution of the sparse vector recovery problem is to solve for every possible permutation of the sparse target vector.

$$\arg\min_{\mathbf{x}} ||\mathbf{x}||_0 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \tag{1.2}$$

Obviously, this solution involves a high cost in terms of computational complexity. This constraint is often relaxed to have some tolerance. Now, the constraint is subject to the the least square with some error ϵ

$$\arg\min_{\mathbf{x}} ||\mathbf{x}||_0 \quad \text{subject to} \quad ||\mathbf{Ax}-\mathbf{y}||_{\mathbf{2}} \leq \epsilon \ . \tag{1.3}$$

An alternative constraint can be added on the least square argument depending on the application, by replacing the error term with the count of non-zero elements L.

$$\arg\min_{\mathbf{x}} ||\mathbf{Ax}-\mathbf{y}||_{\mathbf{2}} \quad \text{subject to} ||\mathbf{x}||_0 \leq L \tag{1.4}$$

Previous set ups corresponds to known problems in the literature. The goal here is to reduce the complexity resulting from applying equation (1.3). Brute force method results in a combinatorial complexity that corresponds to the binomial coefficient. Assuming that there are N realizations in the sparse vector. Typically, a $\binom{N}{L}$ computations needed to check every single sparse permutation of the vector $\mathbf{x}$.

1.2 Related Work

To exploit the redundancy of a non-structured dictionary, many classes of algorithms have been created to solve the sparse vector recovery problem. Following is a summary of the main five classes [2].

1.2.1 Brute Force

At certain situations, it is reasonable to check for sparsity in each single entry and constraint it with any of the above constraints. However, this approach is not practical as problem dimensions expands. However, this method will be the used as a reference to evaluate complexity. This method is guaranteed to reach the exact number of sparsity in the target vector for the price of having the highest computational cost [3].

1.2.2 Greedy Pursuit

Greedy pursuit is an iterative algorithm aims to find a set of weights that achieve the minimum difference compared to the target signal. A well-known method that belongs to this class of algorithms is called matching pursuit (MP) [4]. MP is an algorithm that defines weights (w) successively until it finds the best inner product that achieves the minimum residual R.

$$R^{\tau} = f - \hat{f}^{\tau} \tag{1.5}$$

$$f = \sum_{i=1}^{M} y_i \quad , \quad \hat{f}^{\tau} = \sum_{i=1}^{N} w_i^{\tau} \mathbf{a_i} \tag{1.6}$$

where $\mathbf{a_i}$ corresponds to the columns of the dictionary matrix $\mathbf{A}$.

1.2.3 Linear Optimization

Linear optimization is a class of optimization problems that deals with linear objective functions. One of the famous algorithms that fall in this category is basis pursuit (BP) [5]. BP aims to solve the optimization problem mentioned in equation (1.2) but with taking l_1-norm instead of l_o-norm.

$$\arg\min_{\mathbf{x}} ||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{Ax} = \mathbf{y} \tag{1.7}$$

This method does not assume a noisy set up. This makes the constraint centered around the basis of the dictionary. Hence the name of the algorithm.

1.2.4 Convex Optimization

Convex optimization class of problems, conversely, deals with any non-linear assumed objective function. Adding noise to equation (1.7) and relaxing the constraint to the least square objective, changes the problem to be

$$\arg\min_{\mathbf{x}} ||\mathbf{x}||_1 \quad \text{subject to} \quad ||\mathbf{Ax} - \mathbf{y}||_{\mathbf{2}} \leq \epsilon \tag{1.8}$$

The above method is known as a basis pursuit denoising (BPDN). It can be thought of BPDN as BP corrupted with noise.

Another widely-used algorithm on this category is the least absolute shrinkage and selection operator (LASSO). LASSO has an objective function similar to the one in equation (1.8) but with a different optimization constraint

$$\arg\min_{\mathbf{x}} ||\mathbf{x}||_1 \quad \text{subject to} \quad ||\mathbf{Ax} - \mathbf{y}||_{\mathbf{2}} \leq L \tag{1.9}$$

Identically, LASSO can be represented using a Lagrange multiplier

$$\arg\min_{\mathbf{x}} ||\mathbf{Ax} - \mathbf{y}||_{\mathbf{2}} + \lambda ||\mathbf{x}||_1 \tag{1.10}$$

It can be seen from equations (1.8-9) that LASSO and BPDN are equivalent problems if ϵ in equation (1.8) was chosen to be the sparsity L.

1.2.5 Bayesian Framework

All the previous classes of algorithms assume a deterministic number of sparsity L and aim to find a point estimate of the target vector $\mathbf{x}$. However, this class of algorithms assumes that the target vector originally followed a prior distribution which favors sparsity. Then, approximate a target vector distributed following a posterior distribution, that is derived using the Bayesian framework.

As reflected from the title, this report uses Bayesian framework to design a prior distribution to recover an estimated sparse vector.

1.3 Bayesian Framework Historically

Thomas Bayes was an English Statistician and philosopher who has lived in the eighteenth century (1701-1761)AD. He first illustrated Bayes theorem in a paper that was published in 1764, where he assumed a conditional distribution of a binomial random variable, let's say X, conditioned on a parameter θ that was uniformly distributed. According to [6], Bayes had not reflect on the posterior distribution. Based on [6] the first scientist who had addressed the wide applicability of the posterior distribution is the French scholar Pierre-Simon Laplace (1749-1827)AD.

Aiming to introduce the Bayesian framework, assume that there exist a random variable X that belongs to a parameterized family of distributions and let's call its density $p(x)$. Then, a number of realizations of X are gathered in a vector $\mathbf{x}$. We are interested in some parameter represented by those realizations (i.e. mean, variance, or standard deviation). Let's call this parameter of interest θ. θ by itself belongs to a parameterized family of distributions let's call its density $\pi(\theta)$. Both of the marginal distributions and their joint distribution belongs to a sample space Ω. Then, by definition, a likelihood distribution that summarize the relation between the realizations $\mathbf{x}$ and the parameter θ can be formalized as $p(\mathbf{x}|\theta)$. Bayes method states that we can find a posterior distribution for the parameter θ that encapsulates the information embedded in the realizations $\mathbf{x}$.

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\pi(\theta)}{p(\mathbf{x})} \tag{1.11}$$

Aiming for a relevant inference of the posterior distribution $p(\theta|\mathbf{x})$, a proper choice of the prior distribution must be addressed. The marginal distribution $p(\mathbf{x})$ can be seen as a mixture of conditional distribution on parameter θ, by the axiom of the total probability [7]. However, $p(\mathbf{x}|\theta)$ does not uniquely determine the mixing distribution $p(\theta)$.

$$p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta \tag{1.12}$$

Hence the distribution of the random variable θ must have a well defined range of values, and must be tractable with the lowest variance possible.

1.4 Motivation

One of the major applications that belong to CS is the MRI imaging. In MRI, the sensing waveform of the sparse vector are sinusoidal harmonics, and the observed vector $\mathbf{y}$ represents a Fourier coefficients. In this paradigm, data are assumed to be complex. Originally, MR imaging was not a subject of interest to DSP community until introducing a technique [8] called (SENSE). It was shown in [8] that there is a redundancy can be exploited for fast signal acquisition. As in [9] this redundancy of the data can be processed to reduce the sampling rate. However, MRI related research have a tendency towards non-Bayesian approaches. This tendency has many justifications. One is that CS originally started with non Bayesian classes of algorithms [10]. Another is that Bayesian framework has not ensure theoretical guarantees. In [11], it was clearly stated that the absence of a theoretical bound has made them avoid using Bayesian framework. Here I quote [11]: "*As the Bayesian method does not offer theoretical guarantees and the brute force method only remains plausible for small scale problems, we have applied only algorithms belonging to greedy pursuit,convex l_1 optimization and non-convex l_q optimization in this study*".

Motivated not only by this article that clearly states that the absence of a theoretical guarantee discouraged them from using a Bayesian method, but in the absence of reliable information, which is the case for many applications in CS, Bayesian method gives an extra degree of freedom in the design. The idea is that using a reliable prior, something that is not readily observable nor tractable, can give an insight about the original data specifications. So, with the help of the CRB lower bound, one can find a theoretical lower bound on the MSE of the estimate. Thus, and for the aforementioned justifications, this attempt can hopefully shows the potential of using the Bayesian framework as one effective methods in like applications.

The organization of this report is as follows. In chapter 2, an introduction to parameter estimation and CRB bound is given. In chapter 3, A formalization of the problem with the prior assumed distribution along with the calculated bounds is given. In chapter 4, Approximated posterior distribution for two different families of models. Also, numerical results of the MSE following both Baysian methods, along with complexity measure trade-off. Finally, chapter 5, is a conclusion with suggestions for future research.

This thesis aims to find a quantitative criterion to evaluate the goodness of an algorithm by defining a metric. This metric compute gains in terms of computational complexity and losses in terms in dBs. Then, it shows a trade-off between gains and losses for both algorithms used in this research.

Chapter 2

PARAMETER ESTIMATION AND PARAMETER BOUNDS

Statistics is a discipline where data is collected for the purpose of interpretation and inference. How the data is collected is not the main concern, rather than drawing conclusions about the parameters that the collected data represents [12]. Statistical inference is attributed to the situations assumed about the parameter of interest. A number of observations x either belongs to a non-parametric family of distributions and called **nonparameteric estimation**, or belongs to a parametric family of distributions and called parametric estimation. Nonparameteric estimation is useful in the absence of any guidance, constraints, or parameter of interest. A well-known field that uses nonparameteric estimation is neural networks or supervised learning. On the other hand, **parametric estimation** is where those observations $\mathbf{x}$ belongs to a parameter of interest. The parameter of interest θ can be assumed to be deterministic or to follow a known family of distributions. The former is often called **point estimation**, and aims to find a plausible value of θ. The latter is often referred to as **Bayesian estimation**. Bayesian estimation aims to find a distribution of θ, rather than the value of θ itself. Figure 2.1 shows a diagram of different statistical branches followed by a well-known optimization disciplines to estimate the desired parameter.

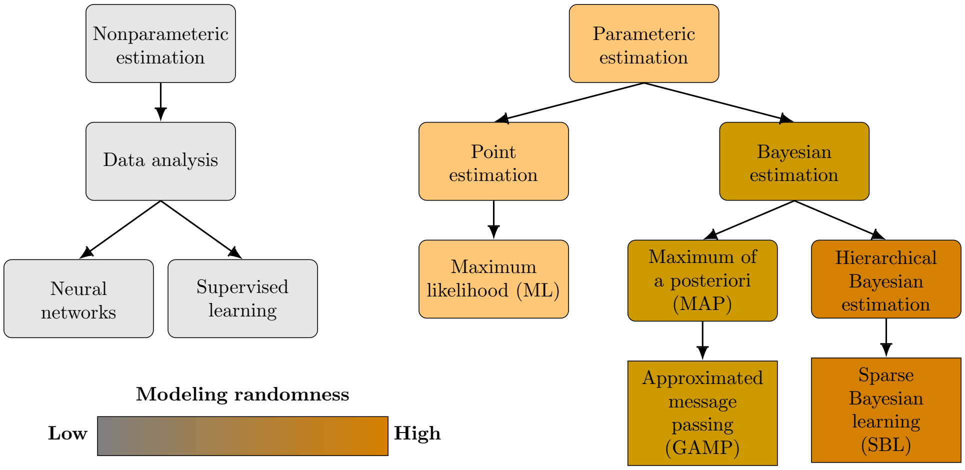

Figure 2.1: Branches Of Statistics And Corresponding Optimization Disciplines.

In general, estimation model must have **parameter space** (Ω) as the domain of values parameter can have, **observation space** ($\mathcal{X}$) as a finite-dimensional space whose points can be denoted by a random vector $\mathbf{X}$, **probabilistic mapping functions** from parameter space to observation space (i.e. cumulative distribution functions (CDFs) and there derivatives probability distribution functions (PDFs)), and finally **estimation rule** which can be denoted as $\hat{\theta}(\mathbf{x})$.

2.1 Estimation Cost and Risk Minimization

After defining a framework of a general estimation model, it is intuitive to search for a way to measure the quality of an estimation. This performance indicator will be between the true parameter value θ and the estimated value taken from the observations $\hat{\theta}(x)$. This measure can be defined as θ_ϵ, which is the difference between the true and estimated value of the parameter.

$$\theta_\epsilon(x) = \hat{\theta}(x) - \theta \tag{2.1}$$

A cost function can be defined as a customized function of the error that measures user satisfaction adequately and also one that results in a tractable problem. Once the cost function is defined, a good estimate can be achieved by minimizing the cost function.

Some of the widely-known cost functions are the squared error cost function

$$C(\theta_\epsilon) = {\theta_\epsilon}^2 \ . \tag{2.2}$$

The absolute value of the error cost function

$$C(\theta_\epsilon) = |\,\theta_\epsilon\,| \ . \tag{2.3}$$

The uniform error cost function

$$C(\theta_\epsilon) = \begin{cases} 0, & |\,\theta_\epsilon\,| \leq \frac{\Delta}{2}, \\ 1, & |\,\theta_\epsilon\,| \geq \frac{\Delta}{2}. \end{cases} \tag{2.4}$$

Risk can be defined as the expected value of any chosen cost function over the joint probability of the parameter and observation spaces $\mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)$, if the parameter follows a stochastic assumption, or the likelihood density $\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$,if the parameter is under the deterministic assumption. The prior distribution $\pi_\theta(\theta)$ and θ values are assumed to be known for each of the two cases. The following risk is for estimation within Bayesian framework.

$$R = \mathbb{E}\{C[\theta, \hat{\theta}(\mathrm{x})]\} = \int_{\text{-}\infty}^{\infty}\int_{\text{-}\infty}^{\infty} C[\theta, \hat{\theta}(\mathrm{x})]\, \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)\, dx\ d\theta \tag{2.5}$$

While the risk for non-Bayesian framework is

$$R = \mathbb{E}\{C[\theta, \hat{\theta}(\mathrm{x})]\} = \int_{\text{-}\infty}^{\infty}\int_{\text{-}\infty}^{\infty} C[\theta, \hat{\theta}(\mathrm{x})]\, \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\, dx\ d\theta \tag{2.6}$$

This risk concept is applicable for a general model of estimation. Naturally a minimum value of risk is desired. The following sections will try to optimize risk for Bayesian and non-Bayesian assumptions.

2.2 Bayesian Risk Analysis

Now, we try to explore integral in equation (2.4) for the three cost functions (2.1-3)

- The squared error risk optimization:

$$R_{ms} = \int_{\text{-}\infty}^{\infty}\int_{\text{-}\infty}^{\infty} [\theta - \hat{\theta}(\mathrm{x})]^2\, \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)\, dx\ d\theta$$

$$R_{ms} = \int_{\text{-}\infty}^{\infty} \mathrm{p}_{\mathrm{x}}(\mathrm{x})\, dx \int_{\text{-}\infty}^{\infty} [\theta - \hat{\theta}(\mathrm{x})]^2\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\theta\ .$$

Since both integrals of the last equation are positive quantities, minimizing the inner integral minimizes risk. This optimization step is called minimum mean squared error (MMSE). Let's define the value achieving the minimum as $\hat{\theta}(\mathrm{x})_{ms}$.

$$R_{ms}(\theta\,|x) = \int_{\text{-}\infty}^{\infty} [\theta - \hat{\theta}(\mathrm{x})]^2\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\theta$$

$$\frac{d}{d\hat{\theta}} R_{ms}(\theta\,|X) = \frac{d}{d\hat{\theta}} \int_{\text{-}\infty}^{\infty} [\theta - \hat{\theta}(\mathrm{x})]^2\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\theta$$

$$= -2\int_{\text{-}\infty}^{\infty} \theta\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta + 2\,\hat{\theta}(\mathrm{x}) \int_{\text{-}\infty}^{\infty} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta = 0$$

$$\Rightarrow \hat{\theta}(\mathrm{x})_{ms} = \int_{\text{-}\infty}^{\infty} \theta\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta = \boxed{\mathbb{E}_{\theta\,|X}\{\theta\}} \qquad (2.7)$$

The term inside the box is called the conditional mean estimator. This methodology of evaluating the estimate is used if we have limited information of the prior distribution, or we have access to the distribution of the parameter. [13].

- The absolute value of the error cost function:

The Bayes estimate of the absolute value cost function can be written as

$$R_{abs} = \int_{\text{-}\infty}^{\infty}\int_{\text{-}\infty}^{\infty} |\,\theta - \hat{\theta}(\mathrm{x})\,|\, \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)\, dx d\theta$$

$$R_{abs} = \int_{\text{-}\infty}^{\infty} \mathrm{p}_{\mathrm{x}}(\mathrm{x})\, dx \int_{\text{-}\infty}^{\infty} |\,\theta - \hat{\theta}(\mathrm{x})\,|\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\theta$$

doing the same as we did in the previous case, we minimize the inner integral

$$\begin{aligned}
R_{abs}(\theta\,|X) &= \int_{\text{-}\infty}^{\infty} |\,\theta - \hat{\theta}(\mathrm{x})\,|\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\theta \\
\frac{d}{d\hat{\theta}} R_{abs}(\theta\,|X) &= \frac{d}{d\hat{\theta}} \int_{\text{-}\infty}^{\infty} |\,\theta - \hat{\theta}(\mathrm{x})\,|\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\theta \\
&= \int_{-\infty}^{\hat{\theta}(\theta)_{abs}} \theta\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta - \int_{\hat{\theta}(\theta)_{abs}}^{\infty} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta = 0
\end{aligned}$$

$$\Rightarrow \hat{\theta}(\mathrm{x})_{abs} : \int_{-\infty}^{\hat{\theta}(\theta)_{abs}} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta = \int_{\hat{\theta}(\theta)_{abs}}^{\infty} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta \tag{2.8}$$

The estimate defined in equation (2.7) is called the median of a posterior density. This method of risk optimization is often used to find more robust estimates in asymmetric distributions [14].

- The uniform error cost function :

  The risk of the uniform function can be formulated as

$$R_{unf} = \int_{\text{-}\infty}^{\infty} \mathrm{p}_{\mathrm{x}}(\mathrm{x})\, dx[1 - \int_{\hat{\theta}(\theta)_{unf} - \frac{\Delta}{2}}^{\hat{\theta}(\theta)_{unf} + \frac{\Delta}{2}} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\theta] \tag{2.9}$$

  To minimize risk in this case, we clearly want to maximize the internal integral in the local region of the estimate (i.e. having a small value of Δ).

$$\arg\max_{\theta} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x}) \tag{2.10}$$

  This methodology of minimizing risk is called a maximization of a posteriori (MAP). This particular method is the most well-known method to find an estimate in the Bayesian framework.

