UNIVERSITY OF CALIFORNIA, SAN DIEGO

Bayesian Sparse Signal Recovery using Scale Mixtures with Applications to Speech

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Control and Robotics)

by

Ritwik Giri

Committee in charge:

Professor Bhaskar D. Rao, Chair
Professor Sanjoy Dasgupta
Professor Kenneth Kreutz Delgado
Professor Lawrence K. Saul
Professor Nuno Vasconcelos

2016

The Dissertation of Ritwik Giri is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Bhaskar D. Rao, for his precious guidance and suggestions over the years and giving me the freedom to pursue the research ideas that interested me. He introduced me to the field of Compressed Sensing and Sparse Signal Recovery and his intuitive understanding of the problem helped me to pursue some useful research directions, which resulted in contribution on both algorithmic and as well as applications side. I would also like to thank the members of my committee, Professors Sanjoy Dasgupta, Kenneth Kreutz-Delgado, Lawrence Saul, and Nuno Vasconcelos. I am especially thankful to Professor Sanjoy Dasgupta for his "Bayesian Learning" class, which introduced me to all things Bayesian, later on which became the core of my dissertation work. I have been fortunate to have awesome labmates at DSP lab (which I have been affiliated with), Bang Nguyen, Elina Nayebi, Soon-En Chiu, Yacong Ding, David Ho and Maher Al-Shoukairi, who have helped me with both their technical and non-technical support. Special thanks to Igor Fedorov, who has not only been an excellent labmate but a good friend, for our collaboration works and also for some really interesting conversations.

I also want to thank my family, specifically my parents, Tapan Kumar Giri and Aparajita Giri, for supporting me all these years. They have always been there for me and motivated me throughout this journey, and without their support this would not have been possible. Finally I want to thank all my friends- Priyanka Barman, Karan Sikka, Varish Diddi, Keyur Karandikar, Shailendra Singh, Sumit Dhoble, Kumar TL, Rakesh Varna and others. This journey would not have been that fun and smooth without support from my friends. Special thanks to Karan for all the helpful advice and suggestions in both professional and personal life, and for being a really good friend for all these years.

Much of the material in this dissertation has been published, submitted for publication, or is in preparation for publication. Chapter 2 is, in part, a reprint of material

published in two articles: ”Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures”, IEEE Trans. Signal Processing (2016); ”Learning distributional parameters for Adaptive Bayesian Sparse signal recovery”, IEEE Computational Intelligence Magazine (2016). In all cases I was the primary author and B.D. Rao supervised the research.

Chapter 3 is, in part, a reprint of material published in two articles, ”Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures”, IEEE Trans. Signal Processing (2016) and, ”Hierarchical Bayesian Formulation of Sparse Signal Recovery Algorithms using Scale Mixture Priors”, Asilomar Conference on Signals, Systems and Computers, 2015. I was the primary author and B.D. Rao supervised the research.

Chapter 4 is, in full, is based on the material as it appears in, ”Learning distributional parameters for Adaptive Bayesian Sparse signal recovery”, IEEE Computational Intelligence Magazine (2016). I was the primary author and B.D. Rao supervised the research.

Chapter 5 is, in part, a reprint of material is in preparation for submission under the title, ”Empirical Bayes based Relative/ Room Impulse Response Estimation” and also based on the material as it appears in, ”Dynamic Relative Impulse Response Estimation using Structured Sparse Bayesian Learning”, 41st IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2016. In both cases, I was the primary author, Fred Mustiere contributed to the research, and Tao Zhang and B.D. Rao supervised the research.

The text of Chapter 6, in full, is based on the material as it appears in: ”Reweighted Algorithms for Independent Vector Analysis (IVA)”, submitted to IEEE Signal Processing Letters. I was the primary author, while B.D. Rao and H. Garudadri supervised the research.

Chapter 7, in full, is a reprint of material as in ”Multivariate Scale Mixtures for Joint Sparse Regularization in Multi-Task Learning”, submitted to IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2017. I was the primary author, and B.D. Rao supervised the research.

Finally, the material in Section 8 is, in full, is a reprint of material published as ”Block Sparse excitation based All-Pole modeling with applications to Speech”, in 39th IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2014. I was the primary author and B.D. Rao supervised the research.

VITA

1990	Born, Kolkata, India
2011	Bachelor of Engineering, Jadavpur University, India
2013	Master of Science, University of California, San Diego
2016	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

- **Ritwik Giri**, Bhaskar D. Rao, *Block Sparse excitation based All-Pole modeling with applications to Speech*, 39th IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2014.
- **Ritwik Giri**, Heesook Choi, Kevin Soo Hoo, Bhaskar D. Rao, *User behavior modeling in a cellular network using MapReduced Latent Dirichlet Allocation*, 15th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL) 2014.
- **Ritwik Giri**, Bhaskar D. Rao, *Bootstrapped Sparse Bayesian Learning for Sparse Signal Recovery*, 48th Annual Asilomar Conference on Signals, Systems and Computers, 2014.
- **Ritwik Giri**, Michael L. Seltzer, Dong Yu, Jasha M. Droppo, *Improving Speech Recognition in Reverberation using a Room-aware Deep Neural Network and Multi-Task Learning*, 40th IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.
- **Ritwik Giri**, Bhaskar D. Rao, *Hierarchical Bayesian Formulation of Sparse Signal Recovery Algorithms using Scale Mixture Priors*, 49th Annual Asilomar Conference on Signals, Systems and Computers, 2015 (Invited paper).
- **Ritwik Giri**, Bhaskar D. Rao, *Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures* IEEE Transactions on Signal Processing (2015).
- Karan Sikka, **Ritwik Giri**, Marian Bartlett, *Joint Clustering and Classification for Multiple Instance Learning*, 26th British Machine Vision Conference (BMVC), 2015.
- **Ritwik Giri**, Bhaskar D. Rao, Fred Mustiere, Tao Zhang, *Dynamic Relative Impulse Response Estimation using Structured Sparse Bayesian Learning*, , 41st IEEE

International conference on Acoustics, Speech and Signal Processing (ICASSP) 2016.

- **Ritwik Giri**, Bhaskar D. Rao, *Learning distributional parameters for Adaptive Bayesian Sparse signal recovery*, IEEE Computational Intelligence Magazine Special Issue on "Model Complexity, Regularization and Sparsity", 2016.
- Igor Fedorov, Alican Nalci, **Ritwik Giri**, Bhaskar D. Rao, *A Unified Bayesian Framework for Sparse Non-negative Matrix Factorization (S-NMF)*, IEEE Transactions on Signal Processing, 2016 (under review).
- Igor Fedorov, **Ritwik Giri**, Bhaskar D. Rao, Truong Q. Nguyen, *Robust Bayesian Simultaneous Block Sparse Signal Recovery with Applications to Face Recognition*, IEEE International Conference of Image Processing (ICIP) 2016.
- **Ritwik Giri**, Bhaskar D. Rao, *Reweighted Algorithms for Independent Vector Analysis (IVA)*, IEEE Signal Processing Letters, 2016 (under review).
- **Ritwik Giri**, Tao Zhang, *Bayesian Blind Deconvolution with Application to Acoustic Feedback Path Modeling*, IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 (Under review).
- **Ritwik Giri**, Bhaskar D. Rao, *Multivariate Scale Mixtures for Joint Sparse Regularization in Multi-Task Learning*, IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 (Under review).
- **Ritwik Giri**, Bhaskar D. Rao, Fred Mustiere, Tao Zhang, *Empirical Bayes based Relative/ Room Impulse Response Estimation* (to be submitted).
- United States Patent Application (Filed): *Dynamic Relative Transfer Function Estimation using Structured Sparse Bayesian Learning*, **Ritwik Giri**, Fred Mustiere, Tao Zhang, 2016.

FIELDS OF STUDY

Major Field: Electrical Engineering

Studies in Statistical Signal Processing, Audio Signal Processing

ABSTRACT OF THE DISSERTATION

Bayesian Sparse Signal Recovery using Scale Mixtures with Applications to Speech

by

Ritwik Giri

Doctor of Philosophy in Electrical Engineering (Intelligent Systems, Control and Robotics)

University of California, San Diego, 2016

Professor Bhaskar D. Rao, Chair

Sparse Signal Recovery (SSR) problem has received a lot of interest in recent times because of its significant impact on many engineering applications. This thesis tackles this important problem in a Bayesian framework and discusses a generalized scale mixture distribution family, Power Exponential Scale Mixture (PESM) and analyzes its usefulness as a candidate for the sparsity promoting prior distribution. We derive a unified MAP estimation or Type I framework for SSR by employing an appropriate member of the PESM family, Generalized t distribution (GT) and show that the unified framework encompasses several popular regularization based SSR algorithms such as the reweighted

ℓ_1 and reweighted ℓ_2 algorithms among others for specific distributional parameters. We also propose an adaptive framework of learning the distributional parameters of GT over the iterations based on the measurements, instead of fixing them beforehand. In addition to that, exploiting the natural hierarchical framework induced by the PESM family, we utilize these priors in a Type II/ Empirical Bayes framework and develop corresponding EM based SSR algorithms. Multivariate extension of our proposed PESM family has also been discussed, which in turn resulted in a unified framework for imposing joint sparsity in a Multi Task Learning (MTL) framework.

We have also shown three specific applications of SSR in audio signal processing, which includes problem specific algorithm enhancements but still utilizes the basic understanding of SSR. For example, by employing a source prior from the M-PESM family in a joint blind source separation problem, we propose a class of reweighted algorithms for Independent Vector Analysis (IVA) with the ability to exploit any intra-source correlation structure. An Empirical bayes based Impulse Response (IR) estimator has also been proposed, which exploits both sparse early reflections and exponential decay reverb tail structure in Room Impulse Response/ Relative Impulse Response as prior information. Sparsity in residual has also been exploited for a speech modeling application, which uses the prior block sparse structure of glottal excitation to find the all pole filter coefficients to model speech efficiently.

Chapter 1

Introduction

Sparse Signal Recovery (SSR), i.e. finding sparse signal representations from overcomplete dictionaries, has become a very active research area in recent times because of its wide range of engineering applications and interesting theoretical nature [26, 46, 53, 55, 77, 144]. For example, in several popular computer vision problems, such as face recognition [171], motion segmentation [56], and activity recognition [173], signals lie in low-dimensional subspaces of a high dimensional ambient space. An important class of methods to deal with this depends on exploiting the notion of sparsity. Following this path, Sparse Representation based Classification (SRC) [171] was proposed, which produced state of the art results in a face recognition task. Sparse coding, which is essentially a variant of SSR, has also been widely used as a promising tool in several image processing based applications with great success [54, 100, 119]. In this thesis, we discuss the sparse recovery problem from a Bayesian perspective and unifies many popular recovery algorithms in a general framework and propose novel extensions of the well established algorithms for different relevant applications specifically in audio signal processing.

1.1 Problem Formulation

The problem of SSR can be formulated as a problem of finding a sparse solution to an underdetermined system of equations $\mathbf{y} = \Phi\mathbf{x}$, which is a linear forward generative model, where $\Phi = [\phi_1, .., \phi_M]$ is an $N \times M$ matrix with $N < M$, and it is assumed that $\text{Spark}(\Phi)$[1] $= N + 1$ [53]. The columns ϕ_i of Φ are often formed from a physically meaningful model and the elements of the vector $\mathbf{x}$ are generally non-zero parameters of interest which are to be identified and the vector $\mathbf{y}$ is the $N \times 1$ measurement vector. The goal is to solve for $\mathbf{x}$, an $M \times 1$ vector, with the requirement that the solution vector

[1] Spark: Given a matrix A we define σ= Spark(A) as the smallest possible number such that there exists a subgroup of σ columns from A that are linearly dependent. .

$\mathbf{x}$ be sparse, i.e. many of its entries will be zero. Finding a sparse solution involves determining the number of columns K (the sparsity index), and the set of column vectors $\{\phi_{k_i}\}_{i=1}^{K}$ that best model $\mathbf{y}$. Even though the forward model is linear, the goal of enforcing sparsity will make the ensuing inverse problem from $\mathbf{y}$ to $\mathbf{x}$ highly nonlinear. Ideally one can recover the optimal sparsest solution $\mathbf{x}_0$ by solving the following ℓ_0 optimization problem [53],

$$\min_x ||\mathbf{x}||_0 \text{ such that } \mathbf{y} = \Phi\mathbf{x}, \tag{1.1}$$

where $\|\mathbf{x}\|_0$ is a measure of the support of $\mathbf{x}$. In practice, measurements are generally corrupted by noise, which motivates the following modified optimization problem,

$$\min_x ||\mathbf{y} - \Phi\mathbf{x}||_2^2 + \lambda||\mathbf{x}||_0, \tag{1.2}$$

where $\lambda > 0$ is related to the measurement noise variance. It can be shown that in the limit as $\lambda \to 0$, the above two problems are equivalent [53].

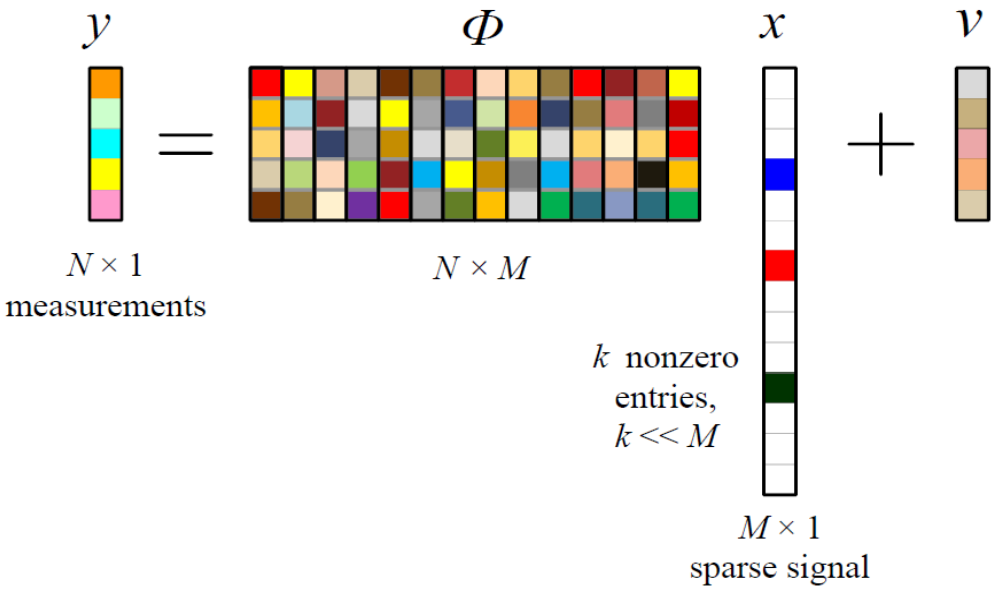

Figure 1.1. Single Measurement Vector (SMV) Model

This problem is known as Single Measurement Vector (SMV) recovery problem, since we are dealing with only measurement. Another extension of this model is the Multiple Measurement Vector (MMV) problem, where multiple measurements are considered simultaneously for recovery. Details of the MMV problem will be discussed later.

1.2 Algorithms

Since, the above optimization problem is not convex and is known to be NP-hard [129], one popular family of algorithms are based on approximating the original penalty factor $||\mathbf{x}||_0$ by a suitable surrogate $g(\mathbf{x})$ leading to the optimization problem,

$$\min_x ||\mathbf{y} - \Phi\mathbf{x}||_2^2 + \lambda g(\mathbf{x}). \tag{1.3}$$

Different choices of the penalty factor $g(\mathbf{x})$, also referred to here as diversity measure, lead to different SSR algorithms [30, 33, 41, 48]. It has been shown that the choice of a strictly increasing, concave penalty factor on the positive orthant, leads to an objective function with local minima being sparse and the sparsest solution is the global minimum under some conditions [145, 169]. Majorization-Minimization [60] can be employed to solve this optimization problem for such penalty functions, and this has led to the development of useful reweighted norm minimization algorithms. Minimizing diversity measures $g(\mathbf{x})$ to recover the sparse representations has been a popular algorithm exploration avenue. In this framework, the SSR problem formulation can also be viewed as a regularization approach to signal reconstruction. A popular approach among this class is the ℓ_p norm minimization based methods. $p = 1$ leads to a tractable and computationally attractive convex optimization problem and the very well known approaches such as Basis Pursuit, LASSO are based on the ℓ_1 framework [47, 156]. Other than the convexity property, ℓ_1

based approaches have been supported by theoretical guarantees of exact recovery given some conditions on the overcomplete dictionary [48], which makes these approaches attractive options. The recently proposed reweighted ℓ_1 and ℓ_2 norm minimization approaches [30, 33, 146] have empirically shown superior recovery performance over ℓ_1 minimization and are considered in this thesis. A generalized framework for ℓ_p regularized unconstrained minimization problems using iterative reweighted algorithms have also been studied in [115].

Greedy algorithms such as, Matching Pursuit (MP) [38], Orthogonal Matching Pursuit (OMP) [139] have also gained lot of interest because of their low computational complexity. However, their recovery performance is strongly affected by the coherence among columns of the dictionary, and also do not have satisfactory performance in noisy scenarios. Recently message passing algorithms [17, 50] have also become very popular because of some very fast implementations which make them really useful for high dimensional applications of sparse recovery.