2.3 Non-Bayesian Risk Analysis

Obviously, the joint assumption in the above risk analysis does not apply for the deterministic case, since there is no prior density. Rather, risk defined in equation

(2.5) is adopted.

$$R_{ml} = \int_{\text{-}\infty}^{\infty} [\theta - \hat{\theta}(\mathrm{x})]^2 \, \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta) \, dx \tag{2.11}$$

Since the expectation is only over the observations x, a minimization of the risk hopefully will lead us to

$$\mathbb{E}\{\hat{\theta}(\mathrm{x})_{ml}\} = \theta \tag{2.12}$$

However, this is a very optimistic assumption, and it does not open a possibility for further analysis. Instead a more robust definition of expectations of the maximum likelihood expectation (MLE) estimate is defined as follows

$$\mathbb{E}\{\hat{\theta}(\mathrm{x})_{ml}\} \triangleq \int_{\text{-}\infty}^{\infty} \hat{\theta}(\mathrm{x}) \, \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta) \, dX \tag{2.13}$$

The mean square criteria of a mean square error using the a general estimate can be decomposed into varaince and bias as follows

$$\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[(\hat{\theta} - \theta)^2] = \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \tag{2.14}$$

$$= \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)} \, [(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta}] + \theta)^2 \tag{2.15}$$

$$= \text{Variance} + \text{Bias}^2 \tag{2.16}$$

Following the above decomposition of the mean squared error. three situations can arise depending on values in equation (2.15):

- If $\mathbb{E}[\hat{\theta}(\mathrm{x})] = \theta$, this is called unbiased estimator.
- If $\mathbb{E}[\hat{\theta}(\mathrm{x})] = \theta + \mathrm{B}$, where B is not a function of θ, which means that this estimator has a known bias.
- If $\mathbb{E}[\hat{\theta}(\mathrm{x})] = \theta + \mathrm{B}(\theta) = \mu_\theta$, the estimator has an unknown bias that is a function of the parameter.

Figure 2.2 shows a contrast between the visual inference of the bias, variance relations.

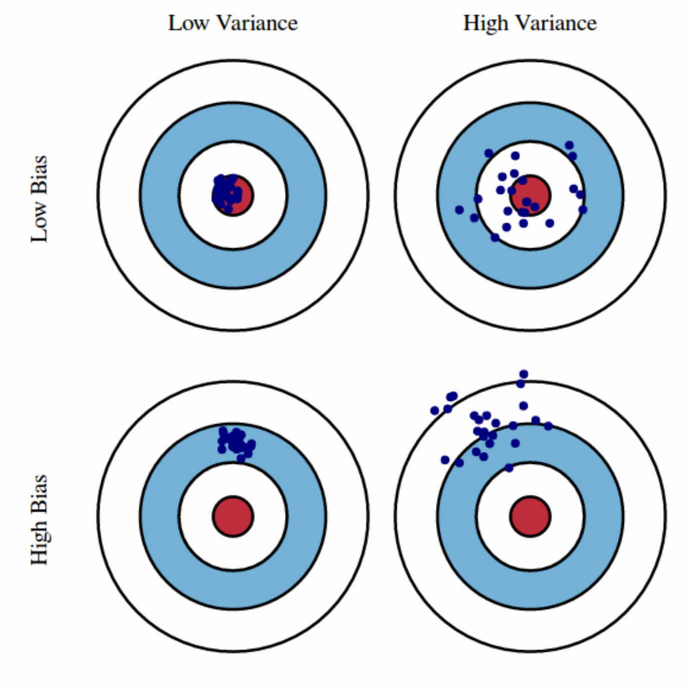

Figure 2.2: Bias Variance Visualization

2.4 Parameter Bounds

Statistical bounds are another way of measuring the quality of an estimate. Bounds derived from the covariance inequality and Ziv-Zakai family bounds are well-known examples of parameter bounds [14]. A well-known bound on the covariance inequality is the Cramér-Rao Bound (CRB).

In this report, a Bayesian framework is applied on two models, one is assumed to be the correct model that data are distributed upon. Hence the name, Correctly specified model. The other model follows another family of distributions, and assumed to be a misspesified model. The remaining part of this chapter is meant to derive CRB type bounds. The deterministic correctly specified bound has been given the name CRB, the Bayesian correctly specified bound and has been given the name

BCRB, the deterministic misspecified bound and has been given the name MCRB. Finally, the Bayesian misspecified bound and has been given the acronym MBCRB. Before digging deep into the CRB bound, certain assumptions are needed.

2.4.1 Inner Product and Expectated Value

Assume we have two random variables $\zeta(\mathrm{x})$ and $\eta(\mathrm{x}, \theta)$ that are functions of a random variable X that is jointly distributed with θ according to joint cumulative distribution function $F(\mathrm{x}, \theta)$. The inner product of these two random variables can be defined as the expectation of their product, sometimes called their correlation:

$$< \zeta(\mathrm{x}), \eta(\mathrm{x}, \theta) > \triangleq \mathbb{E}[\zeta(\mathrm{x}), \eta(\mathrm{x}, \theta)] = \int_{\text{-}\infty}^{\infty} \zeta(\mathrm{x}) \eta(\mathrm{x}, \theta) dF(\mathrm{x}, \theta)$$

where $dF(\mathrm{x}, \theta) = \mathrm{p}_{\mathrm{x}, \theta}(\mathrm{x}, \theta)\, dxd\,\theta$

2.4.2 Inner product and Cosine Between Two Vectors

$$< \zeta, \zeta > = ||\zeta||^2, \qquad \cos(\theta) = \frac{< \zeta, \eta >}{||\zeta||\ ||\eta||}$$

From the previous two facts, and taking into account the two that $||.||^2$ is a positive quantity and $\cos^2(\theta) \leq 1$, the covariance inequality(i.e. Cauchy-Schwarz inequality) can be defined as follows:

$$\mathbb{E}[\zeta^2] \geq \frac{\mathbb{E}^2[\zeta\eta]}{\mathbb{E}[\eta^2]}$$

Worth noting from the covariance inequality that the cosine between any two elements of the inner product space equals unity if and only if $\zeta \propto \eta$. In other words, equality of the covariance inequality can be reached if the two random variable are parallel(i.e. proportional) to each other.

It is important to say that we are going to use the covariance inequality to establish a

Cramér-Rao type bound. It can be shown that our choice of score function, along with some regularity conditions, facilitates this whether the model is correctly specified or misspecified [15].

2.5 Cramér-Rao Bound For Correctly Specified Models

Cramér-Rao bound is derived from the covariance inequality and gives us the minimum variance bound that can be achieved for the risk function. In order to establish this bound, proper values need to be assigned for ζ and η. Let $\zeta(\hat{\theta}(\mathrm{x}), \theta)$ and $\eta(\mathrm{x}, \theta)$. where ζ is related to the variance we want to find its lower bound, and η is the customized score function chosen to find the tightest bound possible. Before proceeding in the derivation of the Cramér-Rao Bound there are some assumptions must be taken into account for CRB and BCRB. Those are mentioned in appendix (A.1-2). Also, the joint probability density $\mathrm{p}_{\mathrm{x},\theta}(\mathrm{x}, \theta) = \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\, \pi_\theta(\theta)$ will be naturally used for the Bayesian case, while the conditional density is used in the derivation of the deterministic case.

2.5.1 CRB

The parameter value is assumed to be deterministic but unknown. Furthermore, the estimator is only a function of the observed data $\hat{\theta} = \hat{\theta}(\mathrm{x})$ and any observed data point is taken from the likelihood density given the true value of the parameter: $\mathrm{x} \sim \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$. Also, it is convenient to define a value of the estimator mean to be

$$\mu_\theta \triangleq \mathbb{E}_{\mathrm{x}|\,\theta}[\hat{\theta}]$$

Now choose the random variables as follows:

$$\zeta = \hat{\theta} - \mu_\theta, \qquad\qquad \eta = \frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\, \theta} \; .$$

Derivation of the bound will be as follows

$$\begin{aligned}\mathbb{E}_{\mathrm{x}|\,\theta}[(\hat{\theta}-\mu_\theta)^2] &\geq \frac{\mathbb{E}^2_{\mathrm{x}|\,\theta}[(\hat{\theta}-\mu_\theta)(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})]}{\mathbb{E}_{\mathrm{x}|\,\theta}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^2]}\\ &= \frac{\mathbb{E}^2_{\mathrm{x}|\,\theta}[\hat{\theta}\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta} - \mu_\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}]}{\mathbb{E}_{\mathrm{x}|\,\theta}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^2]}\ .\end{aligned}$$

Focusing on the numerator, it is desired to get rid of its dependence on the estimator, since it is the estimate we want to lower bound:

$$\begin{aligned}&\Rightarrow \mathbb{E}_{\mathrm{x}|\,\theta}[\hat{\theta}\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta} - \mu_\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}]\\ &= \int_{-\infty}^{\infty} \hat{\theta}(\mathrm{x})\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}\,\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\,dx - \int_{-\infty}^{\infty} \mu_\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}\,\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\,dx\\ &= \int_{-\infty}^{\infty} \hat{\theta}(\mathrm{x})\,\frac{\partial\,\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}\,\frac{\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}dx - \mu_\theta \int_{-\infty}^{\infty} \frac{\partial\,\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}\,\frac{\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}dx\\ &= \frac{\partial}{\partial\,\theta}\int_{-\infty}^{\infty} \hat{\theta}(\mathrm{x})\,\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\,dx - \mu_\theta\,\frac{\partial}{\partial\,\theta}\int_{-\infty}^{\infty} \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\,dx\\ &= \frac{\partial}{\partial\,\theta}(\mu_\theta) - \mu_\theta\,\frac{\partial}{\partial\,\theta}(1)\\ &= \frac{\partial \mu_\theta}{\partial\,\theta}\ .\end{aligned}$$

Hence the CRB bound can be written as follows:

$$\Rightarrow \mathbb{E}_{\mathrm{x}|\,\theta}[(\hat{\theta}-\mu_\theta)^2] \geq \left.\frac{(\frac{\partial \mu_\theta}{\partial\,\theta})^2}{\mathbb{E}_{\mathrm{x}|\,\theta}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^2]}\right|_{\theta=\theta_T} = CRB(\theta_T) \qquad (2.17)$$

Observe that for an unbiased estimator ($\mu_\theta = \theta$), the bound is given by the inverse Fisher information content and the variance is the MSE.

2.5.2 BCRB

The Bayesian case is slightly different from the previous one in terms of the prior assumption. Now, there is a prior density $\pi_\theta(\theta)$, and although the estimator is still a function of the observed data $\hat{\theta} = \hat{\theta}(\mathrm{x})$, the data is distributed according to the joint probability density $\mathrm{x} \sim \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)$. The values of random variables and the covariance inequality will slightly change. We choose the random variables as follows:

$$\zeta = \hat{\theta} - \theta, \qquad\qquad \eta = \frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}\ .$$

Then the inequality will look like

$$\mathbb{E}_{\mathrm{x},\theta}[(\hat{\theta}-\theta)^{2}] \geq \frac{\mathbb{E}^{2}_{\mathrm{x},\theta}[(\hat{\theta}-\theta)(\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta})]}{\mathbb{E}_{\mathrm{x},\theta}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta})^{2}]}$$

Similarly the derivation will try to get rid of the estimator from the left term of the bound.

$$\begin{aligned}\mathbb{E}_{\mathrm{x},\theta}[(\hat{\theta}-\theta)^{2}] &\geq \frac{\mathbb{E}^{2}_{\mathrm{x},\theta}[(\hat{\theta}-\theta)(\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta})]}{\mathbb{E}_{\mathrm{x},\theta}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta})^{2}]}\\ &= \frac{\mathbb{E}^{2}_{\mathrm{x},\theta}[\hat{\theta}\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta} - \theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}]}{\mathbb{E}_{\mathrm{x},\theta}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta})^{2}]}\end{aligned}$$

As what has been done to CRB we use the assumptions provided in appendix (A.1) to manipulate the numerator:

$$\begin{aligned}&\Rightarrow \mathbb{E}_{\mathrm{x},\theta}[\hat{\theta}\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta} - \theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}]\\ &= \mathbb{E}_{\mathrm{x},\theta}[\hat{\theta}\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}] - \mathbb{E}_{\mathrm{x},\theta}[\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}]\\ &= \mathbb{E}_{X}[\hat{\theta}(\mathrm{x})[\mathbb{E}_{\theta\,|X}[\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}]]] - \mathbb{E}_{\mathrm{x},\theta}[\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}]\\ &= 0 - \mathbb{E}_{\mathrm{x},\theta}[\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}]\end{aligned}$$

The first element is zero because of the fact that the expected value of the score function for the joint probability density goes to zero (4-A.1)

$$\begin{aligned}\Rightarrow -\mathbb{E}_{\mathrm{x},\theta}\left[\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}\right] &= -\int_{\text{-}\infty}^{\infty}\int_{\text{-}\infty}^{\infty}\theta\,\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}\,\mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)\,dxd\,\theta\\ &= -\int_{\text{-}\infty}^{\infty}\mathrm{p}_{\mathrm{x}}(\mathrm{x})\,dx\int_{\text{-}\infty}^{\infty}\theta[\frac{\partial \ln \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}{\partial\,\theta} + \frac{\partial \ln \mathrm{p}_{\mathrm{x}}(\mathrm{x})}{\partial\,\theta}]\,\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\,d\,\theta\\ &= -\int_{\text{-}\infty}^{\infty}\mathrm{p}_{\mathrm{x}}(\mathrm{x})\,dx\int_{\text{-}\infty}^{\infty}\theta\,\frac{\partial\,\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}{\partial\,\theta}\,\frac{\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}{\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}d\,\theta\\ &= -\int_{\text{-}\infty}^{\infty}\mathrm{p}_{\mathrm{x}}(\mathrm{x})\,dx\int_{\text{-}\infty}^{\infty}\theta\,\frac{\partial\,\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}{\partial\,\theta}\,d\,\theta\end{aligned}$$

The inner integral can be evaluated using the integration-by-parts method:

$$\begin{aligned}
&= -\int_{-\infty}^{\infty} \mathrm{p_x(x)}\, dx \left[\theta\, \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x}) - \int_{-\infty}^{\infty} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta \right] \\
&= -\int_{-\infty}^{\infty} \mathrm{p_x(x)}\, dx \left[0 - \int_{-\infty}^{\infty} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta \right] \\
&= +\int_{-\infty}^{\infty} \mathrm{p_x(x)}\, dx \int_{-\infty}^{\infty} \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\, d\,\theta = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)\, dxd\,\theta = 1
\end{aligned}$$

Hence the BCRB bound can be written as follows:

$$\mathbb{E}_{\mathrm{x},\theta}[(\hat{\theta}-\theta)^2] \geq \left. \frac{1}{\mathbb{E}_{\mathrm{x},\theta}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta})^2]} \right|_{\theta=\theta_T} = BCRB(\theta_T) \qquad (2.18)$$

2.6 Cramér-Rao Bound For Misspecified Models

Parameter bounds assumed usually a perfect match of the true distribution of the data. However, this is mostly wrong due to many reasons. In this chapter, we try to find a theoretical bound for a misspecified model for deterministic and random true parameter. Hence, there are two distributions that must be taken into consideration for this part: $\mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)$ as the true data distribution and $\mathrm{q}_{\mathrm{x},\theta}(\mathrm{x},\theta)$ as the assumed distribution.

$$\mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta) \neq \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta) \quad \text{is allowed} \quad (MCRB)$$

$$\mathrm{q}_{\mathrm{x},\theta}(\mathrm{x},\theta) \neq \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta) \quad \text{is allowed} \quad (MBCRB)$$

Note that, the true distribution $\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$ is possibly independent of θ and a function of the data only (i.e. $\mathrm{p} = \mathrm{p_x(x)}$), because the real purpose for this approach is to allow for the possibility that the model parametrization is incorrect. Also, if the real distribution actually depends on the parameter θ, then the value θ takes on will be fixed for our measurement and outside our control. Hence assuming the real distribution (p) is only a function of observations (x) is sufficient for the purpose of studying misspecification.