1.2.1 Bayesian Methods

In addition to the regularization framework, another options for SSR algorithm development is the Bayesian framework [13, 16, 71, 82, 88, 158, 159]. In a Bayesian framework, the sparsity constraint is incorporated by choosing a suitable sparse prior on the coefficient vector $\mathbf{x}$. In a Bayesian setting, there are two popular avenues for algorithm development: a Type I MAP based approach, and a Type II Evidence Maximization approach involving a Hierarchical model. Most of the approaches discussed above, based on (1.3), can be interpreted and cast in a suitable Type I framework. A Type II framework has been considered in [88, 157], where a Relevance Vector Machine is adapted to the problem at hand. In [164, 166, 172] a Type II optimization problem has been transformed into a Type I problem by employing a suitable penalty function and

reweighted norm minimization algorithm is developed to solve the resulting optimization problem. Following the Type II framework, a Laplacian prior which corresponds to ℓ_1 norm minimization can also be represented in a Hierarchy using a Gaussian Scale Mixture (GSM) representation [13, 59]. In the statistics community, the well known Bayesian Lasso [137] also makes use of the equivalence of a hierarchical Gaussian-Exponential prior to the Laplace prior, and conducts a fully Bayesian inference (via Markov chain Monte Carlo or MCMC sampling algorithms). Demi-Bayesian Lasso [15] solves the same problem using a Type II approach. It has been shown empirically that a Type II methods performs consistently better than Type I, i.e the MAP estimation approach, and theoretical analysis in support for this superiority has recently begun to appear. However, much remains to be done and this work is an attempt in this direction. In [118], the two different frameworks are analyzed in a generalized Hierarchical Bayesian setting which motivates us to analyze these two frameworks for the specific SSR problem to gather additional insights by exploiting domain knowledge.

1.2.2 Choice of Prior

In a Bayesian framework, the sparsity constraint is incorporated by choosing a suitable sparsity inducing prior on the coefficient vector $\mathbf{x}$. Most of the approaches discussed above, based on (1.3), can be formulated in a Bayesian Maximum A Posteriori (MAP) estimation framework with a specific choice of sparsity inducing prior distribution, $p_X(\mathbf{x}) \propto \exp(-g(\mathbf{x}))$ with assumption of Gaussian measurement noise with variance λ. This leads to the discussion as to what classes of distributions are suitable for sparsity promoting prior, and is there any specific property they need to satisfy which will make them suitable to promote sparsity. Since Gaussian density can not be used to represent sparse priors, we need densities that are more peaked, with heavier tails than Gaussian, i.e. it is close to zero most of the time, but occasionally takes relatively large values.

This class of distributions is known as super Gaussian distributions [134]. The most commonly used definition of sub and super Gaussianity involves the sign of the fourth cumulant, i.e. the kurtosis relative to a Gaussian of equal variance. If the kurtosis exceeds that of the Gaussian, then a random variable X, or its density $f_X(x)$, is said to be super Gaussian. Likewise if the excess kurtosis is negative, then X is said to be sub Gaussian. Recently Gaussian Scale Mixtures (GSM) [14, 101, 113, 132, 157, 165] and Laplacian Scale Mixtures (LSM) [2, 65] have been proposed as a suitable class of distributions to promote sparsity. In this thesis, we introduced a more general scale mixture framework, the Power Exponential Scale Mixture (PESM) family, for SSR algorithm development and we provide new theoretical insights and enhanced algorithms. In [132], Type I and Type II frameworks for SSR were introduced using two forms of density representation, a convex representation and a GSM representation, to provide an unified treatment. We build on this work and employ a generalized scale mixture representation to establish connections and develop enhancements to popular SSR algorithms, as well as treat both ℓ_1 and ℓ_2 variants in an unified manner.

1.3 Extension: MMV Problem and Joint Sparsity

Though the single measurement problem was in the forefront of the initial research activities in the sparse recovery/ compressed sensing field, recently Multiple Measurement Vector (MMV) problem has gained a lot of interest because of its relevance in different applications [39]. In many engineering applications often multiple measurements are available to solve the recovery problem. The SMV problem can easily extended for to the following MMV model,

$$\mathbf{Y} = \Phi\mathbf{X} + \mathbf{V} \tag{1.4}$$

where $\mathbf{Y} = [\mathbf{Y}_{:,1},,\mathbf{Y}_{:,L}] \in \mathbb{R}^{N\times L}$ is constructed using L measurement vectors, $\mathbf{X} = [\mathbf{X}_{:,1},,\mathbf{X}_{:,L}] \in \mathbb{R}^{M\times L}$ is the desired solution matrix, and $\mathbf{V}$ is an unknown noise matrix. A key assumption in the MMV model is that the sparsity profile of every desired coefficient vector, i.e., every column in $\mathbf{X}$ is same hence they have identical support. Since $\mathbf{X}$ has a large number of rows which are completely zero, the notion of joint sparsity is introduced, which essentially means that the columns of the desired coefficient matrix $\mathbf{X}$ are jointly sparse. It has been shown that compared to the SMV case, the successful recovery rate of the support can be greatly improved using multiple measurement vectors [90].

Most of the algorithms discussed for the SMV can be extended in a straight forward manner to solve the MMV problem. From a Bayesian perspective to enforce joint sparsity, a multivariate super Gaussian distribution can be employed as the sparsity promoting prior over each row of $\mathbf{X}$ [168, 179]. Following this approach, in this thesis we also introduced the multivariate extension of generalized scale mixture family, namely Multivariate Power Exponential Scale Mixture (M-PESM) and provided unified inference framework which encompasses well know MMV recovery algorithms. Since in real life applications there could be correlation structure present among the entries in each nonzero row of $\mathbf{X}$, our unified MMV recovery framework also has the ability to exploit any present correlation structure, which often leads to superior recovery performance.

1.4 Applications

Sparse signal recovery has been successfully deployed in a variety of engineering applications including,

- Signal Representation [49, 123]
- EEG/MEG source localization [76, 81, 163]
- Bandlimited extrapolations and spectral estimation [25, 51]

- Array Signal Processing [122, 151]
- Speech Coding [40, 67]
- Sparse Channel Equalization [20, 38]
- Compressive Sampling [28, 29]
- Magnetic Resonance Imaging [62, 116]
- Financial Data analysis [35, 58]
- Audio signal processing [68, 73]

and many more.

In this thesis, we focus on specific applications in audio signal processing/ source separation. Specifically in Chapter 5 we propose a novel empirical bayes based room/ relative impulse response estimator which exploits the sparsity notion and show the efficacy of our proposed estimator in an adaptive echo cancellation task and in blocking matrix construction task for adaptive beamformer Generalized Sidelobe Canceler (GSC). In Chapter 8 we use the notion of block sparsity in a speech modeling problem, and show its efficacy over traditional and well known LPC model. Finally in Chapter 6 we show the application of joint sparsity in Independent Vector Analysis (IVA) framework for joint blind source separation problem by employing our proposed multivariate generalized scale mixture: M-PESM as a source prior.

1.5 Contributions and Organization

- In Chapter 2, one of the major contributions of this work, the introduction of a more general Scale Mixture framework, the Power Exponential Scale Mixture (PESM) family, for SSR algorithm development has been discussed. The PESM

representation includes the popular GSM and LSM as special cases and provides a mechanism to provide a unified view of the popular ℓ_1 and ℓ_2 frameworks currently employed. We also establish the conditions under which a distribution which is symmetric with respect to the origin will have a PESM representation, which generalizes the result known for GSM. This work will emphasize the generalized t (GT) distribution family of priors, a member of PESM, since it has a wide range of tail shapes, and also includes the heavy tailed super gaussian distributions. GT family of distributions have been mentioned in statistics literatures for design of robust regressors for several financial modeling tasks, where the heavy tail nature of GT helps to model the outliers [27, 127]. In this work we show when a GT distribution will be suitable to promote sparsity, i.e. for what values of the distributional parameters, member of a GT family will be a super Gaussian distribution.

- In Chapter 3, we summarize two types of Bayesian frameworks, i.e. Type I and Type II for SSR in detail, along with providing connections to traditional norm minimization approaches by suitable choice of sparse prior distributions. Of particular importance is the treatment of the diversity measure used in connection with the reweighted ℓ_1 algorithm as well as an unified treatment of both ℓ_1 and ℓ_2 based approaches. We formulate and unify three well known diversity minimization based SSR algorithms in the PESM framework and derive the Type I and Type II versions of them. Of particular interest is the Type II counterpart of the reweighted ℓ_1 algorithm [30]. We also analyze the difference between Type I and Type II inference procedures and our analysis shows the fundamental difference between these two frameworks and also helps to understand a potential reason for the empirical superiority of Type II methods over Type I.

- In Chapter 4, we provide an alternative derivation the unified MAP estimation framework for SSR, limited only to the GT distribution family, and propose an adaptive paradigm, where the distributional parameters of a GT member is not fixed beforehand, and they are adapted over iterations based on the measurements in a nested gradient descent based algorithm. Our proposed approach shows merit over the other competing approaches with fixed distributional parameters.

- As an application in audio signal processing, in Chapter 5, we propose an Empirical Bayes based estimation approach for Room Impulse Response (RIR) and Relative Impulse Response (ReIR), namely Structured Sparse Bayesian Learning (S-SBL), where the regularization has been incorporated by exploiting the prior knowledge of the system. Similarity in the structure of both RIRs and ReIRs enable us to use our proposed approach for both the Echo cancellation and Blocking Matrix construction tasks. Specifically, unified treatment of sparse early reflection and exponentially decaying reverberation tail in a prior distribution using an Empirical Bayesian framework is the main novelty of our work. Our approach also models any ambient measurement noise and leads to a much more robust estimator of the IR. We also study the Mean Squared Error properties of our estimator, and show that under some conditions our proposed estimator is actually minimizing a weighted MSE.

- As an application, in Chapter 6, we propose a multivariate extension of PESM, namely M-PESM as the source prior for IVA in a Joint Blind Source Separation (JBSS) task. This class of distributions also helps us to exploit both the higher order (greater than second order) dependencies within a SCV and also any intra-source correlation (second order dependency), present across the datasets. We also show that two popular variants of IVA in literature, are special cases of this

unified framework. By employing a specific member (Multivariate Generalized t distribution) of M-PESM as the source prior, our unified framework leads to two novel Reweighted algorithms for IVA.

- In Chapter 7, we discuss M-PESM in details and show its usefulness in promoting joint sparsity in a Multi Task Learning framework. Though in [179], authors have considered incorporating correlation structure in a MMV problem in a Type II setting, we provide the similar option of exploiting the correlation structure in a unified Type I framework, by employing M-PESM as joint sparsity inducing prior.

- Finally for another application, in Chapter 8, we use the notion of block sparsity in residual and show how by modeling the glottal excitation as block sparse, we can model speech better than the traditional LPC approach.

Chapter 2

Power Exponential Scale Mixtures (PESM)

2.1 Introduction

Scale mixture distributions namely GSM and LSM have gained a lot of attention in recent years because of their ability to represent complex heavy tailed super Gaussian distributions in a simple hierarchical manner [65, 101, 113, 132, 168]. In the statistics community, robustness has been the major reason for the use of scale mixtures. In regression analysis, the method of least squares often fails because of the outliers in the data, which motivates the use of heavy tailed distributions to model the outliers. Their heavytailed nature also makes them suitable to promote sparsity in recovery problem. As one of the main contributions in this thesis, we introduce a generalized scale mixture namely, PESM for Sparse Signal Recovery (SSR). In this section we analyze PESM in details and identify conditions under which a symmetric distribution can be represented as a PESM. We also discuss Generalized t (GT) distribution family in details, which is a member of PESM, and show when a member of GT family is suitable to be a sparsity inducing prior distribution in Bayesian Sparse Signal Recovery algorithms.

2.2 PESM: When and Why?

Power Exponential (PE) distributions were first introduced by Box and Tiao (1962) in the context of robust regression to deal with non-normality. PE distribution is symmetric about the origin and a zero mean PE distribution has the following parameterized form:

$$f_{PE}(x;p,\gamma) = \frac{p\, e^{-\frac{|x|^p}{\gamma}}}{2\gamma^{1/p}\Gamma(\frac{1}{p})} \quad \text{where, } p,\gamma > 0. \tag{2.1}$$

Where, p is known as the shape parameter and γ is knows as the scale parameter. It is evident from the above given form, that $p = 2$ results in the normal distribution, whereas $p = 1$ connects to the well-known double exponential or Laplacian distribution.

$p < 2$ leads to distribution with heavier tails than the Gaussian distribution.

PESM family of distributions refers to distributions that can be represented in a hierarchy using scale mixture of PE distribution.

$$p_X(x) = \int f_{PE}(x; p, \gamma)\alpha(\gamma)d\gamma \tag{2.2}$$

where, $\alpha(.)$ is a density function on $\mathbb{R}$ and p is a positive shape parameter. Now, in a bayesian framework sparsity on coefficient vector is imposed by choosing a sparse i.e, a supergaussian (heavy tailed) prior on x. Choice of distributional parameter p along with different suitable mixing densities, i.e. $\alpha(\gamma)$, will lead to different distributions including the super Gaussian distributions. Because of the scale mixture representation, the generation of the random variable X can be viewed in a hierarchy, i.e. generate γ using $\alpha(\gamma)$ followed by generating X using $f_{PE}(x; p, \gamma)$. As special cases, the choice of $p = 2$ leads to GSM which has been very popular in the literature, and $p = 1$ leads to the LSM. Interestingly, a Laplacian distribution $p_X(x) = \frac{a}{2}e^{-a|x|}$ can be represented as a GSM with exponential mixing density, i.e. $\alpha(\gamma) = \frac{a^2}{2}\exp(-\frac{a^2}{2}\gamma)u(\gamma)$, where $u(.)$ is the unit step function [13]. This means, any LSM can also be represented as a GSM with an extra layer of hierarchy.

More explicitly,

$$\begin{aligned} p_X(x) &= \int_0^\infty p(x|\gamma)\alpha(\gamma)d\gamma \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\gamma}}\exp(-\frac{x^2}{2\gamma}) \times \frac{a^2}{2}\exp(-\frac{a^2}{2}\gamma)d\gamma \\ &= \frac{a}{2}e^{-a|x|} \end{aligned} \tag{2.3}$$

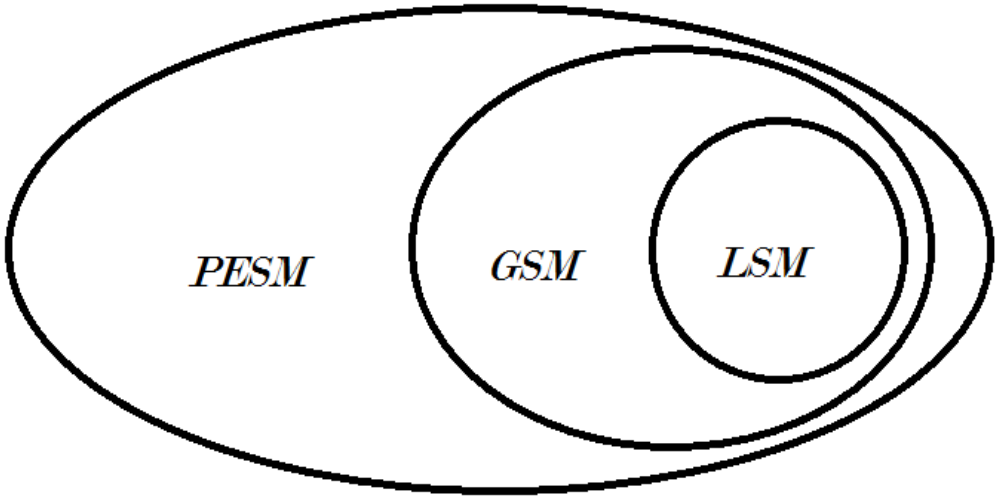

Figure 2.1. PESM: Generalized scale mixtures

2.2.1 When PESM?

Here, we will answer the question as to when a distribution can be represented as a PESM. The derivation uses the result dealing with integral representations, discussed in [132, 162]. We summarize our main result in the following theorem,

Theorem 2.2.1 *A distribution $p_X(x)$, which is symmetric around the origin has a PESM representation with shape parameter p if and only if $p_X(x^{1/p})$ is completely monotonic on $(0,\infty)$.*

Proof: Before proceeding with the proof of the theorem, recall the definition of a completely monotone function [22],

Definition 2.2.2 *A function $f(x)$ is completely monotone on $(0,\infty)$ if, f is infinitely differentiable and $(-1)^n f^{(n)}(x) \geq 0, \;\; n = 0,1,...$ for every $x \in (0,\infty)$ where, $f^{(n)}(x)$ denotes the n^{th} order derivative of f.*

Now to prove the first part of Theorem 2.2.1, let's assume that X is a random variable with a distribution $p_X(x)$ which has a PESM representation. Hence,

$$p_X(x) = \int_0^\infty PE(x;p,\gamma)d\alpha(\gamma) = \int_0^\infty \frac{p\, e^{-\frac{|x|^p}{\gamma}}}{2\gamma^{1/p}\Gamma(\frac{1}{p})}d\alpha(\gamma) \tag{2.4}$$

where, $\alpha(\gamma)$ could be interpreted as the cumulative distribution function of the scale mixing density. Let,

$$g(x) = p_X(x^{1/p}) = \int_0^\infty \frac{p\, e^{-\frac{x}{\gamma}}}{2\gamma^{1/p}\Gamma(\frac{1}{p})}d\alpha(\gamma) \quad \text{for, } 0 \le x < \infty. \tag{2.5}$$

Hence from the definition of completely monotone, it is straightforward to see that $g(x)$ is completely monotone on $(0,\infty)$, since its derivatives have alternating signs.