2.6.1 MCRB

Similar to the case of CRB, θ is a deterministic value, and the estimate is a function of the data (i.e. $\hat{\theta} = \hat{\theta}(\mathrm{x})$). However, in the misspecified case, $\mathrm{x} \sim \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta) \neq \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$ is allowed. In addition, a new definition of the mean must be introduced:

$$\mu_{\mathrm{p}} \triangleq \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})] = \int_{-\infty}^{\infty} \hat{\theta}_{\mathrm{q}}(\mathrm{x})\, \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\, dx$$

where $\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[.]$ means that this expectation is with respect to the true distribution $\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$, and $\hat{\theta}_{\mathrm{q}}(\mathrm{x})$ has subscript "q" to indicate that this estimator is based on the assumed distribution $\mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$. The choice of random variables will be:

$$\zeta = \hat{\theta}_{\mathrm{q}}(\mathrm{x}) - \mu_{\mathrm{p}}, \qquad \eta = \frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} - \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}]$$

Then the inequality will look like

$$\mathbb{E}_{\mathrm{p}(\mathrm{x},\theta)}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x}) - \mu_{\mathrm{p}})^2] \geq \frac{\mathbb{E}^2_{\mathrm{p}(\mathrm{x},\theta)}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x}) - \mu_{\mathrm{p}})(\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} - \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}])]}{\mathbb{E}_{\mathrm{p}(\mathrm{x},\theta)}[(\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} - \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}])^2]} \tag{2.19}$$

using the score function that involves that KL divergence in appendix (A.1-2), the above inequality can be written as

$$\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x}) - \mu_{\mathrm{p}})^2] \geq \frac{\mathbb{E}^2_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x}) - \mu_{\mathrm{p}})(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} + \frac{\partial \mathrm{D}}{\partial \theta})]}{\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} + \frac{\partial \mathrm{D}}{\partial \theta})^2]} \tag{2.20}$$

Numerator can be simplified to be:

$$\begin{aligned}
&\left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x}) - \mu_{\mathrm{p}})(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} + \frac{\partial \mathrm{D}}{\partial \theta})]\right]^2 \\
&= \left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}] + \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\frac{\partial \mathrm{D}}{\partial \theta}] - \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\mu_{\mathrm{p}}\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}] - \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\mu_{\mathrm{p}}\frac{\partial \mathrm{D}}{\partial \theta}]\right]^2 \\
&= \left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}] + \frac{\partial \mathrm{D}}{\partial \theta}\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})] - \mu_{\mathrm{p}}\,\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}] - \mu_{\mathrm{p}}\frac{\partial \mathrm{D}}{\partial \theta}\right]^2 \\
&= \left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}] + \mu_{\mathrm{p}}\frac{\partial \mathrm{D}}{\partial \theta}\right]^2
\end{aligned}$$

Denominator can be simplified to be:

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}\left[\left(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}+\frac{\partial \mathrm{D}}{\partial\,\theta}\right)^{2}\right]\\
&=\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^{2}]+\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[2(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})(\frac{\partial \mathrm{D}}{\partial\,\theta})]+\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \mathrm{D}}{\partial\,\theta})^{2}]\\
&=\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^{2}]+2\,\frac{\partial \mathrm{D}}{\partial\,\theta}\,\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})]+\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \mathrm{D}}{\partial\,\theta})^{2}]\\
&=\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^{2}]+2\,\frac{\partial \mathrm{D}}{\partial\,\theta}(-\frac{\partial \mathrm{D}}{\partial\,\theta})+(\frac{\partial \mathrm{D}}{\partial\,\theta})^{2}\\
&=\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^{2}]-(\frac{\partial \mathrm{D}}{\partial\,\theta})^{2}
\end{aligned}
$$

Hence, the MCRB bound after simplification will be:

$$
\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x})-\mu_{\mathrm{p}})^{2}]\geq\left.\frac{\left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}]+\mu_{\mathrm{p}}(\frac{\partial \mathrm{D}}{\partial\,\theta})\right]^{2}}{\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}\left[(\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta})^{2}\right]-\left[(\frac{\partial \mathrm{D}}{\partial\,\theta})^{2}\right]}\right|\triangleq MCRB(\theta,q:p)
\tag{2.21}
$$

Note that if the model is correctly specified (i.e. $q=\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$), then:

$$
\begin{aligned}
&\mu_{q}=\mu_{\theta},\ \frac{\partial \mathrm{D}}{\partial\,\theta}=0,\ \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\,\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}]=\frac{\partial \mu_{\theta}}{\partial\,\theta}\\
&\quad\to MCRB(\theta,q:p)=CRB(\theta)
\end{aligned}
$$

2.6.2 MBCRB

Similar to BCRB, it is more advantageous to work with the MSE instead of the variance (i.e. pursuing the value of the estimator θ instead of its mean μ_{p}, and introducing the new score function for misspecified models. Thus, random variables will be as follows:

$$
\zeta=\hat{\theta}_{\mathrm{q}}(\mathrm{x})-\theta,\qquad\qquad \eta=\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}-\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}]\ .
$$

Then, the inequality becomes

$$\mathbb{E}_{\mathrm{p}(\mathrm{x},\theta)}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x})-\theta)^{2}] \geq \frac{\left[\mathbb{E}_{\mathrm{p}(\theta)}\left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x})-\theta)(\frac{\partial\ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta}-\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\frac{\partial\ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta}])]\right]\right]^{2}}{\mathbb{E}_{\mathrm{p}(\theta)}\left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[(\frac{\partial\ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta}-\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\frac{\partial\ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta}])^{2}]\right]}$$

Note that the inner expectation is identical to the MCRB case, which means we can evaluate the conditional bound then we average out the outer expectation. Hence the MBCRB bound can be written as follows:

Note that inequalities (10 and 11) involve terms that could be function of the estimator in their right hand side :

$$\begin{array}{lr}\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta_{\mathrm{T}})}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\,\dfrac{\partial\ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta}] & MCRB\\ \mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\,\dfrac{\partial\ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta}] & MBCRB\end{array}$$

Which means by definition that they are not a Cramér-Rao Type inequalities. The given choice of the appropriate regularity conditions, however, will result in expressions independent of the estimator, and therefore an inequality of the Cramér-Rao type[16].

$$\mathbb{E}_{\mathrm{p}(\mathrm{x},\theta)}[(\hat{\theta}_{\mathrm{q}}(\mathrm{x})-\theta)^{2}] \geq \left.\frac{\left[\mathbb{E}_{\mathrm{p}(\theta)}\left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}[\hat{\theta}_{\mathrm{q}}(\mathrm{x})\,\frac{\partial\ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta}]+\mu_{\mathrm{p}}(\frac{\partial \mathrm{D}}{\partial\theta})\right]\right]^{2}}{\mathbb{E}_{\mathrm{p}(\theta)}\left[\mathbb{E}_{\mathrm{p}(\mathrm{x}|\,\theta)}\left[(\frac{\partial\ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\theta})^{2}\right]-(\frac{\partial \mathrm{D}}{\partial\theta})^{2}\right]}\right| \triangleq MBCRB(\theta,p:q) \tag{2.22}$$

However, derivations of bounds in this report in vector format, not in scalar one as shown above. Table 2.1 shows all important terms and a the lower bound on MSE for all four cases.

Table 2.1: The CRB Bounds.

Cramer Rao Bound				
Type	Est error (ζ)	Score Function (η)	**FIM**	CRLB Type Bound
CRB	$\hat{\theta}_{\mathbf{p}}(\mathbf{x}) - \theta_{\mathbf{t}}$	$\frac{\partial \ln \mathrm{p}(\mathrm{x} \mid \theta)}{\partial \theta^*}$	$\mathbb{E}_p[(\frac{\partial \ln \mathrm{p}(\mathrm{x} \mid \theta)}{\partial \theta^*})^2]$	$\frac{\partial \mu_{\theta_{\mathrm{t}}}}{\partial \theta_{\mathrm{t}}} (\mathbf{FIM})^{-1} \frac{\partial \mu_{\theta_{\mathrm{t}}}}{\partial \theta_{\mathrm{t}}}^{\mathbf{H}}$
MCRB	$\hat{\theta}_{\mathbf{q}}(\mathbf{x}) - \mu_{\mathrm{p}}(\theta)$	$\frac{\partial \ln \mathrm{q}(\mathrm{x} \mid \theta)}{\partial \theta^*}$- $\mathbb{E}_{\mathrm{p}_{\mathrm{x} \mid \theta}} \left\{ \frac{\partial \ln \mathrm{q}(\mathrm{x} \mid \theta)}{\partial \theta^*} \right\}$	$\mathrm{J}_{(\mathrm{p:q})}(\theta)$	$\Xi_{(\mathrm{p:q})}(\theta)^{\mathbf{H}} \, \mathrm{J}_{(\mathrm{p:q})}(\theta)^{\mathbf{-1}} \, \Xi_{(\mathrm{p:q})}(\theta)$
BCRB	$\hat{\theta}_{\mathbf{p}}(\mathbf{x}) - \theta$	$\frac{\partial \ln \mathrm{p}(\mathrm{x} \mid \theta)}{\partial \theta^*}$	$\mathbb{E}_p[(\frac{\partial \ln \mathrm{p}(\mathrm{x} \mid \theta)}{\partial \theta^*})^2]$	$\mathbf{FIM}^{-1}$
MBCRB	$\hat{\theta}_{\mathbf{q}}(\mathbf{x}) - \mu_{\mathrm{p}}(\theta)$	$\frac{\partial \ln \mathrm{q}(\mathrm{x} \mid \theta)}{\partial \theta^*}$- $\mathbb{E}_{\mathrm{p}_{\mathrm{x} \mid \theta}} \left\{ \frac{\partial \ln \mathrm{q}(\mathrm{x} \mid \theta)}{\partial \theta^*} \right\}$	$\mathrm{J}_{(\mathrm{p:q})}(\theta)$	$\mathbb{E}_{p_\theta}[\Xi_{(\mathrm{p:q})}(\theta)^H] E_{p_\theta}^{-1}[\mathrm{J}_{(\mathrm{p:q})}(\theta)] \, \mathbb{E}_{\mathrm{p}_\theta}[\Xi_{(\mathrm{p:q})}(\theta)]$

Chapter 3

PROBLEM FORMULATION

3.1 Linear System Revisited

Let's investigate the linear system in equation (1.1)

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

$$\begin{bmatrix} \vdots \\ \mathbf{y} \\ \vdots \end{bmatrix} = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & \mathbf{A} & . & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix} \begin{bmatrix} . \\ \mathbf{x} \\ . \end{bmatrix}$$

$\mathbf{y} \in \mathbb{C}^{M\times 1}$ a vector of observed responses

$\mathbf{A} \in \mathbb{C}^{M\times N}$ a matrix of deterministic and known measurements

$\mathbf{x} \in \mathbb{C}^{N\times 1}$ the target sparse vector

The solution of this linear system can have one of the following situations depending on the rank of **A**. Assuming a linearly independent **A** we have the three following situations:

- **Overdetermined linear system** ($M > N$):

$$\begin{bmatrix} \vdots \\ . \\ \mathbf{y} \\ . \\ \vdots \end{bmatrix} = \begin{bmatrix} a_{11} & . & a_{1n} \\ . & . & . \\ . & \mathbf{A} & . \\ . & . & . \\ a_{m1} & . & a_{mn} \end{bmatrix} \begin{bmatrix} . \\ \mathbf{x} \\ . \end{bmatrix}$$

This system has a thin and tall **A** (i.e. more equations than unknown). In this situation, the system has no exact solution.

- **Perfectly determined linear system** ($M = N$):

$$\begin{bmatrix} \vdots \\ \mathbf{y} \\ \vdots \end{bmatrix} = \begin{bmatrix} a_{11} & . & a_{1n} \\ . & \mathbf{A} & . \\ a_{m1} & . & a_{mn} \end{bmatrix} \begin{bmatrix} . \\ \mathbf{x} \\ . \end{bmatrix}$$

This system has a square **A**. In this situation, the system has a unique solution. The solution can be found by multiplying both side by the inverse of the **A**.

- **Underdetermined linear system** ($M < N$):

This system has a short and fat **A**(i.e. fewer equations than unknowns). In this situation, the system has infinitely many solutions. Classically, an approximate solution is found by using least square criterion.

$$\begin{bmatrix} \vdots \\ \mathbf{y} \\ \vdots \end{bmatrix} = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & \mathbf{A} & . & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix} \begin{bmatrix} \vdots \\ . \\ \mathbf{x} \\ . \\ \vdots \end{bmatrix}$$

3.1.1 Sparse Vector Recovery For Noisy Undetermined System

The problem of interest is formalized as follows

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w} \tag{3.1}$$

where $\mathbf{A} \in \mathbb{C}^{M \times N}$ is a matrix of deterministic and known measurements with ($M < N$), and $\mathbf{x} \in \mathbb{C}^{N \times 1}$ is vector representation of sparse signals, $\mathbf{w} \in \mathbb{C}^{M \times 1}$ is noise modeled following a circularly symmetric complex Gaussian distribution (i.e. $\mathbf{w} \sim \mathcal{CN}(\mathbf{w}; \mathbf{0}, \mathbf{I}_N \sigma_{w_i}^2))$, where CN denotes a complex Gaussian distribution and $\mathbf{y} \in \mathbb{C}^{M\times 1}$ is a vector representation of observed responses.

$$\begin{bmatrix} \vdots \\ \mathbf{y} \\ \vdots \end{bmatrix} = \begin{bmatrix} a_{11} & . & . & . & a_{1n} \\ . & . & \mathbf{A} & . & . \\ a_{m1} & . & . & . & a_{mn} \end{bmatrix} \begin{bmatrix} \vdots \\ . \\ \mathbf{x} \\ . \\ \vdots \end{bmatrix} + \begin{bmatrix} \vdots \\ \mathbf{w} \\ \vdots \end{bmatrix}$$

in this underdetermined assumption $\mathbf{A}$ is short and fat and $\mathbf{x}$ has independent and identically distributed (iid) realizations with a known sparsity L. Bayesian framework is concerned about modeling sparsity in a way that is tractable. The goal here is to estimate a sparse vector $\mathbf{x}$ based on an observation vector $\mathbf{y}$. Let's assume a general model of the prior distribution denoted by density $p(.)$ if correctly specified, and $q(.)$ if not. Figure 3.1 shows the interaction between the aforementioned assumptions of distributions.

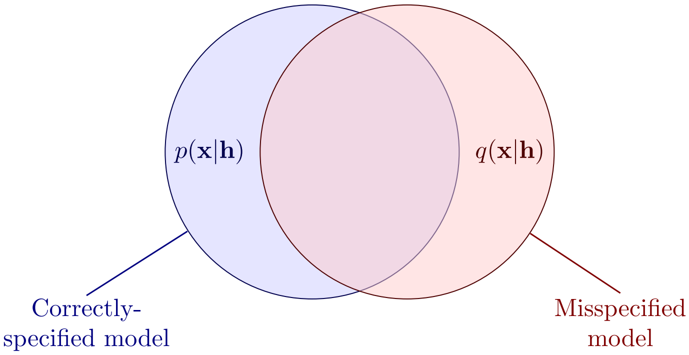

Figure 3.1: Assumed Distributions To Model Sparsity Promoting Prior Distribution

Generally, posterior distribution for both assumptions can be optimized following

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\mathbf{h})p(\mathbf{h}) \tag{3.2}$$

$$q(\mathbf{x}|\mathbf{y}) \propto q(\mathbf{y}|\mathbf{x})q(\mathbf{x}|\mathbf{h})q(\mathbf{h}) \tag{3.3}$$

where $\mathbf{h}$ represents a vector set of arbitrary hyperparameters (i.e. prior parameters of the prior distribution).

3.2 Prior Distributions

To enforce sparsity in the Bayesian framework, proper prior distributions must be assumed. A Bernoulli Gaussian (BG) model is assumed to be the correctly specified distribution and Gamma Normal (GN) model is assumed to be a misspecified model.

3.2.1 Bernoulli Gaussian Model

The Probability Distribution Function (PDF) of a BG model is given by:

$$p(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{N}(1-\lambda_i)\delta(x_i) + (\lambda_i)\mathcal{CN}(x_i; \theta_i, \Phi_i) \tag{3.4}$$

where $\mathbf{h} \triangleq [\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\Phi}]$. In our approach we adopt standardized Complex Gaussian assumption which makes the mean vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_N]^T = \mathbf{0}$, the covariance matrix $\boldsymbol{\Phi} = \mathbf{I}_N$, and $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$, the sparsity-promoting random variable vector following Bernoulli distributed trials, be

$$\boldsymbol{\lambda} \sim \prod_{i=1}^{N} \mathcal{B}(1, \lambda_i) \tag{3.5}$$

where $\lambda_i \triangleq \frac{||\mathbf{x}||_0}{N} = \frac{L}{N}$ a deterministic sparsity ratio of the number of non-zero entries to the full size of $\mathbf{x}$ and $\mathcal{B}$ denotes a Bernoulli density. Figure 3.2 shows generated BG distributions with different values of sparsity $\boldsymbol{\lambda}$

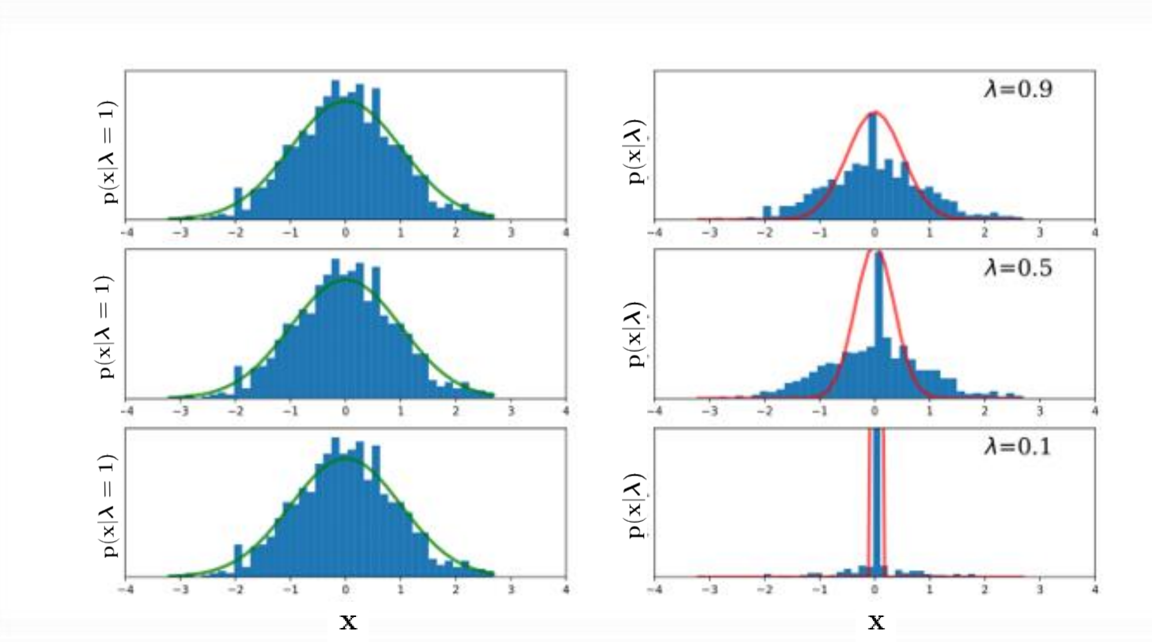

Figure 3.2: Identical Gaussian Distribution With Corresponding Sparsity Values.

3.2.2 Gamma Normal Model

The PDF of a Gamma-Normal model is given by:

$$q(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{N} \mathcal{CN}(x_i; \theta_i, \alpha_i^{-1}) \tag{3.6}$$

where $\mathbf{h} \triangleq [\boldsymbol{\theta}, \boldsymbol{\alpha}]$ as in (3.4). We assume zero-mean for all entries, i.e., $\boldsymbol{\theta} = \mathbf{0}$. We also assume that α_i is an independent identically distributed (i.i.d) random variable that follows Gamma distribution with location and shape parameters, a and b given by:

$$\boldsymbol{\alpha} \sim \prod_{i}^{N} \text{Gamma}(\alpha_i; a_i, b_i) = \prod_{i}^{N} \frac{b_i^{a_i}}{\Gamma(a_i)} \alpha_i^{a_i - 1} e^{-b_i \alpha_i} \ . \tag{3.7}$$

One of the reasons GN is chosen as a sparsity promoting model is its tractability, meaning that given hyperparameters a,b of a prior Gamma, $q(x_i)$ can be found to be

$$q(x_i) = \int_0^\infty q(x_i|\alpha_i) q(\alpha_i) d\alpha_i = \frac{b^a}{\pi \Gamma(a)} \int_0^\infty \alpha_i^{a-1+1} e^{-\alpha_i(|x_i|^2 + b)} d\alpha_i$$

$$= \frac{b^a \Gamma(a+1)}{\pi \Gamma(a)(b+|x_i|^2)^{a+1}} \ . \tag{3.8}$$

Setting $a = 1$ in equation (3.8) will result a closed form density of $q(x_i)$ to be

$$q(x_i) = \frac{b\Gamma(2)}{\pi\Gamma(1)(b+|x_i|^2)^2} = \frac{b}{\pi(b+|x_i|^2)^2} \tag{3.9}$$

On this Bayesian framework, the setting of hyperparameters is essential to achieve multiple objectives. First, hyperparameters a and b must be chosen to favor sparsity. Figure 3.3 shows how the pdf of the Gamma model with different values of hyperparameters changes. Second, as shall be seen later on, the derived bound for this model will still be a function of b. Meaning that there is a trade-off between sparsity constrain the bound behaviour. A complete analysis and discussion will follow in the next chapter.

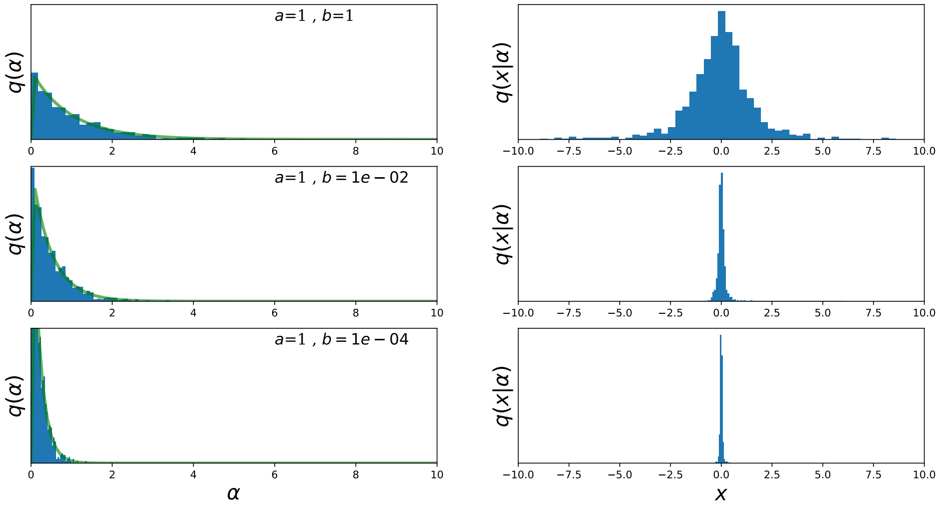

Figure 3.3: Gamma Normal Distribution With Various Values Of Hyperparameters

It can be inferred from Figure 3.3 that X has an assumed mean of zero and an assumed prior on the variance. Since a is location parameter, it is chosen to be 1 or some value around it as in equation (3.7). The shape parameter b determines steepness of the tail. Having a small value of b lower encourages the majority realizations α to be around zero, and vice versa. The desired objective here is to have smooth sparsity with lower losses possible on the bounds of the estimate with fast and guaranteed convergence of the algorithm used.

3.2.3 Bayesian Cramér-Rao Bound

The development of bounds for both models comes from [17]. Since BG is assumed to be the true generative model, general BCRB bound ideally is the tightest bound possible. It is important to note that all distributions in this work are in the complex domain, i.e. the distributions are real valued functions of complex variables. Therefore, complex gradient methods based on Wirtinger calculus will be used. Specifically, complex-valued $\mathbf{x}$ is treated as a function of $\mathbf{x}$ and its conjugate $\mathbf{x}^*$ then derivative is taken with respect to $\mathbf{x}^*$ as in [18]. Hence, a complex-valued vector $\mathbf{x}$ yields parameters as follows,

$$\boldsymbol{\theta} = \begin{bmatrix} \mathbf{x}^T & \mathbf{x}^H \end{bmatrix}^T \tag{3.10}$$

$\mathbf{x}$ and $\boldsymbol{\theta}$ are used interchangeably while extracting bounds since they don not change in the corresponding probability distributions. Worth mentioning that $\mathbf{x}$ and $\boldsymbol{\theta}$ represents the sparse vector and not to be confused with the ones in equations(3.4, 3.6).

The non-Bayesian CRB can be written as

$$\mathbb{E}_{p_{\mathbf{y}|\boldsymbol{\theta}}}\{[\hat{\boldsymbol{\theta}}_p(\mathbf{y}) - \boldsymbol{\theta}][\hat{\boldsymbol{\theta}}_p(\mathbf{y}) - \boldsymbol{\theta}]^H\} \geq \mathbf{J}_{(p)}^{-1}(\boldsymbol{\theta}) \tag{3.11}$$

where

$$\mathbf{J}_{(p)}(\boldsymbol{\theta}) \triangleq -\mathbb{E}_{p_{\mathbf{y}|\boldsymbol{\theta}}}[\frac{\partial^2 \ln \mathrm{p}(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{\mathrm{T}}}] \tag{3.12}$$

Then, we obtain the likelihood from (3.1) as:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{CN}(\mathbf{A}\mathbf{x}, \sigma_w^2 \mathbf{I}_M), \tag{3.13}$$

which gives the following expected value of the second derivative of the log likelihood,

$$\mathbb{E}_{p_{\mathbf{y}|\boldsymbol{\theta}}}\left\{\Big(\frac{\partial^2 \ln \mathrm{p}(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^*\partial\boldsymbol{\theta}^{\mathrm{T}}}\Big)\right\} = -\frac{1}{\sigma_w^2}\begin{bmatrix}\mathbf{B} & \mathbf{0}_{N\times N}\\ \mathbf{0}_{N\times N} & \mathbf{B}^T\end{bmatrix} \tag{3.14}$$

where $\mathbf{B} = \mathbf{A}^H\mathbf{A}$. Hence, BCRB is given by:

$$\boxed{\mathbb{E}_{p_{\boldsymbol{\theta}}}\left\{||\mathbf{x}-\hat{\mathbf{x}}||_2^2\right\} \geq \frac{\sigma_w^2}{\mathbf{tr}\{\mathbf{B}\}}} \tag{3.15}$$

$\mathbf{tr}\{\cdot\}$ refers to the trace of a matrix.

3.2.4 Misspecified Bayesian Cramér-Rao Bound For GN Model

Obtaining MBCRB for an arbitrary class of distributions requires defining a set of parameters. The maximum a posteriori (MAP) estimate $\hat{\boldsymbol{\theta}}_q$ is obtained from the misspecified density as follows:

$$\hat{\boldsymbol{\theta}}_q(\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} q(\mathbf{y},\boldsymbol{\theta}) \tag{3.16}$$

This estimate has a conditional mean given by:

$$\boldsymbol{\mu}_p(\boldsymbol{\theta}) = \mathbb{E}_{p_{\mathbf{y}|\boldsymbol{\theta}}}\left[\hat{\boldsymbol{\theta}}_q\right], \tag{3.17}$$

Then, a slightly different score function than the one used in BCRB to reflect the misspecified nature [19] is given by:

$$\boldsymbol{\eta}(\mathbf{y},\boldsymbol{\theta}) = \frac{\partial \ln \mathrm{q}(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^*} \text{-}\ \mathbb{E}_{\mathrm{p}_{\mathbf{y}|\boldsymbol{\theta}}}\{\frac{\partial \ln \mathrm{q}(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^*}\}\ . \tag{3.18}$$

Finally, the estimation error that is approximated using Taylor series as in [17] is given by:

$$\begin{aligned}\boldsymbol{\zeta}(\mathbf{y},\boldsymbol{\theta}) \approx -&\left(\frac{\partial^2 \ln \mathrm{q}(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^*\partial\boldsymbol{\theta}}^T + \frac{\partial^2 \ln q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^*\partial\boldsymbol{\theta}}^T\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\mu}_p}\right)^{-1}\\ &\times\left(\frac{\partial \ln \mathrm{q}(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^*} + \frac{\partial \ln \mathrm{q}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^*}\right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\mu}_p}\end{aligned} \tag{3.19}$$

It is assumed that MAP estimate is in the vicinity of $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_p(\boldsymbol{\theta})$. Then, we use the MBCRB bound derived for the MSE in [20] as follows,

$$\begin{aligned}&\mathbb{E}_{p_{\mathbf{y},\boldsymbol{\theta}}}\{[\hat{\boldsymbol{\theta}}_q(\mathbf{y}) - \boldsymbol{\mu}_p(\boldsymbol{\theta})][\hat{\boldsymbol{\theta}}_q(\mathbf{y}) - \boldsymbol{\mu}_p(\boldsymbol{\theta})]^H\} \geq \\ &\mathbb{E}_{p_{\boldsymbol{\theta}}}[\boldsymbol{\Xi}_{(\mathrm{p:q})}(\boldsymbol{\theta})^H]\, \mathbb{E}^{-1}_{p_{\boldsymbol{\theta}}}[\mathbf{J}_{(\mathrm{p:q})}(\boldsymbol{\theta})]\, \mathbb{E}_{p_{\boldsymbol{\theta}}}[\boldsymbol{\Xi}_{(\mathrm{p:q})}(\boldsymbol{\theta})]\end{aligned} \tag{3.20}$$

where

$$\mathbb{E}_{p_{\boldsymbol{\theta}}}\Big\{\, \boldsymbol{\Xi}_{(\mathrm{p:q})}(\boldsymbol{\theta})\, \Big\} = \mathbb{E}_{p_{\mathbf{y},\boldsymbol{\theta}}}\Big\{\boldsymbol{\zeta}(\mathbf{y},\boldsymbol{\theta})\boldsymbol{\eta}^H(\mathbf{y},\boldsymbol{\theta})\Big\} \tag{3.21}$$

and,

$$\mathbb{E}_{p_{\boldsymbol{\theta}}}\Big\{\, \mathbf{J}_{(\mathrm{p:q})}(\boldsymbol{\theta})\, \Big\} = \mathbb{E}_{p_{\mathbf{y},\boldsymbol{\theta}}}\Big\{\boldsymbol{\eta}(\mathbf{y},\boldsymbol{\theta})\boldsymbol{\eta}^H(\mathbf{y},\boldsymbol{\theta})\Big\} \tag{3.22}$$

Following the detailed derivation for the SBL model in [17], an MBCRB for GN model is given as follows:

$$\boxed{\mathbb{E}_{p_{\boldsymbol{\theta}}}\Big\{||\mathbf{x} - \hat{\mathbf{x}}||_2^2\Big\} \geq \sigma_w^2\ \mathbf{tr}\{\mathbf{HBH}\}} \tag{3.23}$$

where

$$\mathbf{H} = \mathbb{E}_{p_{\boldsymbol{\theta}}}\Big\{(\mathbf{B} + \sigma_w^2\boldsymbol{\Sigma}(\mathbf{x}))^{-1}\Big\} \tag{3.24}$$

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

$$[\boldsymbol{\Sigma}(\mathbf{x})]_k = \frac{2b}{(|x_k|^2 + b)^2}, k = 1, \ldots, N \tag{3.25}$$

Figure 3.4 shows the general Bayesian CRB in blue, and the MBCRB for GN model in red for low application SNR range. Figure 3.5 shows same bounds for a higher range of SNR. The SNR is defined, given an initial vector $\mathbf{x}$, to be SNR $\triangleq \frac{||\mathbf{x}_0||^2}{M\sigma_w^2}$.

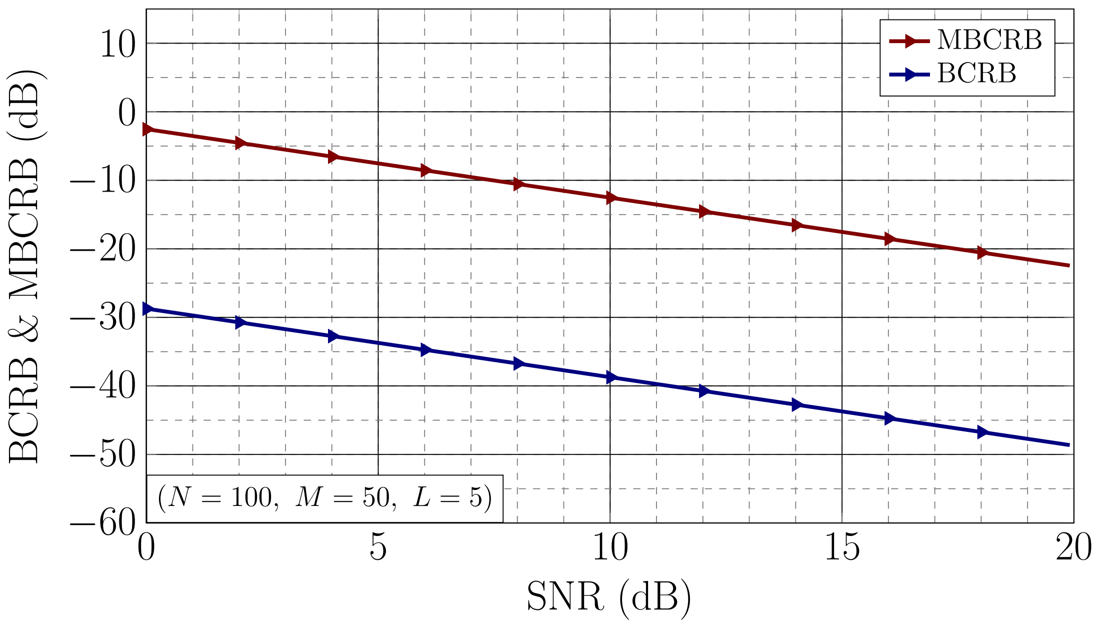

Figure 3.4: BCRB And MBCRB Derived Bounds

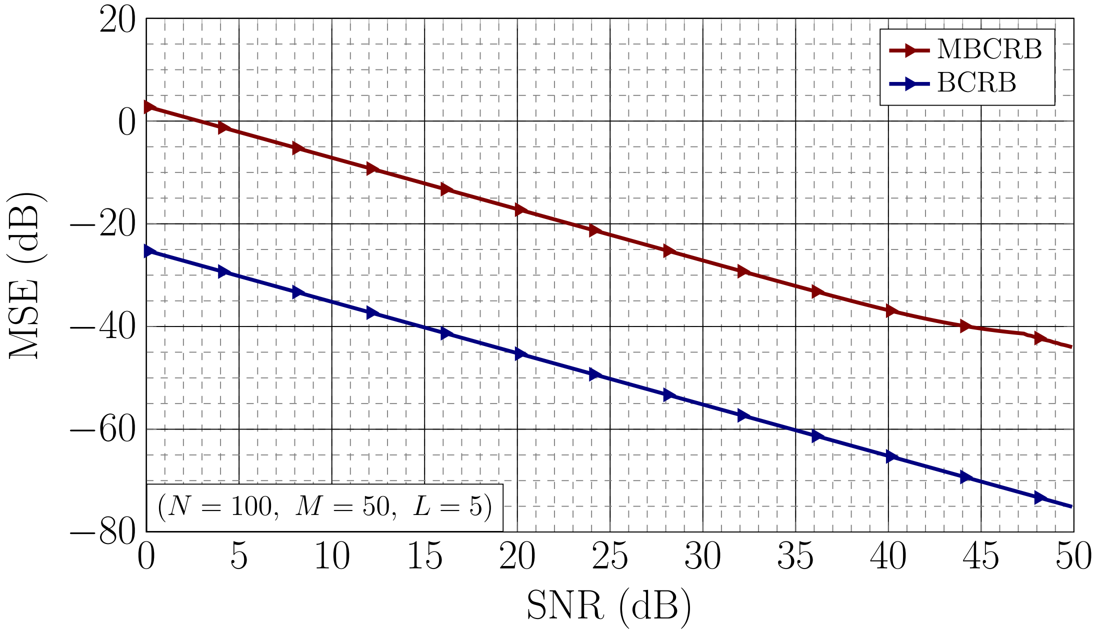

Figure 3.5: BCRB And MBCRB Derived Bounds

The following chapter presents two algorithms to estimate a sparse vector. One is to estimate a sparse vector assuming BG model. The other estimate it assuming GN model. It presents their MSE performances and Complexity measures for comparison.

Chapter 4

METHODOLOGY AND RESULTS

Previously, two models were introduced to favor sparsity on prior distributions $p(x)$ and $q(x)$. In this chapter, two widely-used algorithms are introduced to estimate a posterior density. First algorithm is Sparse Bayesian Learning (SBL) [21]. The way SBL recover $\mathbf{x}$ is by applying variational expectation maximization (VEM) in [22] on GN model. Second algorithm is Generalized Approximate Message Passing (GAMP) [23]. GAMP is a linear mixing algorithm that has proven to provide a decent estimation to any general independent identically distributed generative distribution.

4.1 Sparse Bayesian Learning

SBL is a hierarchical Bayesian model (HBM) developed to promote sparsity in a soft manner. This soft sparsity promotion was originally developed to ensure tractability compared to BG model that lead to non-informative prior. SBL technique and GN model constitute a duality that exploits useful conjugation of Gamma and Normal distributions to optimize posterior density. SBL depends on a widely-used optimization technique called Expectation maximization (EM).