Conversely, suppose $g(x)$ is completely monotone on $(0,\infty)$. From Bernstein's theorem [22, 162], we can write,

$$g(x) = \int_0^\infty e^{-\frac{x}{\gamma}}d\alpha(\gamma) \tag{2.6}$$

for some non decreasing $\alpha(\gamma)$ on $(0,\infty)$. Hence, we get a PESM representation,

$$p_X(x) = g(x^p) = \int_0^\infty e^{-\frac{x^p}{\gamma}}d\alpha(\gamma). \tag{2.7}$$

This completes our proof. □

It is interesting to note that this result is also consistent with the result established for GSM, i.e. with shape parameter of PE, $p = 2$ in [92, 132, 135]. It also provides a new relevant result, that is under what condition a distribution can be represented as an LSM, i.e. PESM with $p = 1$.

2.2.2 Why PESM?

Among Scale Mixtures, GSM in particular has gained a lot of interest over the years in the literature and the proposed PESM framework is an interesting generalization for SSR purposes. It has been shown in [135], that GSM representations always lead to a heavytailed supergaussian distribution and the following lemma was established.

Lemma 2.2.3 *A symmetric pdf $p(x)$ is strongly super-gaussian if $-\log(p(\sqrt{x}))$ is concave on $(0,\infty)$ and strongly sub-gaussian if $-\log(p(\sqrt{x}))$ is convex on $(0,\infty)$.*

Unlike GSM, PESM representation can also be used for subgaussian densities along with supergaussian densities. For example consider a density $p(x) \propto \exp(-x^3)$ (ignoring the normalization constant). It is evident that, this is a subgaussian distribution with negative kurtosis and does not have a GSM representation since it does not satisfy Theorem 2.2.1 for shape parameter $p = 2$. We can still represent $p(x)$ as a PESM with the shape parameter $p = 6$. To verify this following the previous theorem we just need to show that $p(x^{1/6}) = \exp(-g(x))$ is completely monotonic on $(0,\infty)$, where $g(x) = \sqrt{x}$. Now,

$$p'(x^{1/6}) = -g'(x)\exp(-g(x)) \tag{2.8}$$

Where, $g'(x) = \frac{1}{2\sqrt{x}} > 0$. Hence $p'(x^{1/6}) < 0$. Following the same route,

$$p''(x^{1/6}) = \exp(-g(x))(g'(x)^2 - g''(x)) \tag{2.9}$$

Where, $g''(x) = -\frac{1}{4x\sqrt{x}} < 0$. Hence $p''(x^{1/6}) > 0$.

It is evident that $p(x^{1/6})$ has derivatives of alternating signs and satisfies Definition 2.2.2. This proves that $p(x)$ can be represented as a PESM but not as a GSM.

Moreover, for the purposes of the SSR work, the general PESM allows one to treat both the LSM (p=1) and GSM (p=2) in a unified manner thereby enabling treatment of ℓ_1 and ℓ_2 based algorithms in a unified manner.

2.2.3 Example of PESM: Generalized t-distribution

In [72], we have shown that the Inverse Gamma (IG) distribution as the scale mixing density $\alpha(\gamma)$ in the scale mixture representation (2.2) for the PESM family leads to a GT distribution, which is a superset of many of the super Gaussian distributions that have been used in practice in several recent works, e.g. Generalized Double Pareto (GDP) [10], Laplacian and Student's t-distribution, among others.

The GT Distribution has the form:

$$GT(x;p,q) = \frac{\eta}{(1+\frac{|x|^p}{q})^{q+\frac{1}{p}}} \tag{2.10}$$

where η is the normalization constant, p and q are the positive distributional parameters. Distributional parameters, p and q can be used to represent different tail behavior using GT distribution. Larger values of p and q correspond to thin tailed distributions whereas smaller values of p and q are associated with heavy tailed distributions [127], suitable for sparse recovery task.

As mentioned above, the GT distribution family can be represented in PESM framework using $\alpha(\gamma) = IG(\gamma;q,q)$ where,

$$IG(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\big(-\frac{\beta}{x}\big) u(x). \tag{2.11}$$

Interesting special case of note is $p = 2$, which leads to a student's t-distribution, a prior that has been used in the popular Sparse Bayesian Learning /Relevance Vector Machine [157] work and can be decomposed as a GSM with Inverse Gamma as the mixing

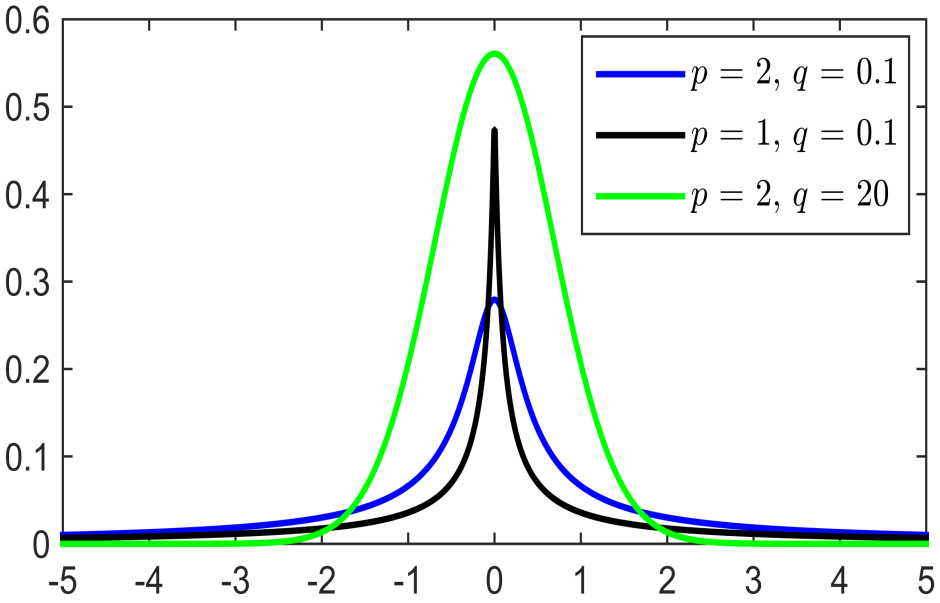

Figure 2.2. Tail behavior of GT distribution for different values of p and q

density. Employing $p = 1$ leads to GDP discussed in [10] which can be represented as a scale mixture of Laplacian following Equation (2.2).

Table 2.1. Variants of Generalized t Distribution

q	p	Distribution
$q \to \infty$	2	Normal
$q \to \infty$	1	Laplacian (Double Exponential)
$q \geq 0$ (degrees of freedom)	2	Student t distribution
$q \geq 0$ (shape parameter)	1	Generalized Double Pareto (GDP)

In this work we will study how the tail nature of GT distributions are controlled by their respective distributional parameters. Kurtosis has been a popular choice of statisticians to analyze the tail behavior of a distribution, which is a function of the fourth

moment [147]. The even moments of the GT family can be computed following [93],

$$E_{GT}[x^r] = \frac{q^{\frac{r}{p}} B(\frac{r+1}{p}, q - \frac{r}{p})}{B(\frac{1}{p}, q)} \tag{2.12}$$

where, $B(.)$ denotes the beta function. From the above moment equation one can deduce that the product of the distributional parameters of GT has to be greater than 4 for its kurtosis to be defined, i.e. $pq > 4$. This often becomes a limiting condition as for most of the heavytailed super Gaussian distributions, kurtosis is not defined. In that case Lemma 2.2.3 provides a feasible option to check the tail nature of a distribution. Here we are interested in knowing for what choice of distributional parameters GT family will represent heavy tailed distributions, i.e. super Gaussian densities. Verifying Lemma 2.2.3 for GT family reveals the following result.

Theorem 2.2.4 *$p_X(x)$, a member of GT family will be a strongly super Gaussian density when its distributional parameters q is bounded and $p \leq 2$.*

Proof: To show that for which values of distributional parameters p and q, a GT distribution will be a strongly super Gaussian distribution, i.e. suitable to represent sparsity inducing prior, we will verify the Lemma 2.2.3 on super Gaussianity.

Lets assume, X is a random variable with distribution $p_X(x)$ from GT family, i.e.

$$p_X(x) = GT(x; p, q) = \frac{\eta}{(1 + \frac{|x|^p}{q})^{q+\frac{1}{p}}}. \tag{2.13}$$

To verify Lemma 2.2.3, we need to check for what values of p and q, $f(x) = -\log p_X(\sqrt{x})$ will be strictly concave on $(0, \infty)$. We will verify the second order condition for concavity

of $f(x)$,

$$f''(x) = \frac{pq+1}{2(q+x^{p/2})^2} x^{p/2-2}(\frac{pq}{2} - q - x^{p/2}). \tag{2.14}$$

For strict concavity on $(0,\infty)$, we need to satisfy $f''(x) < 0$, which leads to, $(\frac{p}{2}-1)q \leq 0$ when q is bounded. Hence, $p_X(x)$, a distribution of GT family, is strongly super Gaussian when, $p \leq 2$ and bounded q. This completes our proof. □

As discussed above, many of the super Gaussian distributions that have been employed to promote sparsity in literatures fall under the above discussed GT family. In Table 2.1, we summarize some special cases that have been used for SSR that arise by different choices of the shape parameters of GT, i.e. p and q along with the resultant popular SSR algorithms. Interestingly when $p = 2$, GT distribution represents a well-known student's t-distribution and q could be interpreted as the degrees of freedom. With the choice of distributional parameters, $p = 2$ and $q \to \infty$, student's t-distribution becomes a Gaussian distribution. This observation is also intuitive from the previous result, Theorem 2.2.4. Kurtosis of a student's t-distribution is defined only when $q > 2$ and it decreases as q increases. As q goes to zero, GT becomes an improper distribution, Jeffreys prior, which has infinite probability mass at the origin. For visualization purposes, in Figure 2.3 we show the nature of the tail of a GT distribution for different values of distributional parameters, i.e. p and q.

2.3 Acknowledgment

The material in this chapter is, in part, a reprint of material published in two articles: "Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures", IEEE Trans. Signal Processing (2016); "Learning distributional parameters for Adaptive Bayesian Sparse signal recovery", IEEE Computational Intelligence Magazine

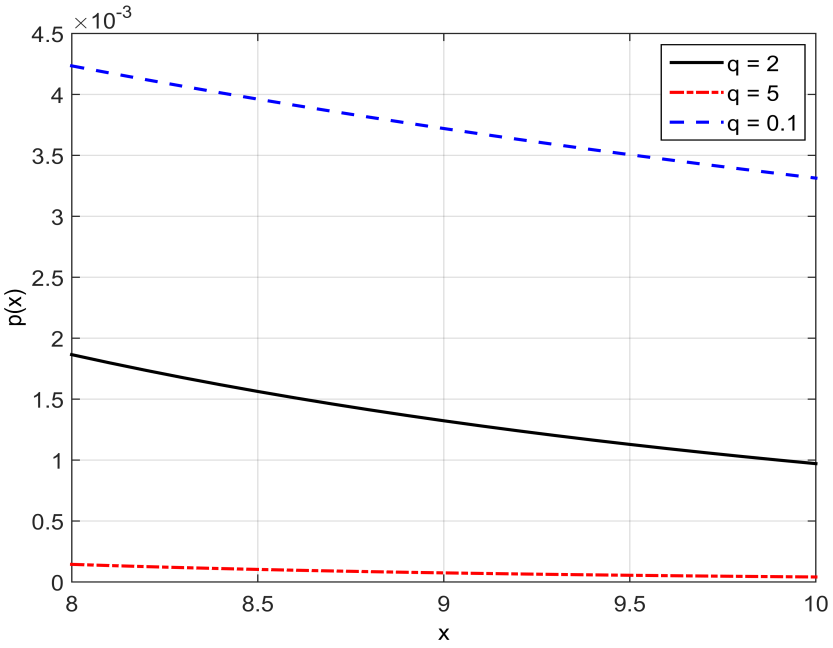

Figure 2.3. Tail behavior of Student's t distribution for different values of degrees of freedom

(2016). In all cases the dissertation author was a primary researcher and B.D. Rao supervised the research.

Chapter 3

Unified Type I and Type II Bayesian Methods for Sparse Signal Recovery

3.1 Bayesian SSR (B-SSR): Type I

Type I inference corresponds to standard MAP estimation technique in B-SSR. In this section we review the Type I framework and derive a Type I algorithm using PESM as the sparse prior. Then we specialize the result using the Generalized t distribution as the sparse prior and also show that the generalized algorithm reduces to well known SSR algorithms.

3.1.1 Background on MAP Estimation (Type I methods)

Having chosen a sparsity enforcing distribution $p(\mathbf{x})$, thereby allowing one to narrow the space of candidate solutions in a manner consistent with application-specific assumptions, a maximum a posteriori (MAP) estimator of $\mathbf{x}$ is then obtained as (Type I estimation)

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \\ &= \arg\max_{\mathbf{x}} [\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x})]\end{aligned} \tag{3.1}$$

Using the Gaussian noise assumption, and a separable prior distribution $p(\mathbf{x}) = \prod_i p(x_i)$, the MAP estimate is obtained by minimizing

$$J(\mathbf{x}) = \|\Phi\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \sum_i g(x_i), \tag{3.2}$$

where $g(x)$ is determined by $\log p(x)$. Incorporating sparsity by enforcing a sparse (supergaussian) distribution as the prior, $p(\mathbf{x})$, reduces to choosing $g(.)$. It has been shown that $g(.)$ which is symmetric, concave and nondecreasing functions on $[0,\infty)$ are useful choices in this context [134]. Now, as discussed above, many of these sparse priors can be represented in a hierarchy and belong to the PESM family.

Table 3.1. Variants of GT distribution and their connection to Type I Algorithms

q	p	Prior Distribution	Penalty Function	SSR Algorithm
$q \to \infty$	2	Normal	$\|\|x\|\|_2$	Ridge Regression
$q \to \infty$	1	Laplacian	$\|\|x\|\|_1$	LASSO
$q \geq 0$	2	Student t distribution	$\log(\varepsilon + x^2)$	Reweighted ℓ_2 (Chartrand's)
$q \geq 0$	1	Generalized Double Pareto	$\log(\varepsilon + \|x\|)$	Reweighted ℓ_1(Candes's)

In order to contrast with the Type II formulation to follow, with the PESM representation one can revisit the equation (3.1) and note that Type I involves integrating out the hyperparameter $\boldsymbol{\gamma}$.

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\max_x p(\mathbf{x}|\mathbf{y}) \\ &= \arg\max_x p(\mathbf{y}|\mathbf{x}) \int p(\mathbf{x}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) d\boldsymbol{\gamma} \end{aligned} \tag{3.3}$$

3.1.2 Unified Type I Inference Procedure

In this section we derive the EM inference procedure for the PESM family in the Type I framework, i.e, we find the MAP estimate of $\mathbf{x}$ where a PESM has been employed for the sparsity inducing prior $p(\mathbf{x})$. Because of the separable prior, the $p(x_i)$ have an independent scale mixture representation,

$$p(x_i) = \int_0^\infty p(x_i|\gamma_i) p(\gamma_i) d\gamma_i \tag{3.4}$$

For MAP estimation of $\mathbf{x}$, we treat the γ_i's as hidden variables and employ an EM algorithm. The complete data log-likelihood can be written as,

$$\log p(\mathbf{y}, \mathbf{x}, \boldsymbol{\gamma}) = \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma}) \tag{3.5}$$

To formulate the Q function, we need to find the conditional expectation of the complete data log-likelihood with respect to posterior of the hidden variables $p(\boldsymbol{\gamma}|\mathbf{x},\mathbf{y})$ which reduces to $p(\boldsymbol{\gamma}|\mathbf{x})$ by virtue of the Markovian property induced by the hierarchy, i.e. $\boldsymbol{\gamma} \rightarrow \mathbf{x} \rightarrow \mathbf{y}$. Since in the M step we need to maximize the Q function with respect to $\mathbf{x}$, we are only concerned with the first two terms in (3.5) and only the second term has dependencies on γ_i. This is the only term we need to be concerned with during the E-step. Now from the scale mixture decomposition and considering the ith component of $\mathbf{x}$,

$$\log p(x_i|\gamma_i) = \log PE(x_i; p, \gamma_i) = -\frac{|x_i|^p}{\gamma_i} + \text{constants} \tag{3.6}$$

Hence, for determining the Q function we need the following conditional expectation, $E_{\gamma_i|x_i}\left[\frac{1}{\gamma_i}\right]$.