4.1.1 Expectation Maximization Derivation

Expectation Maximization algorithm is an iterative algorithm to maximize the log likelihood of the observed data [24]. So assume we have a sequence of random variable Y and another sequence of a hidden random variable X and we would like to optimize the likelihood of the observed date as follows

$$L(\boldsymbol{\alpha}) = \sum_{y_i} \log p(y_i|\boldsymbol{\alpha}) = \sum_{y_i} \log \left[\sum_{x_i} p(y_i, x_i|\boldsymbol{\alpha}) \right] . \tag{4.1}$$

It is suggested to use an iterative way to go around the fact that the log cannot be pushed inside the sum

$$L(\boldsymbol{\alpha}) = \sum_{y_i} \log \left[\sum_{x_i} p(y_i, x_i|\boldsymbol{\alpha}) \right] = \sum_{y_i} \log \left[\sum_{x_i} q(x_i) \frac{p(y_i, x_i|\boldsymbol{\alpha})}{q(x_i)} \right]$$

$$\geq \sum_{y_i} \sum_{x_i} q(x_i) \log \frac{p(y_i, x_i|\boldsymbol{\alpha})}{q(x_i)} \triangleq Q(\boldsymbol{\alpha}, \mathbf{q}) \tag{4.2}$$

The inequality in(4.2) is an application of Jensen's inequality which states that the arithmetic mean is greater than or equal to the geometric mean:

$$\log \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \log(x_i) \ . \tag{4.3}$$

Going back to the defined function Q (i.e. auxiliary function). It can be written in the following manner [7]

$$Q(\boldsymbol{\alpha}, \mathbf{q}) = \sum_{i=1}^{N} E_{q_i} \left[\log p(y_i, x_i|\boldsymbol{\alpha}) \right] + H(q_i) \tag{4.4}$$

where H is the entropy of a distribution that is defined

$$H(\mathbf{q}) = -\sum_{x_i} q(x_i) \log q(x_i) \tag{4.5}$$

Now let's go back to the original upper bound the likelihood of $\boldsymbol{\alpha}$ to define another parameter that is useful to understand the algorithm

$$\begin{aligned}
L(\boldsymbol{\alpha}) &= \sum_{y_i} \log \left[\sum_{x_i} p(y_i, x_i|\boldsymbol{\alpha}) \right] \\
&= \sum_{y_i} \log \left[\sum_{x_i} q(x_i) \frac{p(y_i, x_i|\boldsymbol{\alpha})}{q(x_i)} \right] \\
&= \sum_{y_i} \log \left[\sum_{x_i} q(x_i) \frac{p(x_i|y_i, \boldsymbol{\alpha}) p(y_i|\boldsymbol{\alpha})}{q(x_i)} \right] \\
&= \sum_{y_i} \log \left[- D(q(x_i)||p(x_i|y_i, \boldsymbol{\alpha})) + \sum_{x_i} p(x_i) p(y_i|\boldsymbol{\alpha}) \right] \\
&= -D(q(x_i)||p(x_i|y_i, \boldsymbol{\alpha})) + \sum_{y_i} \log p(y_i|\boldsymbol{\alpha})
\end{aligned}$$

$$= -D(q||p_{X|y,\boldsymbol{\alpha}}) + L(\boldsymbol{\alpha}) \triangleq L(\boldsymbol{\alpha}, \mathbf{q}) \tag{4.6}$$

Now we are ready to form a general lower bound for an arbitrary hidden distribution $q(x_i)$:

$$L(\boldsymbol{\alpha}, \mathbf{q}) \geq Q(\boldsymbol{\alpha}, \mathbf{q}) \tag{4.7}$$

The iterative choice of $\boldsymbol{\alpha}$ will ensure some kind of convergence but will not guarantee a global maxima unless the initial choice was chosen carefully.

EM algorithm always chooses $q(x_i)$ to be $p(x_i|y_i, \boldsymbol{\alpha})$ because our goal is to maximize the likelihood.

- **E-step**

  given a choice of $\boldsymbol{\alpha}^t$ the expectation step will be a maximization of the lower bound (i.e. auxiliary function) as follows

$$\arg\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}^t, p_{\mathbf{x}|\mathbf{y},\boldsymbol{\alpha}^t}) = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} E_{p_{\mathbf{x}|\mathbf{y},\boldsymbol{\alpha}^t}} \left[\log p(y_i, x_i|\boldsymbol{\alpha}^t) \right] \tag{4.8}$$

- **M-step**

  We use the result of the optimization with respect to $\boldsymbol{\alpha}$ in the expectation step to define a new set of $\boldsymbol{\alpha}$'s by

$$\arg\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}^{t+1}, p_{\mathbf{x}|\mathbf{y},\boldsymbol{\alpha}^t}) = \arg\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}^t) \tag{4.9}$$

4.1.2 SBL VEM Algorithm

In our analysis, both variables $\mathbf{x}, \boldsymbol{\alpha}$ are assumed to be hidden. This case has only an E-step [22]. Hence, EM algorithm is further approximated to be VEM under some conditions. According to [22] stationarity and independence allows us to use VEM.

Following VEM, posterior of hidden parameter can be seen as independent marginals.

$$q(\mathbf{x}, \boldsymbol{\alpha}|\mathbf{y}) \approx q(\mathbf{x})q(\boldsymbol{\alpha}) \tag{4.10}$$

$$\log q(\mathbf{x}) \approx \langle \log q(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha})_{q(\boldsymbol{\alpha})} \rangle + \text{constant} \tag{4.11}$$

$$\log q(\boldsymbol{\alpha}) \approx \langle \log q(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha})_{q(\mathbf{x})} \rangle + \text{constant} \tag{4.12}$$

where $\langle . \rangle$ denotes a varitional expected value for the subscript distribution. From that a posterior of each marginal can be approximated as follows

$$q(\mathbf{x}) = \mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Phi}) \tag{4.13}$$

$$q(\boldsymbol{\alpha}) = \prod_{i=1}^{N} Gamma(\tilde{a}_i, \tilde{b_i}) \tag{4.14}$$

where the parameters $\{\boldsymbol{\mu}, \boldsymbol{\Phi}, \tilde{a}_i, \tilde{b_i}\}$ are derived using equations (4.10-12)

$$\boldsymbol{\mu} = \sigma_w^2 \boldsymbol{\Phi} \mathbf{A} \mathbf{y} \tag{4.15}$$

$$\boldsymbol{\Phi} = [\sigma_w^2 \ \mathbf{B} + \boldsymbol{\Sigma}]^{-1} \tag{4.16}$$

$$\tilde{a_i} = a_i + 1 \tag{4.17}$$

$$\tilde{b_i} = b_i + |\mu_i|^2 + \Phi_{ii} \tag{4.18}$$

Hyperparametes a_i,b_i is up to the designer belief. Usually they are chosen to avoid being stuck in a local minima. Also, they are used to characterize the desired application. In this step, hyperparameters are chosen to be

$$a_i = 10^{-6} \tag{4.19}$$

$$b_i = \begin{cases} 10^{-1} & i \in \mathbb{S} \\ 10^{-6} & i \in \mathbb{S}^c \end{cases} \tag{4.20}$$

where $\mathbb{S}$ is a set of non-zero objects of a cardinality $|\mathbb{S}| = L$. Figure 4.1 shows GN density with particular hyperparameters. They are chosen to promote soft sparsity.

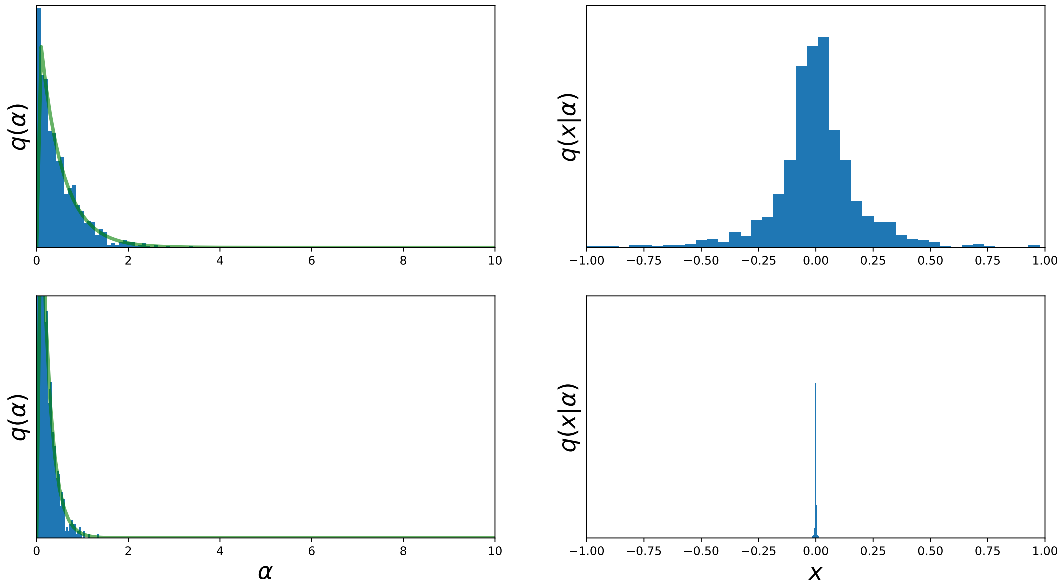

Figure 4.1: Gamma Normal Distribution With The Specified Hyperparameters

A complete algorithm of SBL-VEM is described below [25]

Algorithm 1 SBL-VEM

```
Initialize:
    $\langle \alpha_i^0 \rangle = \tilde{a}_i / \tilde{b}_i^0,\ \mathbf{\Sigma}^0 \triangleq \mathrm{diag}\{\alpha_{\mathbf{i}}{}^0\},$
    $i = 1, \ldots, N,\ \tau = 0, \epsilon = 10^{-2}$

  repeat
    (1) $\tau = \tau + 1$
    (2) $\mathbf{\Phi}^\tau = [\sigma_w^2 \mathbf{B} + \mathbf{\Sigma}^{\tau-\mathbf{1}}]^{-1}$
    (3) $\boldsymbol{\mu}^\tau = \sigma_w^2 \mathbf{\Phi}^\tau \mathbf{A}\mathbf{y}$
    (4) $\tilde{b}_i^\tau = b_i + |\mu_i^\tau|^2 + \Phi_{ii}^\tau$
    (5) $\langle \alpha_i^\tau \rangle = \tilde{a}_i / \tilde{b}_i^\tau$
    (6) update $\mathbf{\Sigma}^\tau = \mathrm{diag}\{\alpha_i^\tau\}$
  until $\left\{ \frac{||\boldsymbol{\mu}^\tau - \boldsymbol{\mu}^{\tau-1}||_2}{||\boldsymbol{\mu}^\tau||_2} \leq \epsilon \right\}$
output $\hat{\mathbf{x}} \sim \mathcal{N}_c(\boldsymbol{\mu}^\tau, \mathbf{\Phi}^\tau)$
```

4.1.3 SBL VEM With A Prior On Hyperparameter b

Previous set up assumed a deterministic value for the hyperparameter $\tilde{\mathbf{b}}$. Another set up can be tested by adding an extra step on the hierarchical model by assuming that $\tilde{\mathbf{b}}$ is a random variable itself as in [25]. To be consistent, lets call it $\boldsymbol{\beta}$ where its distribution

$$q(\boldsymbol{\beta}) = \prod_{i=1}^{N} q(\beta_i) \tag{4.21}$$

$$q(\beta_i) = \begin{cases} Gamma(\beta_i; c, d) & i \in \mathbb{S} \\ \delta(\beta_i - 10^{-6}) & i \in \mathbb{S}^c \end{cases} \tag{4.22}$$

Again, hyperparameters are chosen to help achieving a better estimation of posterior density. In this case there are initialized to be $c = d = 10^{-6}$.

Consequently, expressions on equations (4.10-12) has changed to

$$q(\mathbf{x}, \boldsymbol{\alpha}|\mathbf{y}) \approx q(\mathbf{x})q(\boldsymbol{\alpha})q(\boldsymbol{\beta}) \tag{4.23}$$

$$\log q(\mathbf{x}) \approx \langle \log q(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})_{q(\boldsymbol{\alpha}), q(\boldsymbol{\beta})} \rangle + \text{constant} \tag{4.24}$$

$$\log q(\boldsymbol{\alpha}) \approx \langle \log q(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})_{q(\mathbf{x}), q(\boldsymbol{\beta})} \rangle + \text{constant} \tag{4.25}$$

$$\log q(\boldsymbol{\beta}) \approx \langle \log q(\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})_{q(\mathbf{x}), q(\boldsymbol{\alpha})} \rangle + \text{constant} \tag{4.26}$$

Similarly, a posterior of marginals can be written as

$$q(\mathbf{x}) = \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Phi}) \tag{4.27}$$

$$q(\boldsymbol{\alpha}) = \prod_{i=1}^{N} Gamma(\tilde{a}_i, \tilde{b}_i) \tag{4.28}$$

$$q(\boldsymbol{\beta}) = \prod_{i=1}^{N} Gamma(\tilde{c}_i, \tilde{d}_i) \tag{4.29}$$

and parameters $\{\tilde{a}_i, \tilde{b}_i, \tilde{c}_i, \tilde{d}_i\}$ are now defined as

$$\tilde{a}_i = a_i + 1 \tag{4.30}$$

$$\tilde{b}_i = \begin{cases} \langle b_i \rangle + |\mu_i|^2 + \boldsymbol{\Phi}_{ii} & i \in \mathbb{S} \\ b_i + |\mu_i|^2 + \boldsymbol{\Phi}_{ii} & i \in \mathbb{S}^c \end{cases} \tag{4.31}$$

$$\tilde{c}_i = c_i + 1 \tag{4.32}$$

$$\tilde{d}_i = d_i + \langle \alpha_i \rangle \tag{4.33}$$

where

$$a_i = 10^{-6} \quad b_i = 10^{-1} \quad c_i = 10^{-6} \quad d_i = 10^{-6} \tag{4.34}$$

$$\langle b_i \rangle = \frac{c_i + a_i}{\tilde{d}_i} \tag{4.35}$$

$$\langle \alpha_i \rangle = \frac{\tilde{a}_i}{\tilde{b}_i} \tag{4.36}$$

The complete algorithm in detail is described below [25]

Algorithm 2 SBL VEM with a prior on b

```
Initialize:
    $\langle \alpha_i^0 \rangle = \tilde{a}_i / \tilde{b}_i^0,\ \mathbf{\Sigma}^0 \triangleq \mathrm{diag}\{\alpha_i^0\},$
    $i = 1, \ldots, N,\ \tau = 0, \epsilon = 10^{-2}$
  repeat
    (1) $\tau = \tau + 1$
    (2) $\mathbf{\Phi}^\tau = [\sigma_w^2 \mathbf{B} + \mathbf{\Sigma}^{\tau - \mathbf{1}}]^{-1}$
    (3) $\mu^\tau = \sigma_w^2 \mathbf{\Phi}^\tau \mathbf{A} \mathbf{y}$
    (4) $\tilde{b}_i^\tau = \begin{cases} \langle b_i^{\tau-1} \rangle + |\mu_i^\tau|^2 + \Phi_{ii}^\tau & i \in \mathbb{S} \\ b_i + |\mu_i^\tau|^2 + \Phi_{ii}^\tau & i \in \mathbb{S}^c \end{cases}$
    (5) $\langle \alpha_i^\tau \rangle = \tilde{a}_i / \tilde{b}_i^\tau$
    (6) $\tilde{d}_i^\tau = d_i + \langle \alpha_i \rangle$
    (7) $\langle b_i^{\tau-1} \rangle = (c_i + a_i) / \tilde{d}_i^\tau$
    (8) update $\mathbf{\Sigma}^\tau = \mathrm{diag}\{\alpha_{\mathbf{i}}{}^\tau\}$
  until $\left\{ \frac{||\boldsymbol{\mu}^\tau - \boldsymbol{\mu}^{\tau-1}||_2^2}{||\boldsymbol{\mu}^\tau||_2^2} \leq \epsilon \right\}$
output $\hat{\mathbf{x}} \sim \mathcal{N}_c(\boldsymbol{\mu}^\tau, \mathbf{\Phi}^\tau)$
```

4.2 Generalized Approximated Message Passing

G-AMP is a generalization of AMP algorithm first proposed in [26]. AMP is an iterative algorithm that leverage central limit theorem (CLT) with a dense measurement matrix $\mathbf{A}$. GAMP is an efficient algorithm when using iid under generalized linear model.

We assume in this set up that we have the following system.

$$\mathbf{y} = \mathbf{z} + \mathbf{w} \tag{4.37}$$

where

$$\mathbf{z} = \mathbf{A}\mathbf{x} \tag{4.38}$$

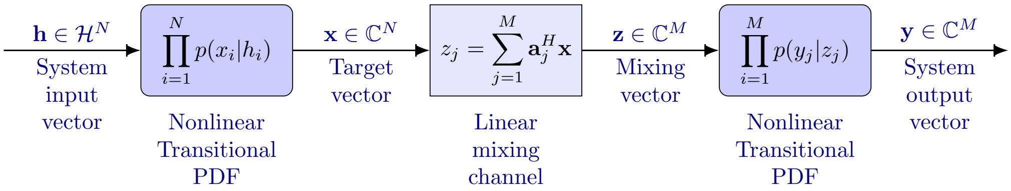
$\mathbf{h} \in \mathcal{H}^N$
System input vector
$\prod_{i=1}^{N} p(x_i|h_i)$
Nonlinear Transitional PDF
$\mathbf{x} \in \mathbb{C}^N$
Target vector
$z_j = \sum_{j=1}^{M} \mathbf{a}_j^H \mathbf{x}$
Linear mixing channel
$\mathbf{z} \in \mathbb{C}^M$
Mixing vector
$\prod_{i=1}^{M} p(y_j|z_j)$
Nonlinear Transitional PDF
$\mathbf{y} \in \mathbb{C}^M$
System output vector

Figure 4.2: A General Estimation Problem With Linear Mixing Channel

Also, lets assume that there is an input vector $\mathbf{h}$ that has components $h_i \in \mathcal{H}$, where the set $\mathcal{H}$ is assumed to be the set of arbitrary hyperparameters that generates some arbitrary random vector $\mathbf{x} \in \mathbb{C}^{\mathbb{N}}$ in component-wise fashion. This can be thought of as an input non-linear channel. Then vector $\mathbf{x}$ is taken to a linear transform matrix $\mathbf{A}$ to generate another random vector $\mathbf{z} \in \mathbb{C}^M$. Finally this unknown vector is the input of another channel where the system output vector of this channel is the results vector $\mathbf{y} \in \mathbb{C}^M$. In this linear set up, this linear mixing process aims to estimate two vectors. First $\mathbf{z}$ is estimated. Then, using estimated value of $\mathbf{z}$, $\mathbf{x}$.

Figure 4.2 shows a diagram describing the previous process. The problem can be summarized by estimating two vectors $\mathbf{x}$ and $\mathbf{z}$, given system input vector $\mathbf{h}$, system output vector $\mathbf{y}$, system transform matrix $\mathbf{A}$, and transitional probability functions $p(\mathbf{x}|\mathbf{h})$, $p(\mathbf{y}|\mathbf{z})$. This algorithm has wide applicability in Bayesian estimation literature due to its generality and low complexity. For instance, in [27] they claim that this model fits in set of hyperparameters on target vector $\mathbf{x}$. A complete description of the procedure and derivations can be found in [23, 27]. However, GAMP algorithm for BG model is described as follows

4.2.1 BG GAMP Algorithm

Back to BG model in equation (3.2) and Figure 3.1, GAMP can have an arbitrary relationship between the observation vector $\mathbf{y}$ and the noiseless components of vector $\mathbf{z}$. Meaning that the conditional distribution can be written as

$$p(y_i|z_i) = \mathcal{N}_c(z_i, \sigma_w^2) \tag{4.39}$$

According to [23], GAMP is fully described by the following four equations

$$g_{\text{out}}(y_i, \hat{z}_i, \mu_i^z; \mathbf{h}) = \frac{y_i - \hat{z}_i}{\mu_i^z + \sigma_w^2} \tag{4.40}$$

$$-\acute{g}_{\text{out}}(y_i, \hat{z}_i, \mu_i^z; \mathbf{h}) = \frac{1}{\mu_i^z + \sigma_w^2} \tag{4.41}$$

$$g_{\text{in}}(\hat{r}_i, \mu_i^r; \mathbf{h}) = \pi(\hat{r}_i, \mu_i^r; \mathbf{h})\nu(\hat{r}_i, \mu_i^r; \mathbf{h}) \tag{4.42}$$

$$\mu_i^r \acute{g}_{\text{in}}(\hat{r}_i, \mu_i^r; \mathbf{h}) = \pi(\hat{r}_i, \mu_i^r; \mathbf{h})\Big(\nu(\hat{r}_i, \mu_i^r; \mathbf{h}) + |\kappa(\hat{r}_i, \mu_i^r; \mathbf{h})|^2\Big) \tag{4.43}$$

$$- \Big(\pi(\hat{r}_i, \mu_i^r; \mathbf{h})\Big)^2 |\kappa(\hat{r}_i, \mu_i^r; \mathbf{h})|^2 \tag{4.44}$$

where

$$\pi(\hat{r}_i, \mu_i^r; \mathbf{h}) \triangleq \frac{1}{1 + (\frac{\lambda_i}{1-\lambda_i} \frac{\mathcal{N}(\mathrm{\hat{r}_i;0,v_i+\mu_i^r}}{\mathcal{N}(\mathrm{\hat{r}_i;0,+\mu_i^r}})^{-1}} \tag{4.45}$$

$$\kappa(\hat{r}_i, \mu_i^r; \mathbf{h}) \triangleq \frac{\hat{r}_i/\mu_i^r}{1/\mu_i^r + 1/v_i} \tag{4.46}$$

$$\nu(\hat{r}_i, \mu_i^r; \mathbf{h}) \triangleq \frac{1}{1/\mu_i^r + 1/v_i} \tag{4.47}$$

and the marginal posterior can be formalized to be

$$p(x_i|\mathbf{y}; \mathbf{h}) = \frac{1}{C_n} p_x(x_i; \mathbf{h}) \times \ \mathcal{N}_c(x_i; \hat{r}_i, \mu_i^r) \tag{4.48}$$

$$= \frac{1}{C_n}(1 - \lambda_i)\delta(x_i) + (\lambda_i)\mathcal{N}_c(x_i; 0, v_i) \times \mathcal{N}_c(x_i; \hat{r}_i, \mu_i^r) \tag{4.49}$$

where

$$C_n \triangleq \int p(x_i; \mathbf{h})\mathcal{N}_c(x_i; \hat{r}_i, \mu_i^r) dx_i \tag{4.50}$$

Each non zero entry of the sparse vector is distributed as follows

$$Pr\{x_i \neq 0|\mathbf{y}; \mathbf{h}\} = \pi(\hat{r}_i, \mu_i^r; \mathbf{h}) \tag{4.51}$$

4.3 Results

After showing the Bayesian bounds along with the algorithms to estimate a posterior PDF, it is time to see results of estimation. There are two main metrics to evaluate each method. First metric is called **computational gain**, which refers to the number of computations needed to achieve the estimate compared to the basic method (Brute force). Second is called **SNR losses**, which refers to the dBs difference between the minimum bound and the averaged mean square error of a particular algorithm at the desired value of a signal to noise ratio. Specifically, if a correct model is assumed, a comparison is initiated between the MSE of an algorithm and BCRB. Otherwise, the comparison is between the MSE and MBCRB. Furthermore, there is a theoretical comparison between BCRB and MBCRB, to evaluate the performance of any algorithm uses that specific family of distributions to the correct family of distribution.

4.3.1 MSE Results

Recall that bounds calculated in Figure 3.4 was assumed for an application with low SNR requirement such as MRI imaging acquisition [9]. Figure 4.3 shows the MSE results along with the derived bounds. It can be seen that SBL VEM β has a better results than SBL VEM b. However, both of them are above MBCRB bound. Also Figure 4.3 shows that GAMP which is built on BG model achieved better results than both algorithms that assumed GN model. Still, GAMP is above BCRB bound.

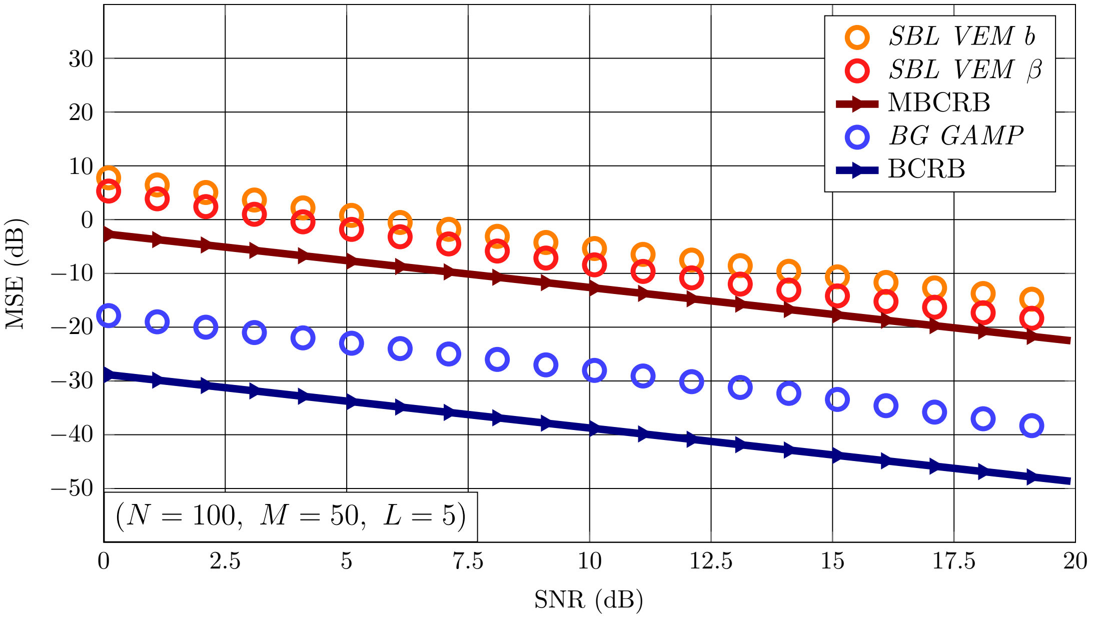

Figure 4.3: MSE Results Along With The Bounds For Low SNR Applications

Although, GAMP performs better than any algorithm that uses GN family of distributions, it is sill within few dBs from the bound. On the other hand, both of the other two algorithms that uses GN family is above the derived misspecified bound, as expected. Recall from Figure 4.1 how the hyperparameter b affected the sparsity level. Figure 4.4 shows the effect of choosing a smaller value of hyperparameter b (i.e. increasing sparsity L).

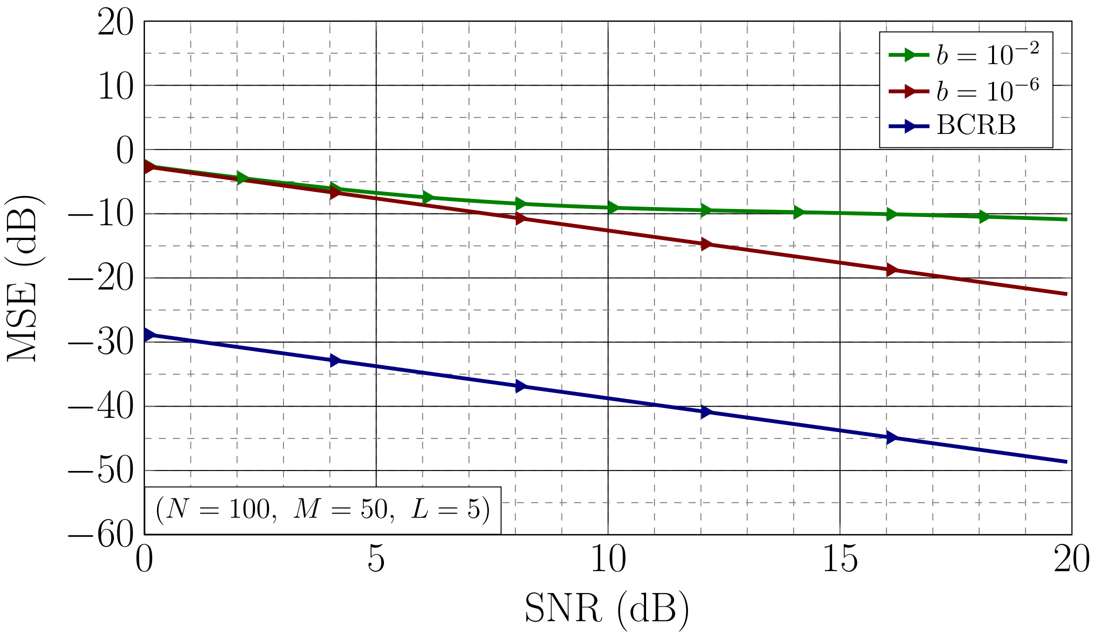

Figure 4.4: MBCRB With Different Choices Of Hyperparameter b

For better appreciation for the derived bounds, an MSE results are calculated for applications that requires higher SNR. Figure 4.5 shows similar results of Figure 4.3 for higher SNR level and Figure 4.6 show the effect of hyperparameter choice on the MBCRB.

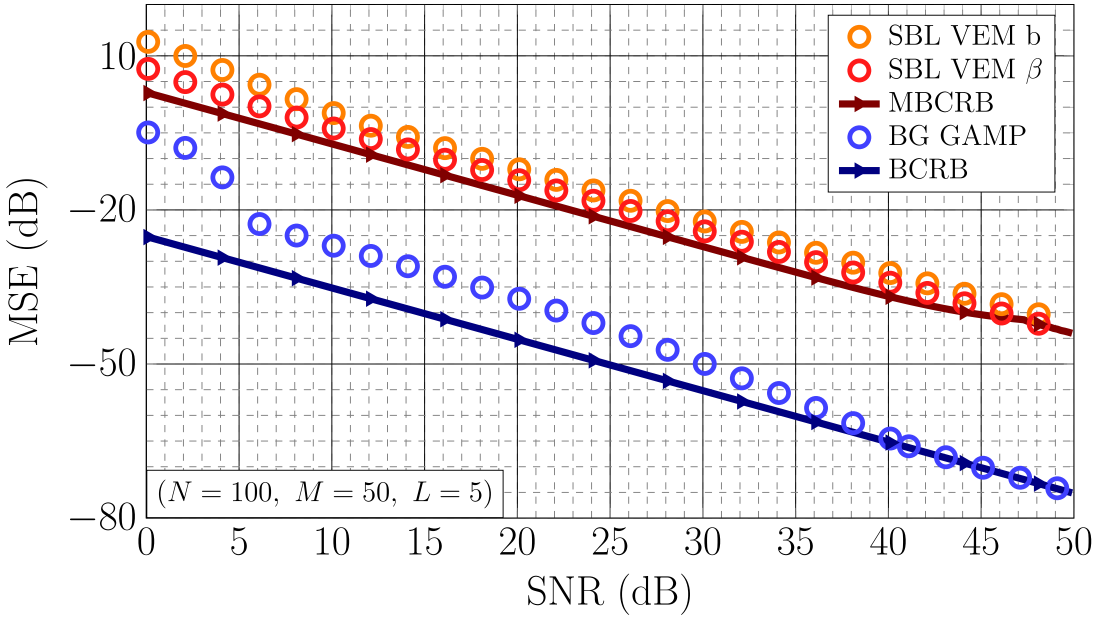

Figure 4.5: MSE Results Along With The Bounds For High SNR Applications

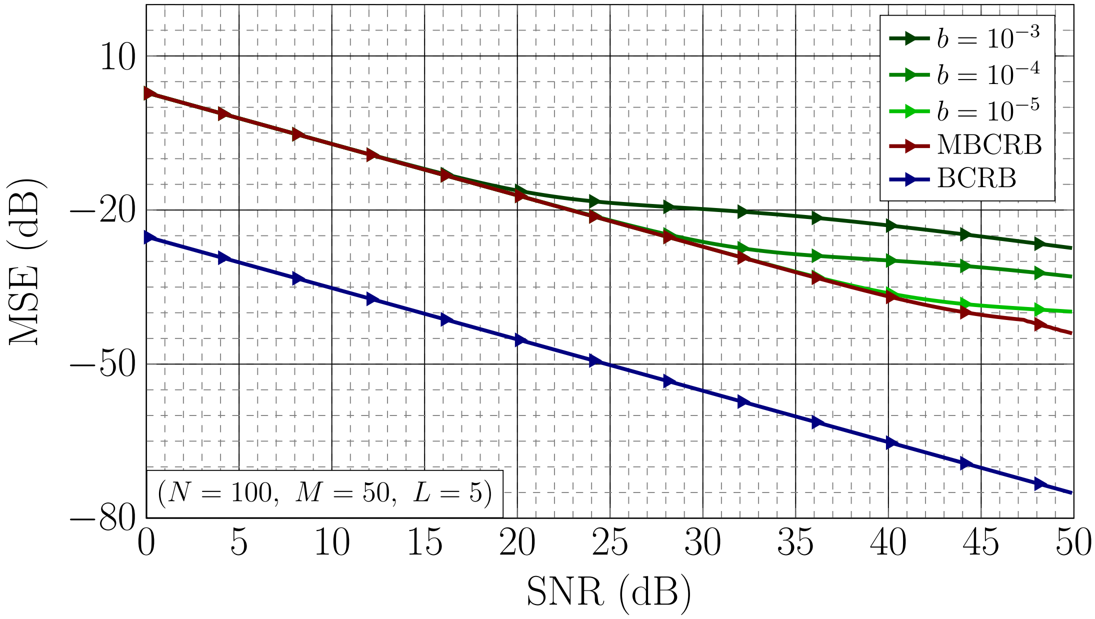

Figure 4.6: MBCRB With Different Choices Of Hyperparameter b

It can be clearly seen from the above results that a bound for both family models can be derived and validated to quantify which algorithm is better for the problem of interest.

To see the MSE results for different dimensions of $\mathbf{x}$, Figures 4.7-12 show the bounds with MSE results for low SNR. It is, again, validated that all MSE results are above specific bounds.