To compute the concerned expectation we will use the following trick. Differentiating inside the integral of the marginal $p(x_i)$,

$$\begin{aligned} p'(x_i) &= \frac{d}{dx_i}\int_0^\infty p(x_i|\gamma_i)p(\gamma_i)d\gamma_i \\ &= -p \times |x_i|^{p-1}\text{sign}(x_i)\int_0^\infty \frac{1}{\gamma_i}p(x_i,\gamma_i)d\gamma_i \\ &= -p \times |x_i|^{p-1}\text{sign}(x_i)p(x_i)\int_0^\infty \frac{1}{\gamma_i}p(\gamma_i|x_i)d\gamma_i \\ &= -p \times |x_i|^{p-1}\text{sign}(x_i)p(x_i)E_{\gamma_i|x_i}\left[\frac{1}{\gamma_i}\right] \end{aligned} \tag{3.7}$$

Hence,

$$E_{\gamma_i|x_i}\left[\frac{1}{\gamma_i}\right] = -\frac{p'(x_i)}{p \times |x_i|^{p-1}\text{sign}(x_i)p(x_i)} \tag{3.8}$$

and enables determining the Q function. Then the M step reduces to,

$$\hat{\mathbf{x}}^{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2\sigma^2}||\mathbf{y} - \Phi\mathbf{x}||^2 + \sum_i w_i^{(k)}|x_i|^p \tag{3.9}$$

Where σ^2 is the variance of the measurement noise and $w_i^{(k)} = E_{\gamma_i|x_i^{(k)}}\left[\frac{1}{\gamma_i}\right]$.

Following the traditional path of EM, the algorithm is an iterative one, i.e, in the E step the weights are computed and in the M step a weighted norm minimization is solved. This alternate procedure is carried out iteratively till convergence.

3.1.3 Special cases of Type I using Generalized t distribution

In this section we specialize the derived unified Type I EM algorithm with the generalized t distribution as $p(x_i)$. We can write $p(x_i) \sim exp(-f(x_i))$ where,

$$f(x_i) = (q + 1/p)\log(1 + \frac{|x_i|^p}{q}) \tag{3.10}$$

Thus,

$$E_{\gamma_i|x_i}\left[\frac{1}{\gamma_i}\right] = \frac{f'(x_i)}{p \times |x_i|^{p-1}\text{sign}(x_i)} \tag{3.11}$$

Substituting the value of $f'(x_i)$ we get,

$$E_{\gamma_i|x_i}\left[\frac{1}{\gamma_i}\right] = \frac{q + 1/p}{q + |x_i|^p} \tag{3.12}$$

So the M step will become,

$$\hat{\mathbf{x}}^{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2\sigma^2}||\mathbf{y} - \Phi\mathbf{x}||^2 + \sum_i w_i^{(k)}|x_i|^p \tag{3.13}$$

Where σ^2 is the variance of the measurement noise and $w_i^{(k)} = E_{\gamma_i|x_i^{(k)}}\left[\frac{1}{\gamma_i}\right] = \frac{q+1/p}{q+|x_i^{(k)}|^p}$.

In following subsections we will show how with specific choices of the distribution parameters of the generalized t, we can derive well known Type I (MAP estimation) based SSR algorithms.

LASSO (ℓ_1-minimization) [156]

Interestingly we see from Table 3.1 that for specific values of the shape parameters ($q \to \infty$ and $p = 1$), a generalized t distribution can be used to represent a double exponential or Laplacian distribution. Now to relate with the unified Type I MAP estimation inference procedure, taking the limit as $q \to \infty$ in (3.12), we get $w_i = 1$. Hence in the M step we are just solving a ℓ_1 penalized regression once as the weights are not changing over iterations, which is essentially the LASSO algorithm.

Reweighted ℓ_1-minimization (Candes et al [30])

The popular reweighted ℓ_1-minimization (Candes et al [30]) is a special case of the MAP estimation approach using a generalized t distribution as sparse prior.

Selecting the parameters of the generalized t as follows; $q = \varepsilon, p = 1$, one obtains,

$$p(x_i|\varepsilon) = GT(1,\varepsilon) = \frac{\eta}{\left(1+\frac{|x_i|}{\varepsilon}\right)^{(\varepsilon+1)}} \tag{3.14}$$

which when substituted in equation (3.2), results in the following cost function,

$$\min_x ||y-\Phi x||_2^2 + \lambda \sum_i \log(|x_i|+\varepsilon) \tag{3.15}$$

In [30], the above mentioned cost function is optimized using a MM approach. Now substituting the distribution parameters in equation (3.12), the weights reduce to $w_i = \frac{1+\varepsilon}{\varepsilon+|x_i|}$. These are the same weights obtained in [30] via a MM method and $p = 1$ in Equation (3.13) results in a weighted ℓ_1 minimization problem with the weights being a function of the previous estimate. This special case of GT has been also called the Generalized Double Pareto (GDP) distribution in the literature [10].

Following the scale mixture decomposition of the GT distribution, since $p = 1$ we can represent the prior as a Laplacian Scale Mixture.

$$p(x) = \int p(x|\gamma)p(\gamma)d\gamma = \int \frac{1}{2\gamma} e^{-\frac{|x|}{\gamma}} p(\gamma)d\gamma, \tag{3.16}$$

where $p(\gamma) = IG(\gamma;\varepsilon,\varepsilon)$. This observation is summarized in the following lemma.

Lemma 3.1.1 *Let $x \sim Laplacian(0,\gamma)$, $\gamma \sim IG(\gamma;\varepsilon,\varepsilon)$, then the resulting marginal density for x is $GT(1,\varepsilon)$.*

Reweighted ℓ_2-minimization ([33,146])

Another popular SSR algorithm, the reweighted ℓ_2 minimization can also be represented in a Bayesian Type I setting by employing a Student t distribution. This heavytailed sparse prior $p(x)$ is again a special case of the generalized t distribution as shown in the table.

$$p(x_i|\varepsilon) = GT(2,\varepsilon) = \frac{\eta}{\left(1+\frac{|x_i|^2}{\varepsilon}\right)^{(\varepsilon+1/2)}} \tag{3.17}$$

The nature of the tail of the student t distribution is controlled by degrees of freedom parameter ε and smaller values of ε correspond to heavier tails. The associated diversity penalty factor is given by $g(x_i) = \log(x_i^2 + \varepsilon)$. For a Type I inference procedure, we can utilize the unified approach discussed above in Section 3.1.3 and substitute the shape and scale parameters $p = 2, q = \varepsilon$ of the generalized t distribution in Equation (3.12) to obtain, $w_i = \frac{\varepsilon+1/2}{\varepsilon+|x_i|^2}$. Since $p = 2$, Equation (3.13) leads to the reweighted ℓ_2 minimization algorithm as discussed in [33].

3.2 B-SSR: Type II (Evidence Maximization)

The success of Type II approaches like SBL for SSR problems motivate the Type II approach for the general PESM family. As special cases, the three Type I algorithms discussed in Section 3.1.3 are explored in the Type II setting. We also analyze the difference between a Type I algorithm and its Type II counterpart which provides insights into the reasons for superior recovery performance of Type II methods.

In a Type II procedure, instead of integrating out the hypeparameters $\boldsymbol{\gamma}$, we estimate them using an evidence maximization method, i.e.

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\mathbf{y}) = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma})p(\mathbf{y}|\boldsymbol{\gamma}) \\ &= \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}) \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\boldsymbol{\gamma})d\mathbf{x}\end{aligned} \tag{3.18}$$

The evidence framework integrates over the coefficient vector $\mathbf{x}$ to obtain the evidence $p(\mathbf{y}|\boldsymbol{\gamma})$. This evidence is weighted by the hyperprior $p(\boldsymbol{\gamma})$ and maximized over $\boldsymbol{\gamma}$. Once $\boldsymbol{\gamma}$ is obtained, the relevant posterior $p(\mathbf{x}|\mathbf{y})$ is approximated, often as $p(\mathbf{x}|\mathbf{y};\hat{\boldsymbol{\gamma}})$, and the mean of the approximated posterior is used as a point estimate. Sparsity is achieved by many of the γ_i being zero [157, 164, 166].

3.2.1 Unified Type II EM algorithm

To solve the above mentioned optimization problem, we again employ the EM algorithm this time by treating $\mathbf{x}$ as the hidden variable. As in previous section, we assume a sparse prior $p(\mathbf{x})$ from the PESM family has been utilized and that the measurement noise is Gaussian with variance σ^2.

Hence the Q function has the form,

$$\begin{aligned} Q(\boldsymbol{\gamma}) &= E_{\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\sigma^2}[\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})] \\ &\approx E_{\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\sigma^2}[\sum_i -\frac{1}{p}\log\gamma_i - \frac{|x_i|^p}{\gamma_i} + \log p(\gamma_i)] \end{aligned} \tag{3.19}$$

Since in the M step we are only concerned with the terms involving $\boldsymbol{\gamma}$, examining them reveals that the E-step requires the computation of the following conditional expectation

$$E_{\mathbf{x}|\mathbf{y};\boldsymbol{\gamma}',\sigma^2}[|x_i|^p] = < |x_i|^p > \tag{3.20}$$

In the M step we will maximize the Q function with respect to γ_i to find the update rules. To illustrate, if we consider a non informative hyperprior, i.e, $p(\gamma_i) = 1$,

$$Q(\boldsymbol{\gamma}) = \sum_i -\frac{1}{p}\log\gamma_i - \frac{< |x_i|^p >}{\gamma_i} \tag{3.21}$$

Taking the derivative of the Q function w.r.t γ_i and setting it to zero results in,

$$\hat{\gamma}_i = p < |x_i|^p > \tag{3.22}$$

Since the E step requires the computation of the conditional expectation given by Equation (3.20), we can either look for a closed form solution or revert to the MCMC technique [137]. We will examine this further for some special cases later.

3.2.2 Difference between Type I and Type II inference methods

Type I and Type II provide two different approaches to solving the SSR problem. Hence it is important to understand the theoretical differences between the two inference procedures to identify their suitability for SSR. In [165], the authors provide evidence for

SBL, using a variational approximation to the prior $p(\mathbf{x})$, that Type II methods attempt to approximate the true posterior $p(\mathbf{x}|\mathbf{y})$. Similar discussion of Type II desirability is provided in [118] in the context of general Bayesian inferencing. We revisit the issue and attempt to corroborate this by exploiting specific attributes of the SSR problem. We first manipulate the Type II objective as shown below.

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{y}) &= \int p(\boldsymbol{\gamma},\mathbf{x}|\mathbf{y})d\mathbf{x} \\ &= \int p(\boldsymbol{\gamma}|\mathbf{x},\mathbf{y})p(\mathbf{x}|\mathbf{y})d\mathbf{x} \\ &= \int p(\boldsymbol{\gamma}|\mathbf{x})p(\mathbf{x}|\mathbf{y})d\mathbf{x} \\ &= p(\boldsymbol{\gamma})\int \frac{p(\mathbf{x}|\boldsymbol{\gamma})}{p(\mathbf{x})}p(\mathbf{x}|\mathbf{y})d\mathbf{x} \end{aligned} \tag{3.23}$$

Lets assume that $\hat{\boldsymbol{\gamma}}$ is the solution of Equation (3.18). It will be sparse for specific choice of $p(\boldsymbol{\gamma})$ as shown in [164, 166].

Now, let $\underline{S}$ be the index of non zero entries and $\overline{S}$ be the index of zero entries. So, we can say $\hat{\boldsymbol{\gamma}}_{\overline{S}} = 0$.

$$\begin{aligned} p(\hat{\boldsymbol{\gamma}}|\mathbf{y}) &= \lim_{\varepsilon\to 0} p(\hat{\boldsymbol{\gamma}}+\varepsilon|\mathbf{y}) \\ &= p(\hat{\boldsymbol{\gamma}}) \lim_{\varepsilon\to 0}\int_{\underline{S}}\int_{\overline{S}} \frac{p(\mathbf{x}_{\underline{S}}|\hat{\boldsymbol{\gamma}}_{\underline{S}}+\varepsilon_{\underline{S}})p(\mathbf{x}_{\overline{S}}|\varepsilon_{\overline{S}})}{p(\mathbf{x}_{\underline{S}})p(\mathbf{x}_{\overline{S}})}p(\mathbf{x}|\mathbf{y})d\mathbf{x} \end{aligned} \tag{3.24}$$

$p(\mathbf{x}_{\overline{S}}|\varepsilon_{\overline{S}})$ is a normal distribution with mean zero and variance $\varepsilon_{\overline{S}}$. Hence when $\varepsilon_{\overline{S}} \to 0$, $p(\mathbf{x}_{\overline{S}}|\varepsilon_{\overline{S}})$ becomes a dirac delta function, i.e. $\delta(x_{\overline{S}})$.

Using the properties of dirac delta functions inside the integration, we obtain

$$p(\hat{\boldsymbol{\gamma}}|\mathbf{y}) = \int_{\underline{S}} \frac{p(\mathbf{x}_{\underline{S}}|\hat{\boldsymbol{\gamma}}_{\underline{S}})}{p(\mathbf{x}_{\underline{S}})} \frac{p(\hat{\boldsymbol{\gamma}})}{p(\mathbf{x}_{\overline{S}}=0)}p(\mathbf{x}_{\underline{S}},\mathbf{x}_{\overline{S}}=0|\mathbf{y})d\mathbf{x}_{\underline{S}} \tag{3.25}$$

Hence from this analysis, we see that we are evaluating a weighted integral of the true

posterior $p(\mathbf{x}|\mathbf{y})$ over the subspaces spanned by the non zero indexes. This shows that in the evidence maximization framework instead of looking for the mode of the true posterior $p(\mathbf{x}|\mathbf{y})$, we approximate the true posterior by $p(\mathbf{x}|\mathbf{y};\hat{\boldsymbol{\gamma}})$ where $\hat{\boldsymbol{\gamma}}$ is obtained by maximizing the true posterior mass over the subspaces spanned by the non zero indexes. This is in contrast to Type I methods that seek the mode of the true posterior and use that as the point estimate of the desired coefficients. Hence, if the true posterior distribution has a skewed peak, then the Type I estimate (Mode) is not a good representative of the whole posterior, where as by going after the true posterior mass, Type II methods will give us a better estimate.

In [166] authors have pointed out another key advantage of Type II framework for the case of ARD (Automatic Relevance Determination) prior which has been used in Sparse Bayesian Learning (SBL) [157], that the dictionary dependency of the Type II priors lead to scale invariant (with respect to the dictionary atoms) recovery of the desired sparse coefficient vector.

Another favorable aspect of the Type II framework is that it inherits the robustness property of a Hierarchical Bayesian modeling framework. It has been shown extensively in the statistics literature [75, 78, 105], that the posterior of a hyperparameter, i.e, $\boldsymbol{\gamma}$, is less affected by the wrong choices of prior than the posterior of the parameter $\mathbf{x}$. In other words, parameters that are deeper in the hierarchy have less effect on the inference procedure, which allows us to be less concerned about the choice of $p(\boldsymbol{\gamma})$. Another virtue is that the hierarchical framework allows for parameter tying and this can greatly reduce the search space for Type II methods by leading to an optimization problem with fewer parameters. This is more evident for problems like the MMV and block sparsity problem [168, 178, 179].

3.2.3 Special case of Unified Type II with different choices of p

As discussed above for the unified Type II approach our concerned posterior is $p(\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\sigma^2)$. For a point estimate of $\mathbf{x}$ we will use the mean of the posterior, $\hat{\mathbf{x}} = \int \mathbf{x} p(\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\sigma^2) d\mathbf{x}$. Now the posterior could be computed as,

$$\begin{aligned} p(\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\sigma^2) &\approx p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|\boldsymbol{\gamma}) \\ &\approx \exp\{-\frac{1}{2\sigma^2}||\mathbf{y}-\Phi\mathbf{x}||_2^2 - \sum_i \frac{|x_i|^p}{\gamma_i}\} \end{aligned} \tag{3.26}$$

The challenge is proper normalization and tractability of the computation of the mean. For the EM algorithm to be successfully implemented, one must also be able to carry out the E-step, Equation (3.20). We now explore this for some specific PESM family members.

Choice of $p = 2$

Choice of $p = 2$ corresponds to Gaussian Scale Mixture, and is very tractable. The GSM based Type II methods have been extensively studied [88, 157, 165] and so we keep the discussion brief. This choice (in Equation (3.26)) leads to a Gaussian posterior given by

$$p(\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\sigma^2) = N(\mu,\Sigma) \tag{3.27}$$

where

$$\mu = \Gamma\Phi^T(\sigma^2 I + \Phi\Gamma\Phi^T)^{-1}\mathbf{y} \tag{3.28}$$

$$\Sigma = \Gamma - \Gamma\Phi^T(\sigma^2 I + \Phi\Gamma\Phi^T)^{-1}\Phi\Gamma \tag{3.29}$$

and $\Gamma = \text{diag}(\boldsymbol{\gamma})$. The EM algorithm can also be readily carried out because the E-step requires the second moment which can be readily obtained using Equation (3.29). The

estimate of $\boldsymbol{\gamma}$ in the M step and the updates of $\boldsymbol{\gamma}$ depend on the mixing density $p(\boldsymbol{\gamma})$ as shown in Equation (3.19) and can be readily carried out for the non-informative prior and for a reasonable large class of priors [132]. The true posterior can be approximated by a Gaussian distribution whose mean and covariance depend on the estimated hyperparameters. Now, for a point estimate of the coefficient vector, we will choose,

$$\hat{\mathbf{x}} = \mu. \tag{3.30}$$

From Equation (3.28), one can see that μ is sparse if $\boldsymbol{\gamma}$ is sparse. To complete the discussion, we discuss the most popular of the Type II methods. In Relevance Vector Machine (Type II) [157], Tipping has shown that the 'true' coefficient prior used in SBL actually follows a student t distribution (GSM with Gamma distribution as mixing density), and discusses in detail how the hierarchical formulation of this prior helps to realize the supergaussian nature. Hence we can see that the corresponding Type II formulation of Reweighted ℓ_2 is SBL with a slight difference. In SBL ε is set to zero which gives us an improper prior $p(x) \sim 1/|x|$ which is sharply peaked at zero. But as discussed in previous literatures, $\varepsilon = 0$ in Type I version, i.e, in Reweighted ℓ_2 increases the number of local minima and convergence to a sub optimal solution becomes more likely. Now to solve the M step for this case we will use the following PESM ($p = 2$) formulation,

Lemma 3.2.1 *Let $x \sim N(0,\gamma)$, $\gamma \sim IG(\varepsilon,\varepsilon)$ Then the resulting marginal density for x is $GT(2,\varepsilon) \simeq Student - t(2\varepsilon)$.*

Details of this inference procedure can also be found in [88, 157], and update rules have been shown in Table 3.2.