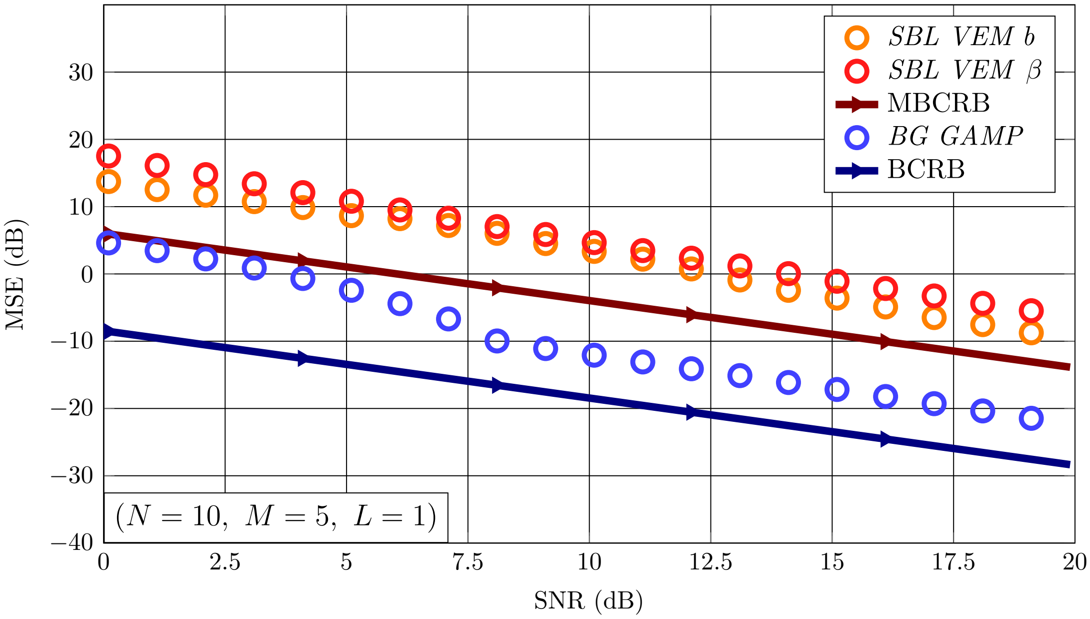

Figure 4.7: MSE Results Along With The Bounds For Vector Size N=10

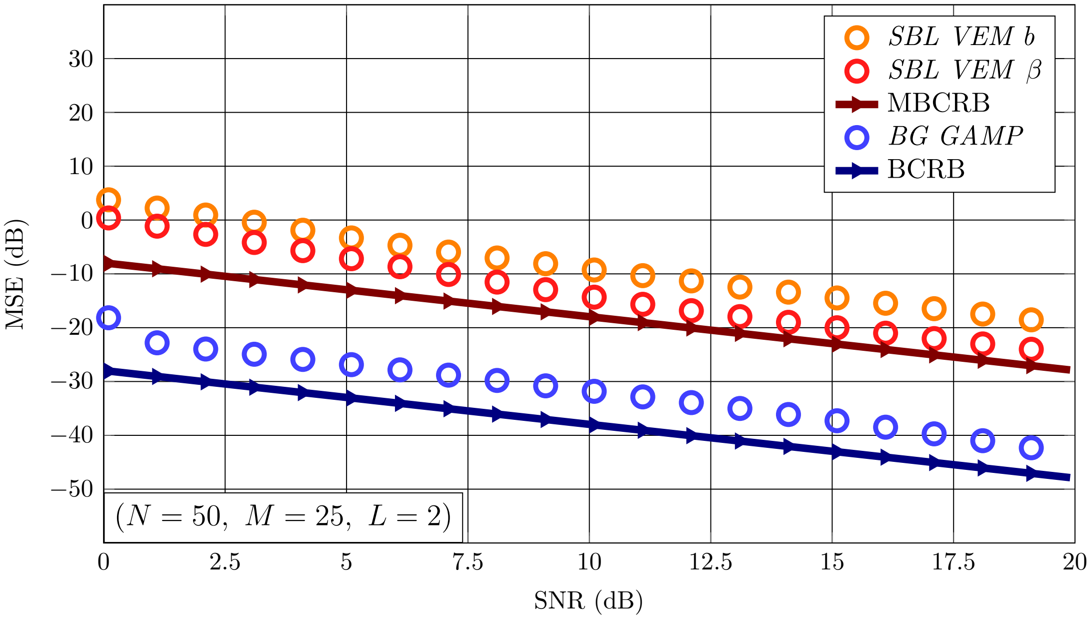

Figure 4.8: MSE Results Along With The Bounds For Vector Size N=50

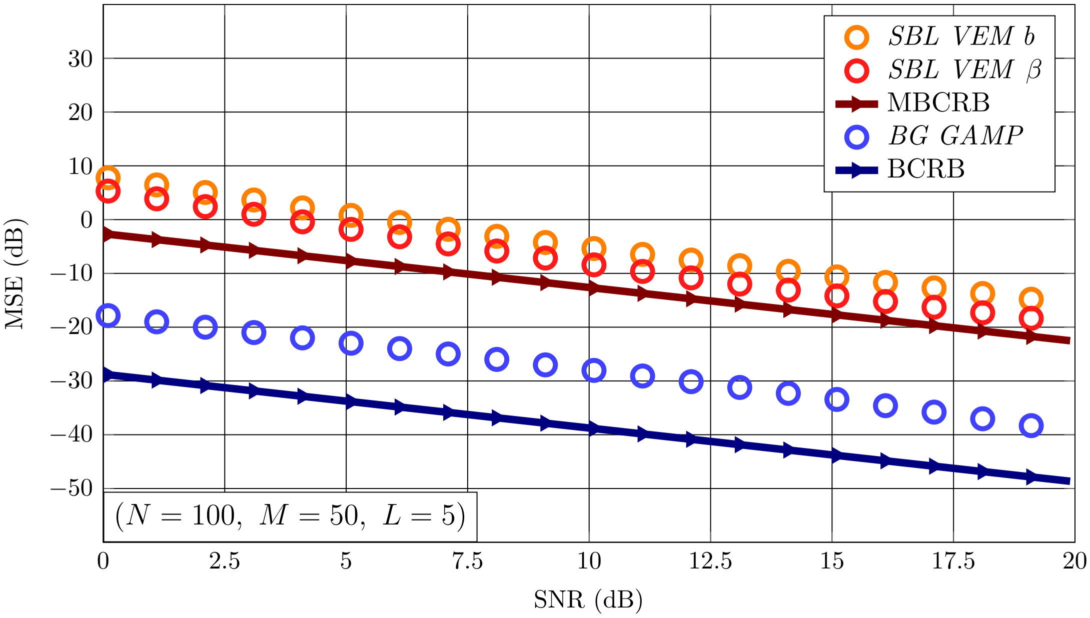

Figure 4.9: MSE Results Along With The Bounds For Vector Size N=100

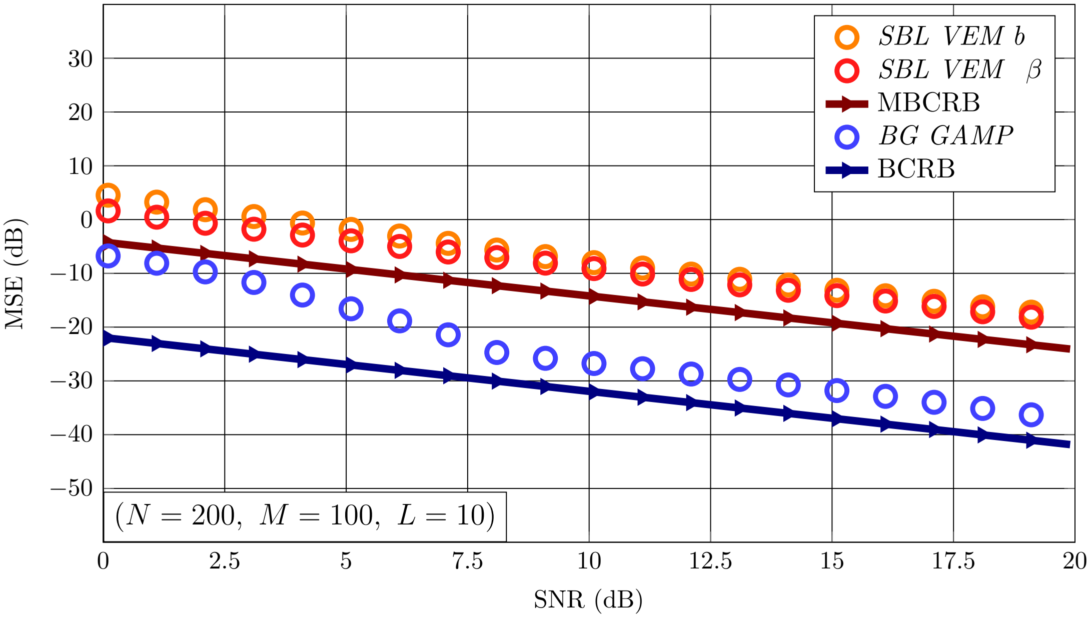

Figure 4.10: MSE Results Along With The Bounds For Vector Size N=200

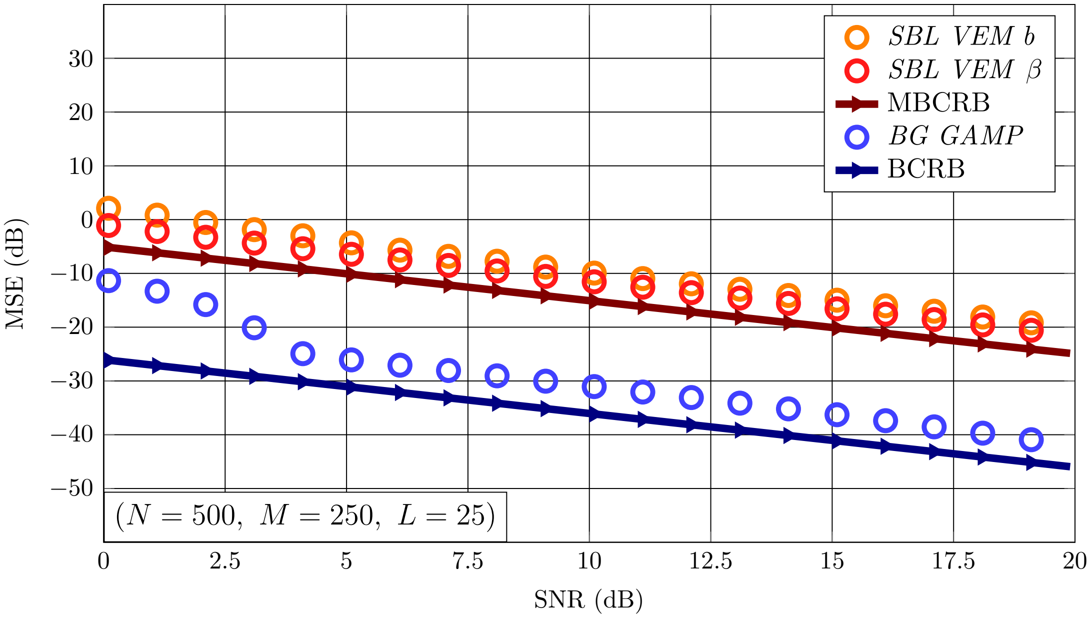

Figure 4.11: MSE Results Along With The Bounds For Vector Size N=500

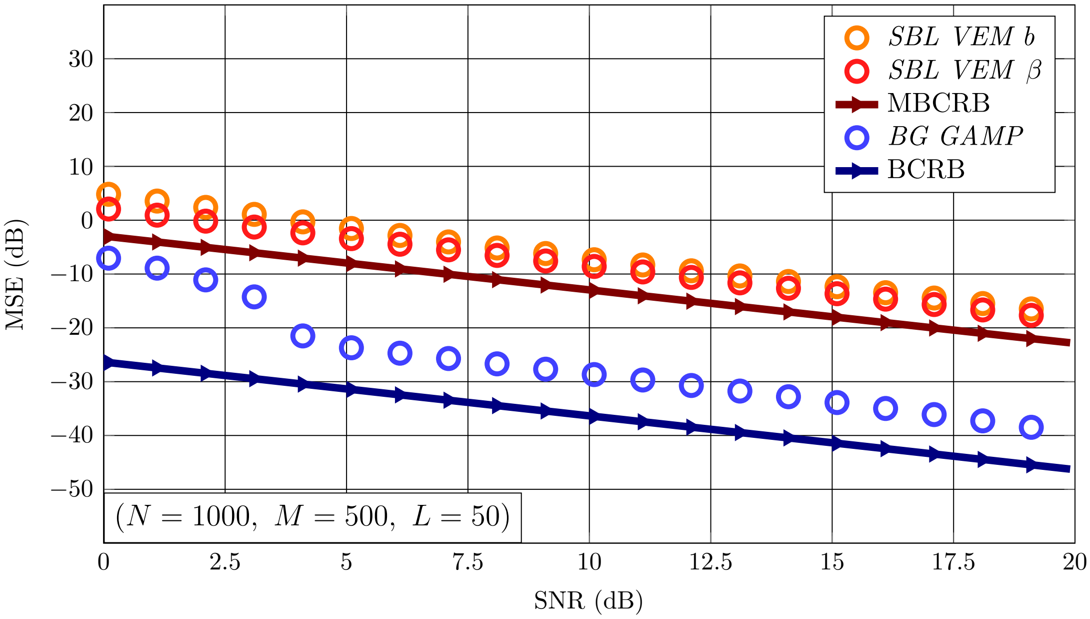

Figure 4.12: MSE Results Along With The Bounds For Vector Size N=1000

4.3.2 Computational Gain

One of the main objectives of using Bayesian framework to model sparsity is to reduce the complexity of checking each entry of the target vector. To better understand this comparison, this section compares computational complexity of both algorithms to brute force. SBL VEM complexity is dominated by the update of the inverted matrix Φ in equation (4.16) which has N dimensions. At worst case scenario the loop will repeat N times before convergence. Hence complexity of SBL VEM is $\mathcal{O}(N^3)$. However, this complexity can be reduced to $\mathcal{O}(N^2M)$ using the matrix inversion lemma as in [25]. For GAMP algorithm, [27] indicates a detailed analysis of the computational complexity that is dominated by the matrix-vector multiplication of the dictionary $\mathbf{A}$. Hence, the worst case complexity is $\mathcal{O}(NM)$.

As seen above, both algorithms are dominated by a dictionary $\mathbf{A}$ dimensions. Table 4.1 shows variations of the dictionary matrix as size expands.

Table 4.1: Computational Complexity Per Size N Of The Three Methods

(N,M,L)	Brute force $\mathcal{O}(N^L)$	SBL VEM $\mathcal{O}(N^2M)$	GAMP $\mathcal{O}(NM)$
$(10, 5, 1)$	10	500	50
$(50, 25, 2)$	2500	62500	1250
$(100, 50, 5)$	$1M$	50000	5000
$(200, 100, 10)$	200^{10}	$4M$	20000
$(500, 250, 25)$	500^{25}	$62.5M$	125000
$(1000, 500, 50)$	1000^{50}	$500M$	$0.5M$

To better evaluate the major improvement in complexity, Figure 4.13 shows the number of computations in log scale with respect to the size N.

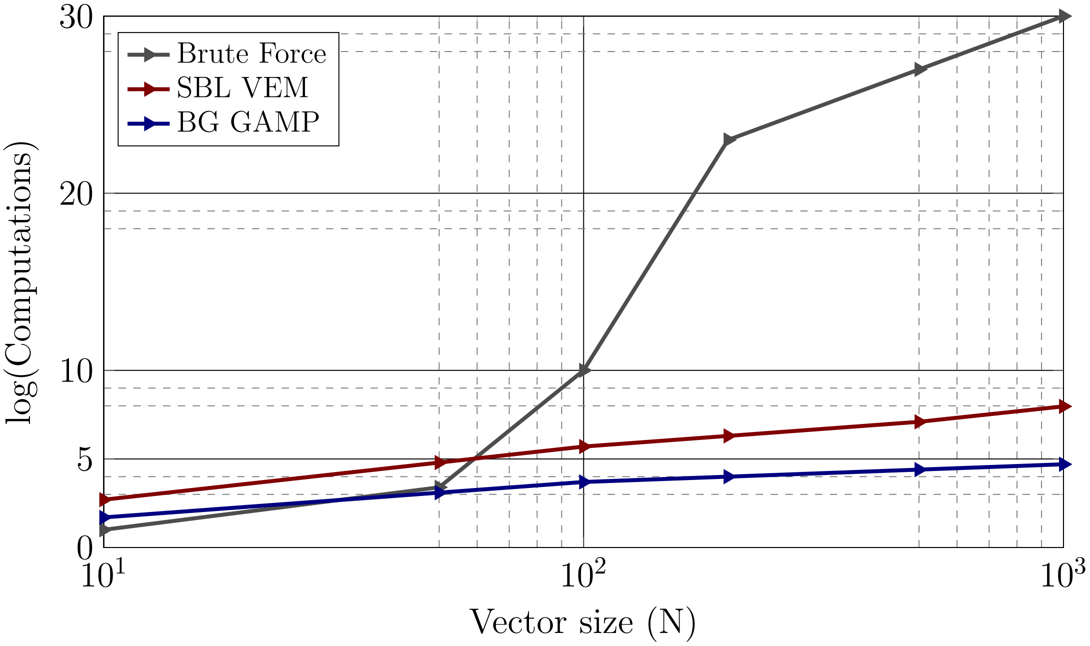

Figure 4.13: Comparative Computational Complexity Per Size N

4.3.3 MSE Accuracy

In chapter 3, a derivation of both models is introduced. Figures 4.15-15 show the difference in dBs each algorithm and the derived bound for its family of distribution.

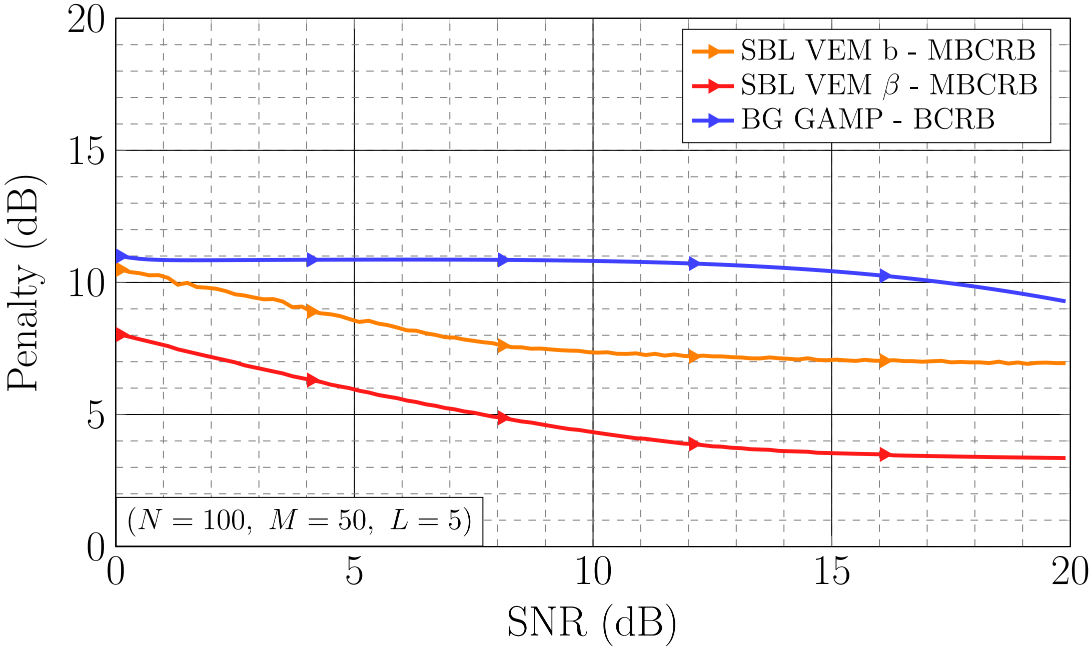

Figure 4.14: The Penalty In dBs Using An Algorithm To The Corresponding Bound For Low SNR