Choice of $p = 1$

With $p = 1$, PESM reduces to a Laplacian Scale Mixture. To successfully carry out the EM algorithm, the E-step requires the computation of $E(|x_i|; \mathbf{y}, \boldsymbol{\gamma}^{(k)})$. A closed form expression does not appear feasible and a more numerical approach may be required. Also, the concerned posterior (Equation 3.26) does not appear to have a simple closed form expression making final inferencing a challenge along with the computation of the mean for the point estimate. An efficient numerical approach needs to be developed and is left for future work.

In this work, we follow an alternate strategy and take advantage of the fact that the LSM family is contained within the GSM family. Since a Laplacian distribution can be written as a member of the GSM family (Equation 2.3) [13, 59], it will be possible to get a closed form posterior using a three layer hierarchy. We will illustrate this for the prior associated with Type I Reweighted ℓ_1-minimization approach and develop a Type II variant. The closed form posterior will be Gaussian and have the same form for the case of $p = 2$ as shown in Equation (3.27). The only difference between $p = 2$ and $p = 1$ lies in the estimation of the hyperparameters.

Type II ℓ_1 variant can also be derived and has been dealt with in previous work [13] and for sake of completeness the update rule is summarized in Table 3.2 along with other Type II algorithms. We will now derive the M step for the case of Type II Reweighted ℓ_1-minimization which can be followed in a straightforward manner for other cases including the ℓ_1 variant.

We have shown in the discussion of Type I Reweighted ℓ_1 that the concerned prior $GT(1, \varepsilon)$ in a Bayesian setting is a Laplacian Scale mixture. This prior can be represented in a 3 layer hierarchy involving a GSM representation for the Laplacian density as summarized below.

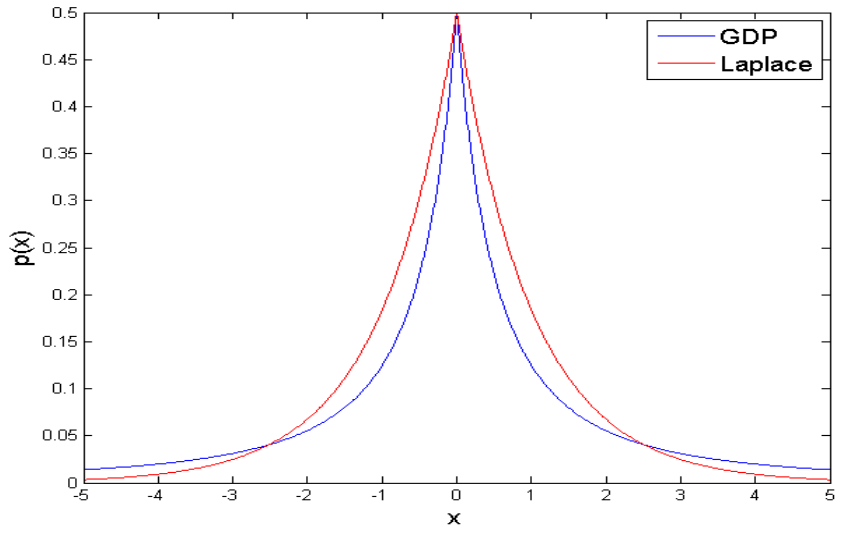

Figure 3.1. Comparison of tail behavior of two distributions: Generalized Double Pareto (GDP) and Laplacian

Lemma 3.2.2 *Let* $x \sim N(0,\gamma)$, $\gamma \sim Exp(\frac{\lambda^2}{2})$ *and* $\lambda \sim Gamma(\varepsilon,\varepsilon)$ *where* $\varepsilon > 0$. *Then the resulting marginal density for* x *is* $GT(1,\varepsilon)$.

Fig. 3.1 compares two corresponding densities, $GT(1,1)$ and Laplace distribution with $\lambda = 1$. It is evident from this figure that the Laplace prior has relatively light tails which contributes to the problem of over-shrinking of the large coefficients. On the other hand, the generalized t distribution has relatively heavier tails and a peak at zero which promotes zero coefficients. This is another reason of the superior recovery performance of Reweighted ℓ_1-minimization over the LASSO, i.e. ℓ_1-minimization, approach.

Table 3.2. Updating Hyperparameters of Type II Algorithms

Type II algorithm	Mixing Density	Update Rules
Type II ℓ_1	$p(\gamma_i\|\lambda) = Exp(\lambda/2)$	$\hat{\gamma}_i = \frac{-1+\sqrt{1+4\lambda(\mu_i^2+\Sigma_{i,i})}}{2\lambda}$, $\hat{\lambda} = \frac{2M}{\sum_i \gamma_i}$
Type II Re-ℓ_1	$p(\gamma_i\|\lambda) = Exp(\lambda^2/2)$, $p(\lambda) = Gamma(\varepsilon,\varepsilon)$	$\hat{\gamma}_i = \frac{-1+\sqrt{1+4\lambda^2(\mu_i^2+\Sigma_{i,i})}}{2\lambda^2}$, $\hat{\lambda} = \frac{-\varepsilon+\sqrt{\varepsilon^2+4(2M+\varepsilon-1)\sum\gamma_i}}{2\sum\gamma_i}$
Type II Re-ℓ_2	$p(\gamma_i\|\varepsilon) = Inv-Gamma(\varepsilon,\varepsilon)$	$\hat{\gamma}_i = \frac{\mu_i^2+\Sigma_{i,i}+2\varepsilon}{2\varepsilon+1}$

Now, for estimation of hyperparameters $\boldsymbol{\gamma}$ and λ in the three layer hierarchy, an EM algorithm will be developed. As in Section 3.2.1, using $(\mathbf{y},\mathbf{x})$ as the complete data, maximizing the conditional expectation of the complete data log likelihood involves maximizing,

$$Q(\boldsymbol{\gamma},\lambda,\sigma^2) = E_{\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\lambda,\sigma^2}[\log p(\mathbf{y},\mathbf{x};\boldsymbol{\gamma},\lambda,\sigma^2)] \tag{3.31}$$

In the E step, for iteration t, we only need to compute the second moment which is straightforward because of the GSM representation of the Laplacian, i.e.

$$E_{\mathbf{x}|\mathbf{y};\boldsymbol{\gamma}^t,\lambda,\sigma^2}[x_i^2] = \Sigma_{(i,i)} + \mu_i^2 \tag{3.32}$$

In the M step, the Q function is maximized with respect to the hyperparameters, $\boldsymbol{\gamma}$ and λ.

$$\begin{aligned} Q(\boldsymbol{\gamma},\lambda) = E_{\mathbf{x}|\mathbf{y};\boldsymbol{\gamma},\lambda,\sigma^2}[&\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\boldsymbol{\gamma}) \\ &+ \log p(\boldsymbol{\gamma}|\lambda) + \log p(\lambda|\varepsilon)] \end{aligned} \tag{3.33}$$

Now using the E step and only retaining the terms that involve $\boldsymbol{\gamma}$ and λ we obtain,

$$\begin{aligned} Q(\boldsymbol{\gamma},\lambda) = &-\frac{1}{2}\sum_i \log\gamma_i - \frac{1}{2}\sum_i \frac{\Sigma_{(i,i)} + \mu_i^2}{\gamma_i} \\ &+ \sum_i (2\log\lambda - \frac{\lambda^2}{2}\gamma_i) + (\varepsilon - 1)\log\lambda - \varepsilon\lambda \end{aligned} \tag{3.34}$$

In the M step, taking the derivative of the Q function w.r.t γ_i and λ and setting to

zero results in.

$$\frac{\partial Q}{\partial \gamma_i} = -\frac{1}{2\gamma_i} + \frac{\Sigma_{(i,i)} + \mu_i^2}{2\gamma_i^2} - \frac{\lambda^2}{2} = 0 \tag{3.35}$$

Solving this quadratic equation we obtain,

$$\hat{\gamma}_i = \frac{-1 + \sqrt{1 + 4\lambda^2(\mu_i^2 + \Sigma_{i,i})}}{2\lambda^2} \tag{3.36}$$

Similarly,

$$\frac{\partial Q}{\partial \lambda} = \frac{2M + \varepsilon - 1}{\lambda} - \lambda \sum_i \gamma_i - \varepsilon = 0 \tag{3.37}$$

Hence,

$$\hat{\lambda} = \frac{-\varepsilon + \sqrt{\varepsilon^2 + 4(2M + \varepsilon - 1)\sum_i \gamma_i}}{2\sum_i \gamma_i} \tag{3.38}$$

We can also estimate the measurement noise variance σ^2 by maximizing the above Q function as shown in [157]. In this work, for simplicity, we will assume that the SNR of the environment is known to us before hand. We can also employ a fixed point optimization technique as shown in [157] to estimate the hyperparameters.

After convergence, one finds that most of the γ_i, i.e. the variance of the normal distribution are driven to zero, which makes the associated coefficient zero and prunes it out from the model.

3.3 Numerical Experiments

In this section we present a set of experiments to evaluate and compare the Type II/Hierarchical framework based methods with those based on regularization framework, i.e. Type I methods (MAP estimation), for the task of sparse signal recovery. The

experimental setup used is quite standard and has been used widely in the SSR literatures.

3.3.1 Problem Specification

The measurement vector $\mathbf{y}$ is generated using a $N \times M = 50 \times 250$ dictionary Φ, whose elements are generated from a i.i.d normal distribution with mean=0 and variance=1. A sparse signal $\mathbf{x}_{gen}$ of length 250 is generated such that $||\mathbf{x}_{gen}||_0 = k$. The support, i.e. the location of the k nonzero elements, is chosen randomly, and the values are chosen from three different distributions:

(I) Uniform ± 1 random spikes. (Sub-Gaussian)

(II) Zero mean unit variance Gaussian.

(III) Student t distribution with degrees of freedom $\nu = 3$. (Super-Gaussian)

The synthetic measurements are generated using $\mathbf{y} = \Phi\mathbf{x}_{gen}$. The generated measurements and the dictionary are then provided as input to the algorithms. The estimated coefficients are compared with the original $\mathbf{x}_{gen}$ that has been used to generate the measurement. For a single instance, the method is credited with a successful recovery if the estimate $\hat{\mathbf{x}}$ satisfies,

$$||\mathbf{x}_{gen} - \hat{\mathbf{x}}||_\infty \leq 10^{-3} \tag{3.39}$$

500 trials are conducted for various fixed combinations of k, i.e. the number of non zero coefficients, and the probability of successful recovery is plotted with respect to k. As expected, the probability of successful recovery decreases as k, i.e. the cardinality of support, increases.

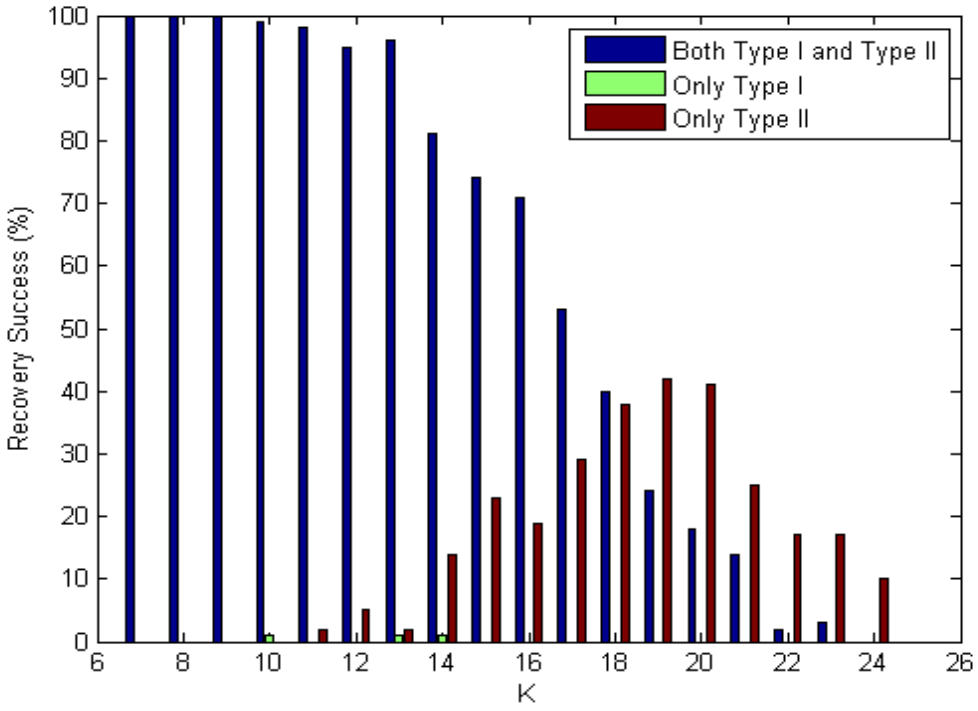

Figure 3.2. Recovery performance for Type I and Type II Reweighted ℓ_1 minimization

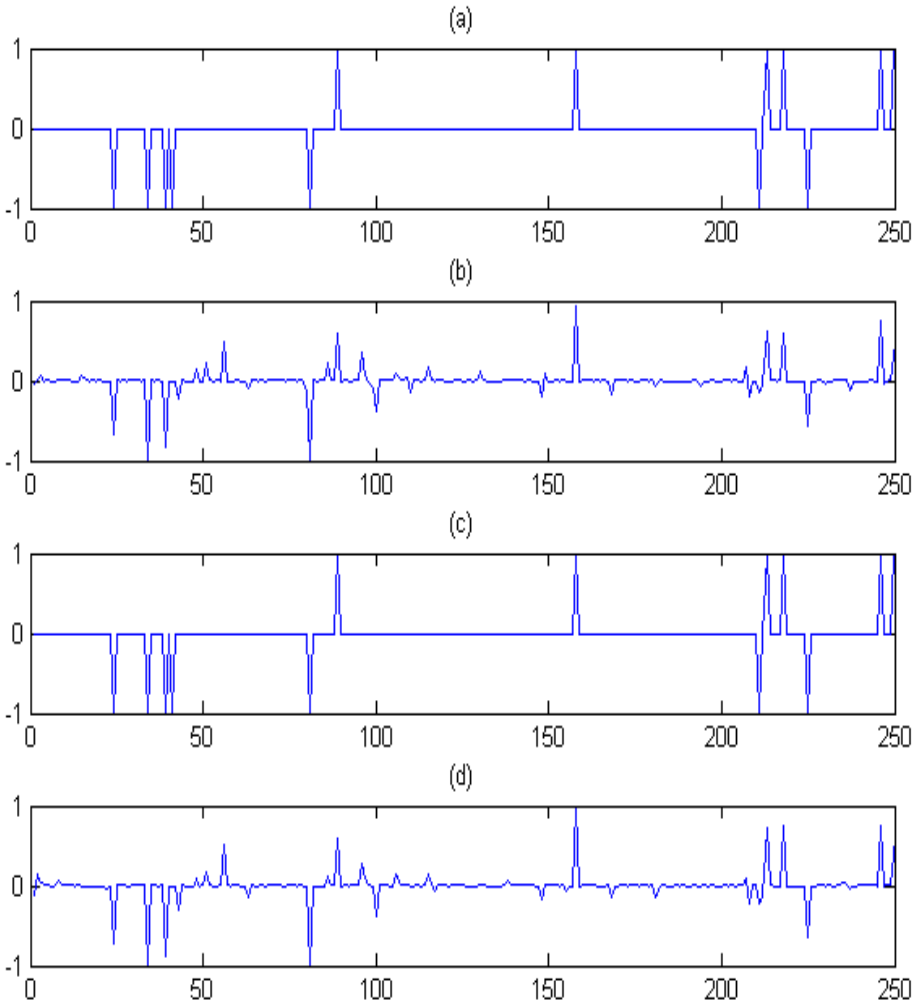

Figure 3.3. Reconstruction of uniform spikes where $k = 13$ using (a) Original Signal, (b) ℓ_1 norm minimization (Type I), (c) Type II ℓ_1 minimization, (d) Candes et al (Type I) Reweighted ℓ_1 minimization

3.3.2 Recovery Performance

Competing Algorithms

Since the main goal of our work is to compare the Type II algorithms with their Type I counterparts, we designed the Type II versions of three well known norm minimization based Type I algorithms and compare their performance. The algorithms in the study are:

- ℓ_1 minimization based SSR. (Basis Pursuit)
- Type II ℓ_1 minimization based SSR. (Fixed $\lambda = 5$)
- Type I Reweighted ℓ_1 minimization. ($\varepsilon = 0.1$ [30])
- Type II Reweighted ℓ_1 minimization (Fixed $\varepsilon = 100$)
- Type I Reweighted ℓ_2 minimization. (ε regularized, optimal update from [33])
- Type II Reweighted ℓ_2 minimization (Fixed $\varepsilon = 0$: SBL)

Performance Comparison

In Figure 3.4, the probability of successful recovery with increasing support cardinality is plotted for the case where the non zero coefficients are from a zero mean, unit variance, Gaussian distribution. It is evident from this plot that for all the algorithms, Type II versions outperform their Type I counterparts. This performance difference is significant in case of ℓ_1 norm minimization. Type I Reweighted ℓ_2 minimization approach works much better compared to other two Type I methods, and the reason being the heuristic update of ε, which helps it to get away from local minima. Hence, ε update in Reweighted ℓ_2 (Type I) is absolutely necessary as we have found out for fixed ε this

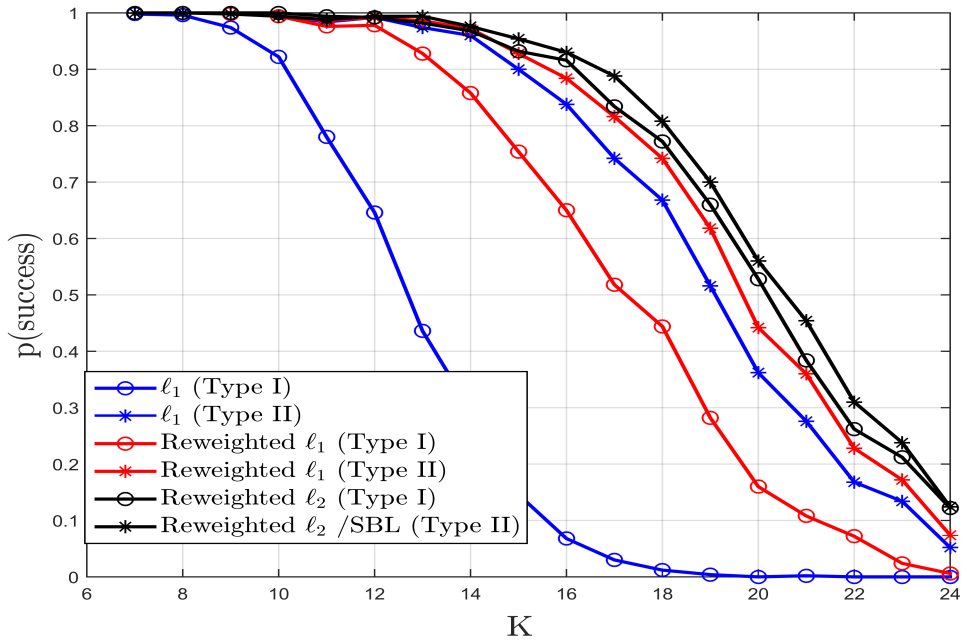

Figure 3.4. Recovery performance with Gaussian distributed non zero coefficients

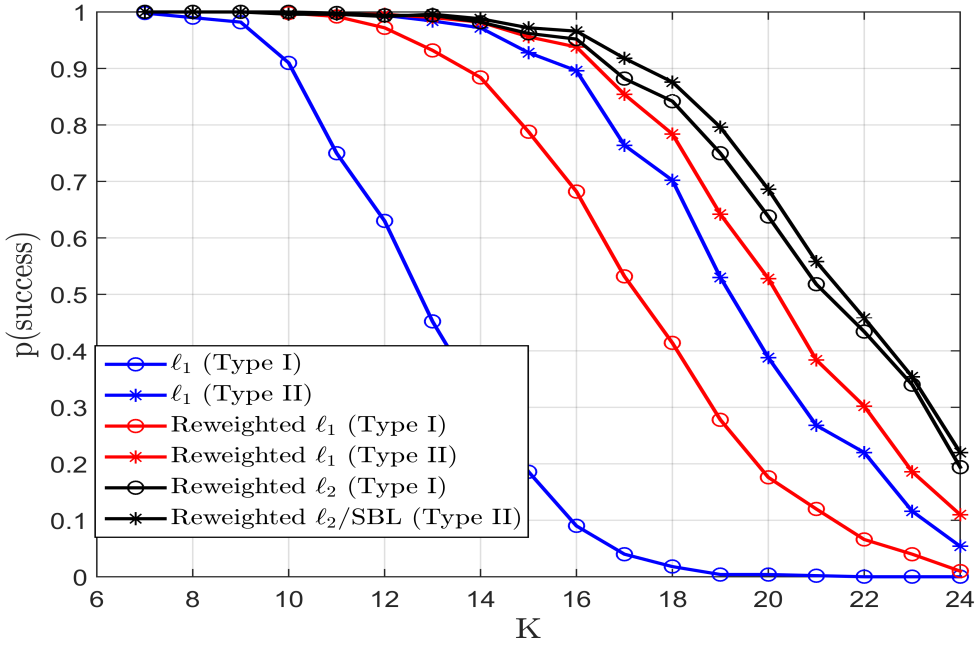

Figure 3.5. Recovery performance with Super Gaussian (Student t) distributed non zero coefficients

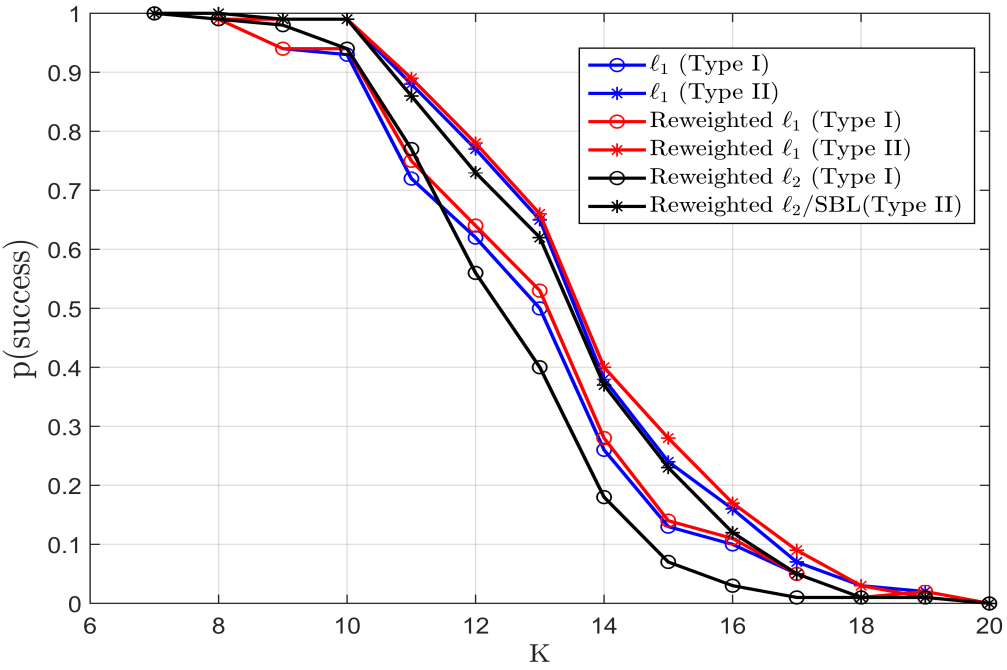

Figure 3.6. Recovery performance with Sub Gaussian distributed non zero coefficients

algorithm's performance decreases significantly. Figure 3.2 shows this comparison for the Reweighted ℓ_1 minimization (Candes et al) in detail. The figure indicates trials when both Type I and Type II method have been successful and when only one of them has been successful and it is evident that for high values of k, Type II outperforms Type I by a significant margin.

In Figure 3.5, the probability of successful recovery with increasing support cardinality is plotted where the non zero coefficients are generated from a student's t distribution with degrees of freedom 3. Again, the empirical superiority of the Type II versions over their Type I counterparts is evident from Figure 3.5. Interesting point to note here, is the performance improvement of Type I and Type II version of Reweighted ℓ_2 algorithm over the others is significant and the reason could be that assumed prior for the non zero coefficients and the true prior have the same tail behavior (student's t) and are better matched.

Finally, we repeat the same set of experiments where the non zero coefficients follow a sub-gaussian distribution, i.e. Uniform ± 1 random spikes, and the plot of the probability of successful recovery with increasing support cardinality is shown in Figure 3.6. Though Type II methods still outperform their Type I counterparts, the performance improvement is less significant compared to the previous two cases. The reason could be that, since the assumed priors are supergaussian, i.e. heavy tails, it is difficult to model the true prior, i.e. sub gaussian density, for the nonzero coefficients. In Figure 3.3, an instance of reconstruction is shown using $k = 13$ along with the original signal. It is evident that both ℓ_1 minimization (Type I) and Candes's Reweighted ℓ_1 minimization (Type I) fail, whereas Type II version of ℓ_1 minimization recovers the original signal. For this instance, the other three SSR algorithms have also been successful in recovering the original signal.

3.4 Conclusion and Discussion

In this chapter, we formulated the SSR problem from a Bayesian perspective and presented two different Bayesian frameworks which encompass all the well known recovery algorithms in practice. We presented a generalized scale mixture family : PESM, which is of prime importance for the design of Hierarchical Bayesian Recovery algorithms, i.e, Type II algorithms. The unified treatment of both ℓ_1 and ℓ_2 norm minimization based algorithms along with the design of Type II version of the Reweighted ℓ_1 minimization algorithm are the main contributions of this work.

We also showed that, in a hierarchical Bayes framework instead of looking for a mode of the true posterior Type II methods actually try to find an approximate posterior such that the mass of the true posterior over the subspace spanned by non zero indexes is maximized. This leads to a better approximation of the true posterior, which is the reason behind the superior empirical results obtained using the Type II framework. Type

II framework also enjoys the robustness property inherited because of its connection with Hierarchical Bayes which allows one to be less concerned about the choice of prior on the hyperparameters.

3.5 Acknowledgment

The material in this chapter is, in part, a reprint of material published in two articles: ”Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures”, IEEE Trans. Signal Processing (2016) and, ”Hierarchical Bayesian Formulation of Sparse Signal Recovery Algorithms using Scale Mixture Priors”, Asilomar Conference on Signals, Systems and Computers, 2015. In all cases the dissertation author was the primary researcher and B.D. Rao supervised the research.

Chapter 4

Learning Distributional Parameters

4.1 Introduction

In the previous chapter we have shown how by employing a sparsity promoting distribution from the GT family, we can derive a unified MAP estimation framework which includes many of the popular SSR algorithms. In this chapter, first we provide an alternative derivation of the MAP estimation framework, albeit limited only to GT family instead of PESM family. In addition, we propose an adaptive framework of learning the distributional parameters of GT over the iterations based on the measurements, instead of fixing them beforehand.

4.2 MAP Estimation with GT prior (Fixed distributional parameters)

In this section we derive the Expectation Maximization (EM) [43] based inference procedure, which is a popular iterative method for finding an MAP estimate in statistical models, i.e. we find the MAP estimate of $\mathbf{x}$ where a GT distribution with fixed distributional parameters p and q, has been employed as the sparsity inducing prior $p(\mathbf{x})$. In contrast to the derivation in our previous work [72], a more direct and simpler derivation is provided, albeit limited to the GT family. For the derivation we utilize the fact that a GT distribution can be decomposed as a PESM. Because of the separable prior, each $p(x_i)$ has an independent scale mixture representation,

$$p(x_i) = \int_0^\infty p(x_i|\gamma_i)p(\gamma_i)d\gamma_i. \tag{4.1}$$

For MAP estimation of $\mathbf{x}$, we treat the γ_i's as hidden variables and employ an EM algorithm. The complete data log-likelihood can be written as,

$$\log p(\mathbf{y},\mathbf{x},\boldsymbol{\gamma}) = \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma}). \tag{4.2}$$

Now we formulate the Q function, i.e. the conditional expectation of the complete data log-likelihood with respect to posterior of the hidden variables $p(\boldsymbol{\gamma}|\mathbf{x},\mathbf{y})$, which reduces to $p(\boldsymbol{\gamma}|\mathbf{x})$, by virtue of the Markovian property induced by the hierarchy, i.e. $\boldsymbol{\gamma} \rightarrow \mathbf{x} \rightarrow \mathbf{y}$. This leads to

$$Q(\mathbf{x},p,q) = E_{\gamma|\mathbf{x}}\big[\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})\big]. \tag{4.3}$$

Since power exponential ($p(x_i|\gamma_i)$) and inverse gamma distribution ($p(\gamma_i)$) are conjugate [45], we can find a closed form for the concerned posterior,

$$p(\gamma_i|x_i) \sim IG(q+\frac{1}{p}, |x_i|^p + q). \tag{4.4}$$

Since in the M step we need to maximize the Q function with respect to $\mathbf{x}$, we only consider the terms that have dependencies on $\mathbf{x}$, i.e. the first two terms in (4.3). But only the second term has dependencies on γ_i, hence this is the only term we need to be concerned with during the E step. Now from the scale mixture decomposition and considering the i^{th} component of $\mathbf{x}$,

$$\log p(x_i|\gamma_i) = \log PE(x_i;p,\gamma_i) = -\frac{|x_i|^p}{\gamma_i} + \text{constants}. \tag{4.5}$$

Hence, for determining the Q function we need the following conditional expectation, $E_{\gamma_i|x_i}\big[\frac{1}{\gamma_i}\big]$ which can be computed using Equation (4.4) as,

$$E_{\gamma_i|x_i}\big[\frac{1}{\gamma_i}\big] = \frac{q+\frac{1}{p}}{|x_i|^p + q}. \tag{4.6}$$

Then the M step reduces to,

$$\hat{\mathbf{x}}^{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2\lambda}||\mathbf{y} - \Phi\mathbf{x}||^2 + \sum_i w_i^{(k)}|x_i|^p \tag{4.7}$$

where λ is the variance of the measurement noise and $w_i^{(k)} = E_{\gamma_i|x_i^{(k)}}\left[\frac{1}{\gamma_i}\right]$.

Following the traditional path of EM, the algorithm is an iterative one, i.e. in the E step the weights are computed and in the M step a weighted norm minimization is solved. This alternate procedure is carried out iteratively till convergence. As discussed in [72], following Table 3.1 if we substitute the distributional parameters p and q in the weight computation stage, Equation (4.7) leads to several popular weighted norm minimization based SSR algorithms such as LASSO [156], Reweighted ℓ_1 norm minimization [30] and Reweighted ℓ_2 norm minimization [33].

4.3 Learning Distributional Parameters

In the previous section we have shown how by choosing specific distributional parameters of GT, we can derive different popular SSR algorithms. From a Bayesian perspective it can be interpreted such as, for each algorithm a fixed prior distribution has been employed. But in real life that assumed prior distribution may be significantly different from the true prior distribution of the data. In this work we propose to learn the distributional parameters from the data allowing to adapt our algorithm to the true prior. As shown before, p and q are distributional parameters of GT and control the tail nature of the prior. Learning them over iterations will help us to model the true tail nature.

Again EM algorithm will be employed and the only difference from the previous section will come in the M step, where we need to optimize the cost function with respect to both the distributional parameters, p and q along with the desired coefficient vector $\mathbf{x}$. For this we revisit the Q function of EM given by (4.3). Unlike the previous case,

all the dimensions of the desired coefficient vector **x** will not share same distributional parameters, i.e., $p(x_i|\gamma_i) = PE(x_i; p_i, \gamma_i)$ and $p(\gamma_i) = IG(\gamma_i; q_i, q_i)$. Now collecting the terms from Q function involving p_i we get,

$$\begin{aligned} Q_1(p_i) &= E_{\gamma_i|x_i}\big[\log p(x_i|\gamma_i)\big] \\ &= \log p_i - E\big[\frac{1}{\gamma_i}\big]|x_i|^{p_i} - \frac{1}{p_i}E[\log\gamma_i] - \log\Gamma(\frac{1}{p_i}) \end{aligned} \tag{4.8}$$

Thus in M step we need to minimize,

$$L_1(p_i) = -Q_1(p_i) = -\log p_i + E\big[\frac{1}{\gamma_i}\big]|x_i|^{p_i} + \frac{1}{p_i}E[\log\gamma_i] + \log\Gamma(\frac{1}{p_i}) \tag{4.9}$$

Taking the derivative at both sides with respect to p_i we get,

$$\frac{\partial L_1}{\partial p_i} = -\frac{1}{p_i} + E\big[\frac{1}{\gamma_i}\big]|x_i|^{p_i}\log|x_i| - \frac{1}{p_i^2}E[\log\gamma_i] - \frac{1}{p_i^2}\Psi(\frac{1}{p_i}) \tag{4.10}$$

Where, $\Psi(x) = \frac{\partial\log\Gamma(x)}{\partial x}$ is the digamma function.