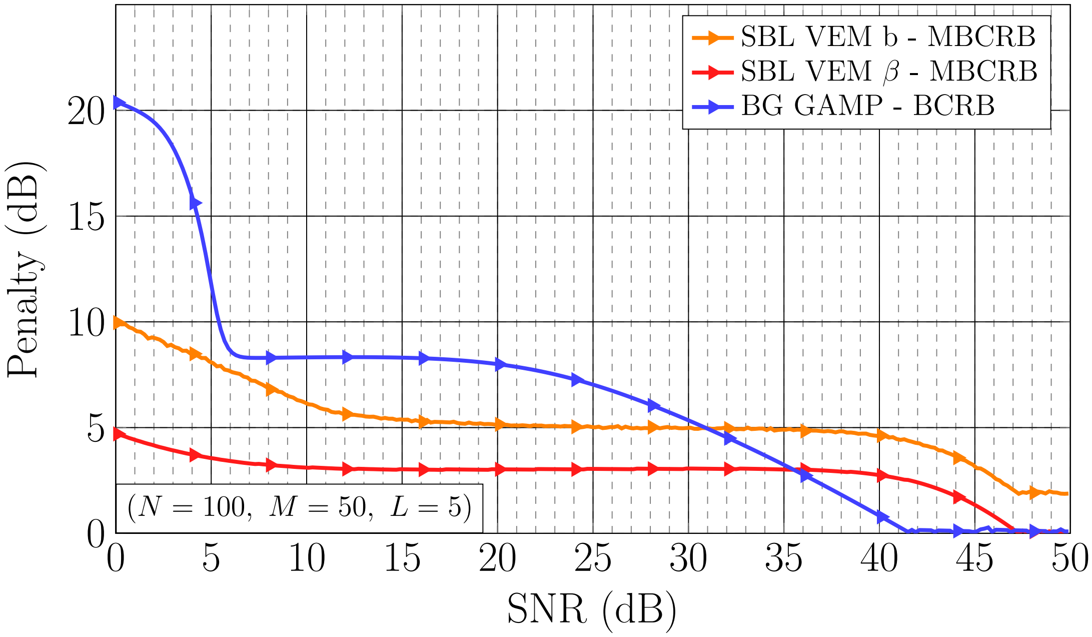

Figure 4.15: The Penalty In dBs Using An Algorithm To The Corresponding Bound For High SNR

4.3.4 Gain-Loss Metric

One of the main objectives of this thesis is quantifying the losses of using the misspecified distribution compared to the correctly specified distribution. It is important to show what has been sacrificed against what has been achieved Figures 4.16-4.17 shows the gain in computational complexity moving away from brute force to either SBL VEM or GAMP versus the Losses occured from moving to the misspecified bound (i.e. GN model). Gain and loss equations are quantified as follows:

$$\text{Gain (SBL VEM)} = \frac{\log(N^L)}{\log(N^2 M)} \tag{4.52}$$

$$\text{Gain (BG GAMP)} = \frac{\log(N^L)}{\log(NM)} \tag{4.53}$$

$$\text{Loss (dBs) (SBL VEM)} = \text{SBL VEM} - \text{BCRB} \tag{4.54}$$

$$\text{Loss (dBs) (BG GAMP)} = \text{BG GAMP} - \text{BCRB} \tag{4.55}$$

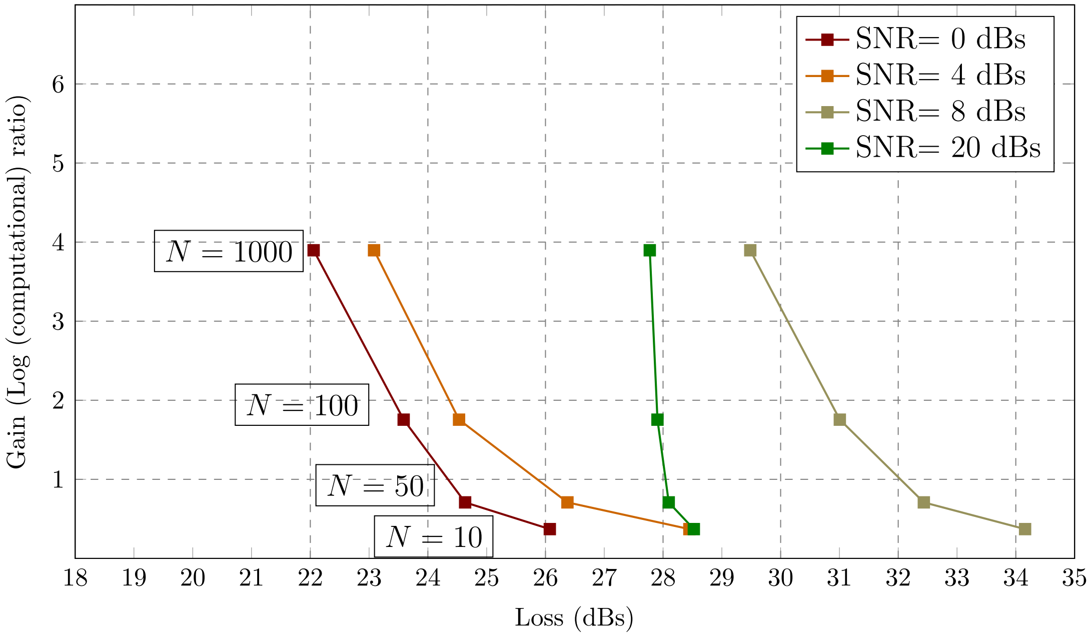

Figure 4.16: SBL VEM Loss From Equation (4.54), SBL VEM Gain From Equation (4.52)

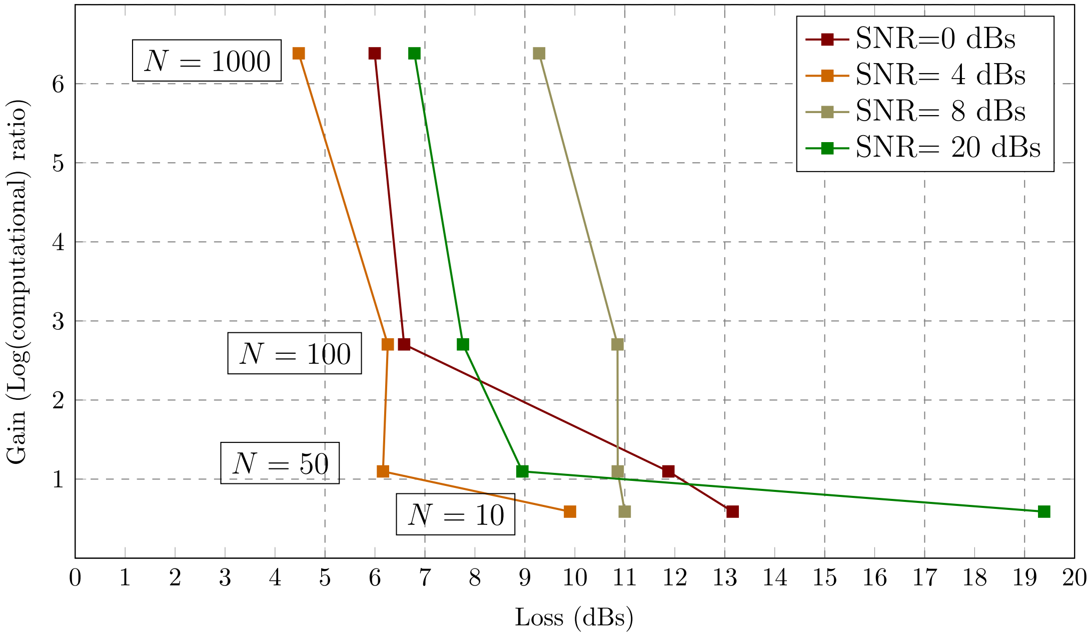

Figure 4.17: GAMP Loss In Equation (4.55), GAMP Gain From Equation (4.53)

Chapter 5

CONCLUSIONS AND FUTURE WORK

In this thesis, a Bayesian framework is adopted to recover a sparse target vector $\mathbf{x}$ with the knowledge of an observed vector $\mathbf{y}$ and a Dictionary $\mathbf{A}$. Sparse vector recovery is a well-known problem in the digital signal processing community. Many applications are interested in modeling this problem, a well-known application is CS. Two family of distributions are assumed to estimate the sparse vector, one is the BG model and assumed to be the correct generative model. The other is the GN model and assumed to have more biased in modeling the sparse vector. The aim of using a Bayesian framework other than estimating a posterior density, is to provide a way to evaluate estimation using a theoretical minimum bound on the MSE. MSE performances for both algorithms validated both bounds. MSE results using GN model was above MBCRB, while MSE results of BG model was above BCRB. Also, this research provided a quantified meteric of gain in terms of computational complexity and loss in terms of dBs. This research showed that MSE results for both algorithms are much more efficient than combinatorial searches(i.e. brute force). This work can be extended to meet any Bayesian framework given that a closed form of a CRB type equation is derived. Work can be extended to guarantee a theoretical bound for other families of distribution. Since GAMP is not-model-specific algorithm, it can be applied to GN model to compare results with SBL VEM. Many algorithms in literature use Bayesian framework without validating their MSE results rather than comparing them with other similar results. This work can guide them through the process to achieve that.

REFERENCES

[1] E. J. Candes and M. B. Wakin, “An introduction to compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, March 2008.

[2] J. A. Tropp and S. J. Wright, “Computational methods for sparse solution of linear inverse problems,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 948–958, June 2010.

[3] A. Miller, *Subset Selection in Regression.* New York: Chapman and Hall/CRC, 2002, vol. 2.

[4] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, Jan. 2001. [Online]. Available: http://dx.doi.org/10.1137/S003614450037906X

[6] S. M. Stigler, “Thomas bayes’s bayesian inference,” *Journal of the Royal Statistical Society. Series A (General)*, vol. 145, no. 2, pp. 250–258, 1982. [Online]. Available: http://www.jstor.org/stable/2981538

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

[8] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, “Sense: sensitivity encoding for fast mri,” *Magnetic resonance in medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[9] J. C. Ye, “Compressed sensing mri: a review from signal processing perspective,” *BMC Biomedical Engineering*, vol. 1, pp. 1–17, 2019.

[10] D. L. Donoho *et al.*, “Compressed sensing,” *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[11] S. Liang, T. Dresselaers, K. Louchami, C. Zhu, Y. Liu, and U. Himmelreich, “Comparison of different compressed sensing algorithms for low snr,” *NMR in Biomedicine*, vol. 30, no. 11, 2017, e3776 NBM-16-0236.R1. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nbm.3776

[12] E. Lehmann and G. Casella, *Theory of Point Estimation*, ser. Springer Texts in Statistics. Springer New York, 2003.

[13] H. Van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory.* Wiley, 2004, no. pt. 1. [Online]. Available: https://books.google.com/books?id=Xzp7VkuFqXYC

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

[15] C. D. Richmond and L. L. Horowitz, “Parameter bounds under misspecified models,” in *2013 Asilomar Conference on Signals, Systems and Computers*, Nov 2013, pp. 176–180.

[16] C. D. Richmond, “On constraints in parameter estimation and model misspecification,” in *2018 21st International Conference on Information Fusion (FUSION).* IEEE, 2018, pp. 1080–1085.

[17] M. Pajovic, “Misspecified bayesian cramér-rao bound for sparse bayesian,” in *2018 IEEE Statistical Signal Processing Workshop (SSP)*, June 2018, pp. 263–267.

[18] D. H. Brandwood, “A complex gradient operator and its application in adaptive array theory,” *IEE Proceedings H - Microwaves, Optics and Antennas*, vol. 130, no. 1, pp. 11–16, February 1983.

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

[20] C. D. Richmond and P. Basu, “Bayesian framework and radar: On misspecified bounds and radar-communication cooperation,” in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, June 2016, pp. 1–4.

[21] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001. [Online]. Available: https://doi.org/10.1162/15324430152748236

[22] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, “The variational approximation for bayesian inference,” *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, November 2008.

[23] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in *2011 IEEE International Symposium on Information Theory Proceedings.* IEEE, 2011, pp. 2168–2172.

[24] S. Borman, “The expectation maximization algorithm - a short tutorial,” 2009.

[25] H. Li, Y. Jiang, J. Fang, and M. Rangaswamy, “Adaptive subspace signal detection with uncertain partial prior knowledge,” *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4394–4405, Aug 2017.

[26] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[27] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” *CoRR*, vol. abs/1010.5141, 2010. [Online]. Available: http://arxiv.org/abs/1010.5141

APPENDIX A

DETAILS ON CRAMÉR-RAO BOUND DERIVATION

A.1 ASSUMPTIONS USING CORRECTLY SPECIFIED BOUNDS

Before proceeding with the derivation, a general statement must be said about the two random variables ζ and η. For all the following derivations regarding CRB bounds, ζ can be seen as the estimated variance between the estimation parameter and the mean of estimation (or for some other assumptions the value of the true parameter). While η can be seen as the score function that will be always customized depending on the desired bound. Score function must be chosen wisely to reach the tightest bound of the estimated variance (or MSE in the). In order to do that: score function shall be chosen to be proportional to the values in the other random variable ζ. Furthermore, score function shall depend on sufficient statistics and shall have zero mean with respect to probability density [19].

Giving the above insight about random variables, the following assumptions and constraints must hold to reach CRB bound:

1. $\eta(x,\theta) = \frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial \theta}$ exists and is absolutely integrable.

2. $\frac{\partial^2 \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial \theta^2}$ exists and is absolutely integrable.

3. Regularity condition:

$$\int_{-\infty}^{\infty} \frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)\, dx$$

4. Checking if the chosen score function meet the mentioned two properties:

   (a) Sufficiency: If T(x) is a sufficient statistics for estimating θ, then the density can be factorized such that $\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta) = a(X).b[T(x),\theta]$, Thus:

$$\frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} = \frac{\partial \ln b[T(x),\theta]}{\partial \theta}$$

   for BCRB:

$$\begin{aligned} \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta) &= \pi_\theta(\theta)\, \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta) \\ \rightarrow \frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial \theta} &= \frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} + \frac{\partial \ln \pi_\theta(\theta)}{\partial \theta} \\ &= \frac{\partial \ln b[T(x),\theta]}{\partial \theta} + \frac{\partial \ln \pi_\theta(\theta)}{\partial \theta} \end{aligned}$$

   (b) score function has a zero mean:

$$\begin{aligned} \mathbb{E}_{x|\,\theta}[\eta(x,\theta)] &= \int_{-\infty}^{\infty} \frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta}\, \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\, dx = \int_{-\infty}^{\infty} \frac{\partial\, \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial \theta} \frac{\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)} dx \\ &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)\, dx = \frac{\partial(1)}{\partial \theta} = 0 \end{aligned}$$

for BCRB:

$$\begin{aligned}\mathbb{E}_{x,\theta}[\eta(x,\theta)] &= \int_{\text{-}\infty}^{\infty}\int_{\text{-}\infty}^{\infty}\frac{\partial \ln \mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)}{\partial\,\theta}\,\mathrm{p}_{\mathrm{x},\theta}(\mathrm{x},\theta)\,dx d\,\theta \\ &= \int_{\text{-}\infty}^{\infty}\mathrm{p}_{\mathrm{x}}(\mathrm{x})\,dx\int_{\text{-}\infty}^{\infty}\frac{\partial[\ln \mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})+\ln \mathrm{p}_{\mathrm{x}}(\mathrm{x})]}{\partial\,\theta}\,\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\,d\,\theta \\ &= \int_{\text{-}\infty}^{\infty}\mathrm{p}_{\mathrm{x}}(\mathrm{x})\,dx\int_{\text{-}\infty}^{\infty}\frac{\partial\,\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}{\partial\,\theta}\frac{\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}{\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})}d\,\theta \\ &= \frac{\partial}{\partial\,\theta}\int_{\text{-}\infty}^{\infty}\mathrm{p}_{\mathrm{x}}(\mathrm{x})\,dx\int_{\text{-}\infty}^{\infty}\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x})\,d\,\theta = \frac{\partial(1)}{\partial\,\theta} = 0\end{aligned}$$

5. for BCRB, an extra assumption must be made regarding the parameter θ:

$$\lim_{\theta\to\pm\infty}\theta\,\mathrm{p}_{\theta|\mathrm{x}}(\theta|\mathrm{x}) = 0\ \forall x$$

(The support of θ must be effectively bounded with respect to the a posteriori density).

A.2 ASSUMPTIONS USING MISSPECIFIED BOUNDS

Analogous to what has been done to the correctly specified case, some of assumptions and constraints must hold to reach the desired bounds: Let $X \sim q_{x|\,\theta_T}(x|\,\theta_T)$ where θ_T is fixed and inaccessible; thus, differentiation is restricted to θ in $\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)$.

1. Exploit Kullback-Leibler measure of divergence:

$$\begin{aligned}
&D(\mathrm{p}_{\mathrm{x}|\,\theta_\mathrm{T}}(\mathrm{x}|\,\theta_\mathrm{T})\,//\,\mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)) \triangleq \int_{\text{-}\infty}^{\infty} \mathrm{p}_{\mathrm{x}|\,\theta_\mathrm{T}}(\mathrm{x}|\,\theta_\mathrm{T}) \ln \frac{\mathrm{p}_{\mathrm{x}|\,\theta_\mathrm{T}}(\mathrm{x}|\,\theta_\mathrm{T})}{\mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)} dx \\
&= E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\ln \mathrm{p}_{\mathrm{x}|\,\theta_\mathrm{T}}(\mathrm{x}|\,\theta_\mathrm{T})] - E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)] \\
&\to \frac{\partial}{\partial\,\theta}\, D(\mathrm{p}_{\mathrm{x}|\,\theta_\mathrm{T}}(\mathrm{x}|\,\theta_\mathrm{T})\,//\,\mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)) \\
&= \frac{\partial}{\partial\,\theta}\, E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\ln \mathrm{p}_{\mathrm{x}|\,\theta_\mathrm{T}}(\mathrm{x}|\,\theta_\mathrm{T})] - \frac{\partial}{\partial\,\theta}\, E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)] \\
&= -\frac{\partial}{\partial\,\theta}\, E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)] = -E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}] \\
&\to \eta(X,\theta) = \frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta} + \frac{\partial}{\partial\,\theta}\, D(\mathrm{p}_{\mathrm{x}|\,\theta_\mathrm{T}}(\mathrm{x}|\,\theta_\mathrm{T})\,//\,\mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)) \\
&= \frac{\partial \ln \mathrm{p}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta} + \frac{\partial \mathrm{D}}{\partial\,\theta}
\end{aligned}$$

2. From the previous definitions, the following two facts follow:

$$E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\mu_\mathrm{p}\,\frac{\partial \mathrm{D}}{\partial\,\theta}] = \mu_\mathrm{p}\,\frac{\partial \mathrm{D}}{\partial\,\theta} \qquad\qquad E_{\mathrm{p}(\mathrm{x}|\,\theta_\mathrm{T})}[\frac{\partial \ln \mathrm{q}_{\mathrm{x}|\theta}(\mathrm{x}|\theta)}{\partial\,\theta}] = -\frac{\partial \mathrm{D}}{\partial\,\theta}$$