Similarly collecting terms involving q_i we get,

$$Q_2(q_i) = E_{\gamma_i|x_i}\big[\log p(\gamma_i)\big] \tag{4.11}$$

Thus in M step we need to minimize,

$$L_2(q_i) = -Q_2(q_i) = -q_i\log q_i + \log\Gamma(q_i) + (q_i+1)E[\log\gamma_i] + E\big[\frac{1}{\gamma_i}\big]q_i \tag{4.12}$$

Taking the derivative at both sides with respect to q_i we get,

$$\frac{\partial L_2}{\partial q_i} = -\log q_i - 1 + \Psi(q_i) + E[\log \gamma_i] + E\big[\frac{1}{\gamma_i}\big] \tag{4.13}$$

Using the conjugacy property we have already computed, $p(\gamma_i|x_i) \sim IG(q_i + \frac{1}{p_i}, |x_i|^{p_i} + q_i)$. This allows us to compute the required conditional expectations,

$$w_i = E_{\gamma_i|x_i}\big[\frac{1}{\gamma_i}\big] = \frac{q_i + \frac{1}{p_i}}{|x_i|^{p_i} + q_i} \tag{4.14}$$

$$c_i = E_{\gamma_i|x_i}[\log \gamma_i] = \log(|x_i|^{p_i} + q_i) - \Psi(q_i + \frac{1}{p_i}) \tag{4.15}$$

Finally we will take the gradient of the cost function (negative log likelihood) w.r.t the desired coefficient vector $\mathbf{x}$ i.e., $\frac{\partial L_3}{\partial \mathbf{x}}$ where,

$$\begin{aligned} L_3(\mathbf{x}) &= -E\big[\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\boldsymbol{\gamma})\big] \\ &= -E\big[-\frac{1}{2\lambda}||\mathbf{y} - \Phi\mathbf{x}||_2^2 - \sum_i \frac{|x_i|^{p_i}}{\gamma_i}\big] \\ &= \frac{1}{2\lambda}||\mathbf{y} - \Phi\mathbf{x}||_2^2 + \sum_i w_i |x_i|^{p_i} \end{aligned} \tag{4.16}$$

Hence,

$$\frac{\partial L_3}{\partial \mathbf{x}} = \frac{1}{\lambda}(\Phi^T\Phi\mathbf{x} - \Phi^T\mathbf{y}) + \boldsymbol{\theta} \tag{4.17}$$

Where $\boldsymbol{\theta}$ is a M dimensional vector and $\theta(i) = w_i p_i |x_i|^{p_i - 1}\text{sign}(x_i)$.

We will employ a gradient descent method to optimize with respect to $\mathbf{x}$, p_i and q_i using Equation (4.17), (4.10) and (4.13). Since updating distributional parameters essentially means changing the prior distribution, instead of being aggressive we propose to take a small step at the right direction, and the gradient descent method allows that.

Also in noisy conditions updating distributional parameters could be affected because of erroneous estimates of $\mathbf{x}$ during initial iterations. Hence we employ a nested gradient descent approach, where we update $\mathbf{x}$ for few iterations keeping distributional parameters fixed, which actually optimizes (4.7) for fixed distributional parameters and then updates p_i and q_i using the recent estimate of $\mathbf{x}$. Again we update $\mathbf{x}$ for few iterations with new distributional parameters and keep continuing this till convergence. The proposed algorithm has been summarized below.

Updating the distributional parameter q has been considered before in [33, 41] using a heuristic approach, whereas in this work we provide a more systematic approach of learning both p and q. Now we will analyze how our proposed updating scheme of q affects for the case when $p_i = 1\ \forall i$, i.e., Reweighted ℓ_1 norm minimization approach [30]. We revisit the computed gradient w.r.t q_i in Equation (4.13) and try to analyze the behavior of the gradient as function of x_i in Figure 4.1. This shows that if magnitude of x_i is close to origin, gradient value increases with decreasing $|x_i|$, which means to model small magnitude of $|x_i|$ (Close to zero) our approach will learn small q_i. Revisiting the weight computation step suggested in [30], $w_i^{(t+1)} = \frac{1}{|x_i^{(t)}|+q_i}$, we see that our proposed algorithm exploits different distributional parameters q_i for different dimensions, which enables to capture the relative difference between dimensions and downweights the influence of the high non zero values of x_i by also increasing its corresponding distributional parameter q_i and viceversa for small values/close to zero of x_i.

4.4 Numerical Experiments

In this section we present a set of experiments to evaluate and compare our proposed adaptive algorithm with other MAP estimation based methods with fixed distributional parameters for the task of sparse signal recovery. The experimental setup used is quite standard and has been used widely in the SSR literatures.

Data: Dictionary, Measurement, and Regularization factor: Φ, $\mathbf{y}$, λ
Output: $\hat{\mathbf{x}}$
Initialization: Initialize $\mathbf{p}, \mathbf{q}, \mathbf{x}$
for $Iter_{out} = 1$ *to max-iter* **do**
  while $Iter_{in} < block\text{-}iter$ **do**
1 Update $\mathbf{x}$ using gradient computed in Equation (4.17).
2 $Iter_{in} = Iter_{in} + 1$
3 **if** $\|\mathbf{x}_{Iter_{in}} - \mathbf{x}_{Iter_{in}-1}\|_2 < 10^{-6}$ **then**
      break;
    end
  end
4 Update p_i using gradient computed in Equation (4.10) $\forall i$.
5 Update q_i using gradient computed in Equation (4.13) $\forall i$.
end

Algorithm 1: Adaptive Bayesian Sparse Recovery Algorithm

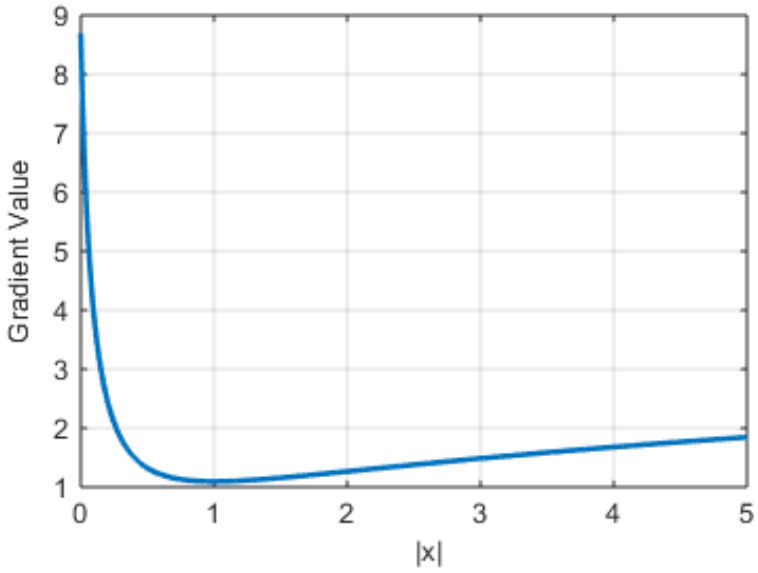

Figure 4.1. Gradient w.r.t q as a function of x

4.4.1 Problem Specification

The measurement vector $\mathbf{y}$ is generated using an $N \times M = 128 \times 256$ dictionary Φ, whose elements are generated from an i.i.d. normal distribution with mean = 0 and variance = 1. A sparse signal $\mathbf{x}$ of length 256 is generated such that $||\mathbf{x}||_0 = k$. The support, i.e. the location of the k nonzero elements, is chosen randomly, and the values are chosen from three different distributions:

(I) Uniform ± 1 random spikes. (Sub Gaussian)

(II) Zero mean unit variance Gaussian.

(III) Laplace distribution with unit variance, i.e. scale parameter = $\frac{1}{\sqrt{2}}$ (Super Gaussian)

The synthetic measurements are generated using $\mathbf{y} = \Phi\mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is the Gaussian measurement noise, whose standard deviation can be controlled for specific SNR. For all our experiments we have used SNR = 10 dB. The generated measurements and the dictionary are then provided as input to the algorithms. The estimated coefficients are compared with the original $\mathbf{x}$ that has been used to generate the measurement. Normalized Mean Square Error (NMSE) has been used as the performance metric which can be computed following,

$$NMSE = \frac{||\mathbf{x} - \hat{\mathbf{x}}||_2^2}{||\mathbf{x}||_2^2}. \tag{4.18}$$

500 trials are conducted for various fixed combinations of k, i.e. the number of non-zero coefficients, and the NMSE is plotted with respect to k. As expected, the reconstruction error (NMSE) increases as k, i.e. the cardinality of support, increases.

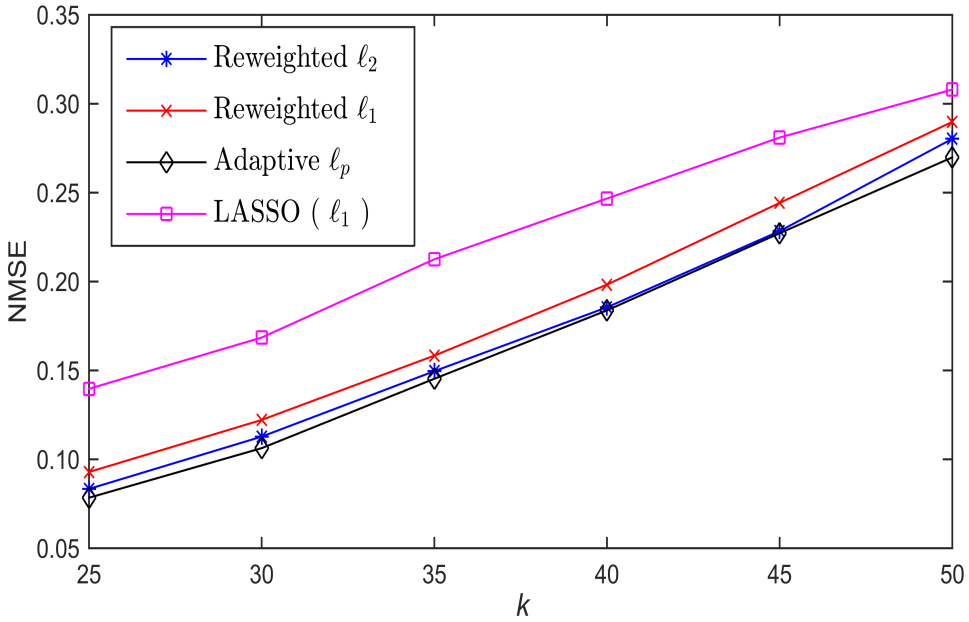

Figure 4.2. Recovery performance with Gaussian distributed non-zero coefficients

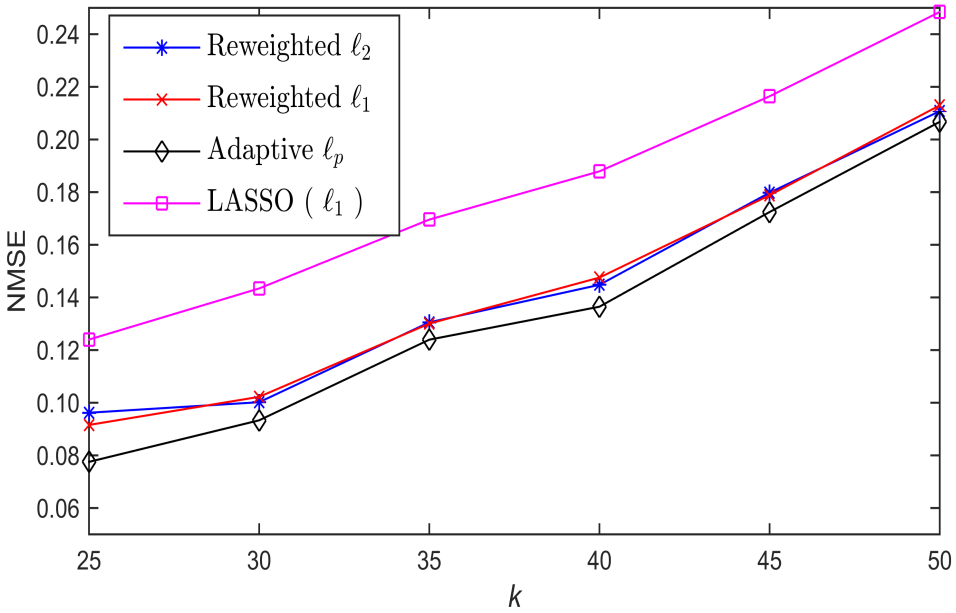

Figure 4.3. Recovery performance with super Gaussian (Laplace) distributed non-zero coefficients

4.4.2 Recovery Performance

Competing Algorithms

The main goal of this work is to show that learning the distributional parameters of the prior distribution will result in better recovery performance than the recovery algorithms with fixed priors. The algorithms in the study are:

- ℓ_1 minimization based SSR. (LASSO)
- Reweighted ℓ_1 minimization. ($\varepsilon = 0.1$ [30])
- Reweighted ℓ_2 minimization. (ε regularized, optimal update from [33])
- Adaptive reweighted ℓ_p minimization based SSR. (Proposed)

Implementation Details

All the competing algorithms including our proposed approach require an estimate of λ, which has been chosen using standard cross validation technique for all the algorithms independently. In our proposed algorithm, we have used $block - iter = 50$ and $max - iter = 5$ for all the cases. Since our goal is to capture the true prior of the non-zero elements we update p_i and q_i when $|x_i| > \delta$. In our experiments we have used $\delta = 0.01$ for all the cases.

Performance Comparison

In Figure 4.2, the average reconstruction error (NMSE) with increasing support cardinality is plotted for the case where the non-zero coefficients are from a zero mean, unit variance, Gaussian distribution. It is evident that our proposed adaptive approach gives the best performance in this case with Reweighted ℓ_2 minimization approach being a very close competitor. It is worth noting that Reweighted ℓ_2 works much better compared

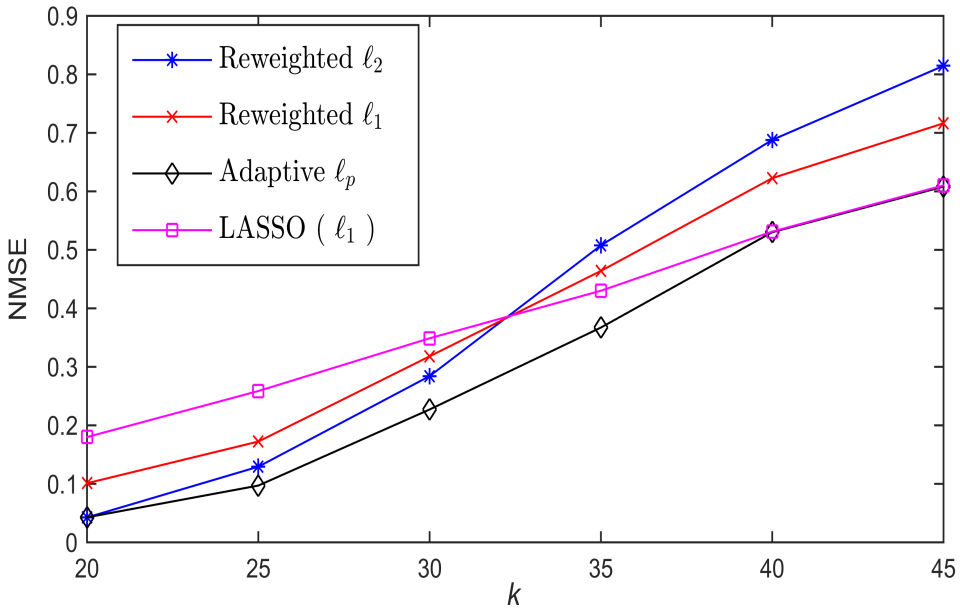

Figure 4.4. Recovery performance with Sub Gaussian distributed non-zero coefficients

to the other two methods with fixed distributional parameters, and the reason being the heuristic update of ε, which helps it to get away from local minima. Hence, ε update in Reweighted ℓ_2 is absolutely necessary as we have found out for fixed ε this algorithm's performance decreases significantly. This heuristic update of ε falls into the adaptive paradigm but our proposed approach provides a systematic approach to adapt both the distributional parameters.

In Figure 4.3, the average reconstruction error (NMSE) with increasing support cardinality is plotted where the non-zero coefficients are generated from a Laplace distribution with variance 1. Again, the empirical superiority of our proposed adaptive approach over other algorithms is evident from Figure 4.3. An interesting point to note here, is that Reweighted ℓ_1 with fixed ε matches the performance of Reweighted ℓ_2 with ε update and the reason could be that true distribution of the non-zero coefficients and the assumed prior for Reweighted ℓ_1 have the similar tail behavior.

Finally, we repeat the same set of experiments where the non-zero coefficients

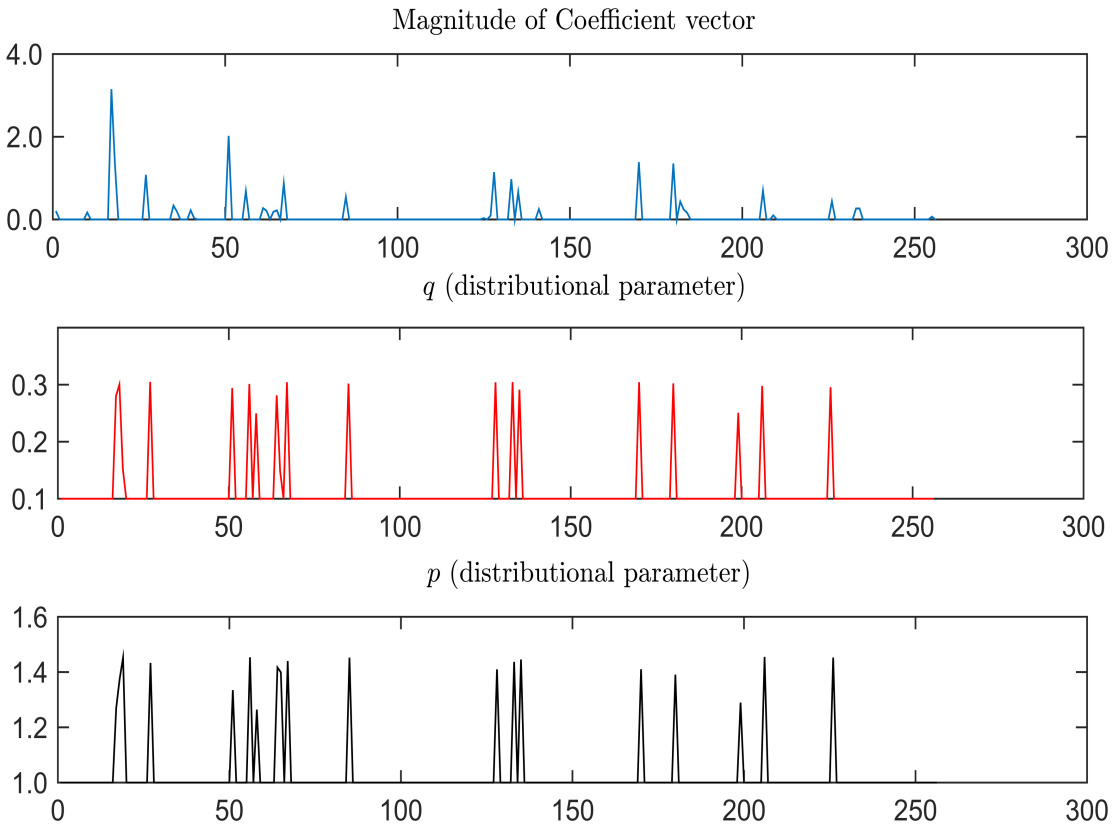

Figure 4.5. Adapted distributional parameters after convergence

follow a sub Gaussian distribution, i.e. Uniform ± 1 random spikes, and the plot of the average reconstruction error (NMSE) with increasing support cardinality is shown in Figure 4.4. The performance improvement of the proposed approach over other reweighted algorithms is significant compared to the previous two cases.

In Figure 4.5 we visualize the adaptation of the distributional parameters step. We consider a case where $k = 30$ and the non-zero entries have been randomly drawn from zero mean and unit variance Gaussian distribution. On top, the absolute value of the true non-zero coefficients have been plotted. Next two figures represent the learned values of q and p after learning them using our proposed algorithm. We see how the adaptation step helps us to learn the corresponding distributional parameters which plays a key role in modeling the tail nature of the sparsity inducing prior distributions.

4.5 Conclusion and Discussion

In this paper, we formulated the SSR problem from a Bayesian perspective and discussed a generalized scale mixture family: PESM in detail. We analyzed the tail behavior of the GT class of distributions, which is a member of PESM family, and discussed when a GT distribution will be suitable to model sparsity. We also showed, how by choosing specific distributional parameters our unified MAP estimation framework leads to several popular SSR algorithms, specifically reweighted algorithms available in the literature. Based on this result, we proposed an adaptive framework where the distributional parameters of GT have not been fixed beforehand, and adapted based on the measurement over iterations. Our extensive empirical results also show the efficacy of this adaptive approach over other MAP estimation based SSR algorithms.

4.6 Acknowledgment

The material in this chapter, in full, is based on the material as it appears in, ”Learning distributional parameters for Adaptive Bayesian Sparse signal recovery”, IEEE Computational Intelligence Magazine (2016). The dissertation author was the primary researcher and B.D. Rao supervised the research.

Chapter 5

Empirical Bayes Based Impulse Response Estimation

5.1 Introduction

The System Identification [114, 141] problem has been very well studied because of its usage in a wide spectrum of engineering applications, which include audio signal processing. Classic and well known system identification approaches are based on a parametric model assumption which can be explained using a small number of parameters [34,140]. Then, a relevant cost function related to the system prediction error is minimized to estimate the model parameters of the concerned system. Bayesian methods for system identification [12,18,24,32] have also been considered and have gained significant interest in recent years. This class of methods is useful because of its ability to incorporate any prior information of the system, i.e. the unknown Impulse Response (IR) which depends on few unknown hyperparameters. These hyperparameters can be estimated from the system measurements using an Evidence maximization approach, which is also known as Empirical Bayes (EB) method [7, 32]. Estimated hyperparameters can then be used in the Bayesian estimator to obtain an estimate of the unknown IR.

In this work, we assume that our unknown system is linear and can be modeled by an FIR, i.e. a finite number of parameters are needed to describe the system. This assumption is very much relevant for the audio signal processing applications that have been considered in this chapter.

Estimation of Room Impulse Responses (RIR's) is a very important problem in the audio community because of its huge number of applications, such as in acoustic echo cancellation (AEC) [175, 177], sound source localization [136, 154], spatial audio rendering [21, 37], and many more [85, 155]. Since modeling the room impulse response between two points in a room is extremely difficult, RIR's are directly measured in a real room environment [95]. A known sound source is played through a speaker and recorded at the desired location using a microphone. The recorded signal and source

recordings are used to estimate the desired RIR, which essentially becomes solving a deconvolution problem. Traditional cross correlation based methods have been popular for the estimation of RIR [98]. It has been shown in [74] that the least square solution without any constraints for the discrete time signals is equivalent to the conventional linear deconvolution. Since standard deconvolution methods do not exploit any any prior information, i.e. the characteristics of the RIR's, they are very sensitive to measurement noise and also suffer from poor temporal resolution [111].

These shortcomings motivate the recent line of works, where some prior information of the RIR's have been incorporated in the estimation framework to make it more robust in noisy scenarios. In [111], authors propose to make the LS solution regularized by incorporating a non negativity and sparsity constraint on the RIR's, whereas in [61], authors propose a Maximum a Posteriori (MAP) estimator for the RIR which incorporates a simple model for ambient noise and also an exponential decaying model for reverberation. In [19], authors have proposed a simultaneous estimation of RIR's using convex regularization that promotes both sparsity and an exponential amplitude envelope.

Relative Impulse Responses (ReIR's) or their frequency-domain counterparts, the Relative Transfer Functions (RTF's) [63] are also important tools in several multichannel audio processing tasks, such as speaker extraction, noise reduction, speech enhancement, and source localization etc [64, 102]. For instance, RTF information can be naturally incorporated in beamforming algorithms, where the RTF is used to design the blocking matrix of an adaptive Generalized Sidelobe Canceler (GSC) [64, 99] to cancel the target signal and produce a noise reference signal. This noise reference signal is then used later for adaptive interference cancellation and post filtering to improve the speech enhancement performance. In this work, we will focus on a two-microphone setup, and aim to estimate the ReIR between these two microphones.

In a realistic acoustic environment, reverberation has to be taken into account in GSC to achieve satisfactory signal cancellation in the output of the blocking matrix. Following this idea, Gannot et al proposed a variant of GSC named as Transfer Function-GSC (TF-GSC) [63] which relies on estimated RTFs. The performance of TF-GSC depends strongly on the quality of the RTF estimate, which is dynamic and changes with the movements of target and of microphones etc. If the RTF estimate is not updated fast enough, or if it is inaccurate, the target signal leaks through the blocking matrix and is canceled by the adaptive filtering stage, which can cause severe signal distortion at the output of GSC.

Like RIR's, ReIR's can also be easily computed in a noiseless environment using a traditional Least Squares (LS) method, as shown in [99], but the LS estimate becomes unstable in the presence of noise. There have been many recent attempts to estimate ReIR's accurately in a noisy environment [96, 125, 150, 153], but most of these solutions require a sufficiently long recording for a good estimate of ReIR (i.e., significantly more than $100 - 200$ ms). In [63], the authors have proposed a method that exploits the non-stationarity of the target speech signal. This method assumes that the noise and the RTF are stationary, or at least much less dynamic, when compared to the target signal. However, this assumption does not hold when there is a speech interferer or if the RTF is highly non stationary. In [121], the authors propose a novel assumption that the ReIRs can be replaced by sparse filters, which regularizes the LS solution. However, in reverberant environments, ReIR's will also exhibit a non-sparse decaying tail [96], which makes this approach detrimental in highly reverberant conditions. Moreover, they do not consider noisy cases. In [96], a novel approach for sparsely reconstructing time domain ReIRs from incomplete RTF measurements is proposed, where the estimation occurs only using high Signal-to-Noise Ratio (SNR) frequency bins.

Since RTF's are highly non stationary, only very short recordings can be used for

the estimation procedure. Hence, existing frequency domain approaches give a biased estimate because of the inaccuracy of the power spectral density estimate, which must be approximated by a finite average [36]. This is the main motivation of focusing on a time domain solution.

In this chapter, we propose an Empirical Bayes based estimation approach: Structured Sparse Bayesian Learning (S-SBL), where the regularization has been incorporated by exploiting the prior knowledge of the system. Similarity in the structure of both RIR's and ReIR's enable us to use our proposed approach for both the echo cancellation and blocking matrix construction tasks. Specifically, unified treatment of sparse early reflection and exponentially decaying reverberation tail in a prior distribution within an Empirical Bayesian framework is the main novelty of our work. Our approach also models ambient measurement noise and leads to a much more robust estimator of the IR. It is also important to note that, though in [19] authors have considered incorporating both the sparse early reflection and exponentially decaying tail in their estimation framework, they require prior information of the reverberation time and the regularization parameter (related to the ambient noise level), which may not be feasible in real life and crossvalidation or some other heuristic approaches are needed to choose these hyperparameters. In contrast to [19], our EB framework estimates the decaying rate and the variance of the ambient noise from the measurement using evidence maximization and eliminates the gruesome task of heuristically choosing these parameters.

The rest of the chapter is organized as follows: In Section 5.2 we introduce the problem and Section 5.3 and Section 5.4 presents the popular existing solutions to that problem in Time Domain and in Frequency Domain respectively, which will be used as our baseline. We present our proposed model along with the inference procedure in Section 5.5. In Section 5.6 we study the MSE properties of proposed estimator. Extensive experimental results over real world recordings are presented in Section 5.7 and in Section

5.8 for echo cancellation and blocking matrix construction tasks respectively. Finally Section 5.9 concludes the chapter and discusses some future directions of this work.

5.2 Problem Formulation

In this section, we will present the Impulse Response (IR) estimation problem in a more generalized setting, following a standard system identification problem. Let's consider a time invariant system, characterized by the IR $h[n]$, where n denotes the discrete time index. A measured signal $x[n]$ can be modeled as a convolution between the source signal $s[n]$ and the IR $h[n]$ with a measurement (additive) noise component $\varepsilon[n]$,

$$x[n] = (h \star s)[n] + \varepsilon[n] \tag{5.1}$$

We can rewrite this system model as a matrix vector product,

$$\mathbf{x} = \mathbf{S}\mathbf{h} + \boldsymbol{\varepsilon} \tag{5.2}$$

Where, $\mathbf{x} \in \mathbb{R}^{N\times 1}$ denotes the stacked measurement vector of size N, $\mathbf{S}$ is the convolution matrix of size $N \times L$ which is constructed using time shifted versions of $s[n]$, $\mathbf{h}$ is the IR of the systems of size $L \times 1$ and $\boldsymbol{\varepsilon}$ denotes the measurement noise vector of size $N \times 1$. For the RIR estimation case, $\mathbf{x}$ can be interpreted as the measurement recording, $\mathbf{h}$ as the true RIR, and $\mathbf{S}$ is the convolution matrix, constructed using the time shifted source signal (probe signal).

For the ReIR estimation case, we will essentially use the same model but with a slightly different convolution matrix. For this case, consider a two channel noisy recording of a target (probe signal) in a diffuse noise environment, whose position is

fixed for a certain time interval. This situation can be represented as:

$$x_L[n] = (h_L \star s)[n] + \varepsilon_L[n] \tag{5.3}$$

$$x_R[n] = (h_R \star s)[n] + \varepsilon_R[n] \approx (h_{rel} \star x_L)[n] + \varepsilon_R[n] \tag{5.4}$$

where h_L and h_R denote the impulse response between the target and the two microphones, $s[n]$ denotes the target speech, and $\varepsilon_L[n]$ and $\varepsilon_R[n]$ denote the noise components. The main problem is to estimate h_{rel}, which denotes the ReIR between the left and right microphone. The oracle solution of this problem in the time domain is, $h_{rel} = h_R \star h_L^{-1}$. To ensure that the solution is causal, a fixed delay of a few milliseconds can be introduced [97, 110], i.e., $h_{rel} = h_R \star h_L^{-1} \star \delta(n-D)$ where D is the delay in samples. The oracle RTF, denoted as H_{RTF}, which is the Fourier Transform of h_{rel}, can also be written as, $H_{RTF}(\theta) = \frac{H_R(\theta)}{H_L(\theta)}$.

We can also write the model for ReIR (5.4) as a matrix vector product,

$$\mathbf{x_R} = \mathbf{X_L}\mathbf{h} + \boldsymbol{\varepsilon_R}. \tag{5.5}$$

It is evident that from a modeling point of view the only difference is that, the convolution matrix $\mathbf{X_L}$ has been constructed using the left microphone recording $x_L[n]$.

Because of this similarity, in the following sections of this chapter we will use the notations shown in Equation (5.2), and $\mathbf{h}$ will denote both RIR and ReIR, depending on the context.

5.3 Time Domain Based Estimators

In this section, we summarize some recent popular time domain based IR estimators.

5.3.1 Traditional Least Square Solution

In a noise-free condition the size-L IR vector $\mathbf{h}$ can be estimated using a Least Square approach, i.e:

$$\hat{\mathbf{h}}_{LS} = \arg\min_{\mathbf{h}} \|\mathbf{x} - \mathbf{S}\mathbf{h}\|_2^2 \tag{5.6}$$

The solution of Equation (5.6) can be easily found by taking the pseudo-inverse:

$$\hat{\mathbf{h}}_{LS} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{x} \tag{5.7}$$

The Least Square (LS) solution without any constraint on the IR is equivalent to the conventional linear deconvolution. This LS solution can also be approximated in an online fashion using an adaptive algorithm such as the Normalized Least-Mean-Square (NLMS).

It is important to note that the above mentioned approach does not employ any constraints on the IR. For band-width limited signals the temporal resolution of linear deconvolution algorithms is limited by the near degeneracy of the columns of the convolution matrix $\mathbf{S}$. The ill-conditioning of the matrix $\mathbf{S}^T\mathbf{S}$ proves to be very detrimental for LS solution and amplifies any noise present in the system, which leads to wildly fluctuating IR estimates [111].

5.3.2 Regularizing Least Square Solution

A workaround to the ill-conditioning problem above is to use diagonal loading to make the matrix $\mathbf{S}^T\mathbf{S}$ well conditioned. The solution then becomes:

$$\hat{\mathbf{h}}_{RLS} = (\mathbf{S}^T\mathbf{S} + \alpha I)^{-1}\mathbf{S}^T\mathbf{x} \tag{5.8}$$

We can show that $\hat{\mathbf{h}}_{RLS}$ is actually the solution of the following optimization

problem:

$$\hat{\mathbf{h}}_{RLS} = \arg\min_{\mathbf{h}} \|\mathbf{x} - \mathbf{S}\mathbf{h}\|_2^2 + \alpha\|\mathbf{h}\|_2^2 \tag{5.9}$$

Subscript RLS denotes Regularized Least Square, which is essentially a ridge regression problem [126]. We will use the RLS method as one of our baseline methods with $\alpha = \frac{0.1}{L} \times \text{trace}(\mathbf{S}^T\mathbf{S})$. (Heuristic Choice)

Another option is to use the prior knowledge of the IR structure based on the statistical theory of room acoustics, to regularize the solution and make it more robust to any present noise. Specifically, sparsity constraint has been imposed on IR in several recent works [110, 121] which leads to the following optimization problem,

$$\hat{\mathbf{h}}_{\ell_1} = \arg\min_{\mathbf{h}} \|\mathbf{x} - \mathbf{S}\mathbf{h}\|_2 + \lambda\|\mathbf{h}\|_1 \tag{5.10}$$

where, $\lambda > 0$ controls the amount of sparsity in the IR $\mathbf{h}$.

Whereas in [19, 96] authors try to impose an exponentially decaying structure on IR to model the reverberation tail. To model the exponential tail, a weighted LASSO problem is solved,

$$\hat{\mathbf{h}}_{W\ell_1} = \arg\min_{\mathbf{h}} \|\mathbf{x} - \mathbf{S}\mathbf{h}\|_2 + \lambda\|\mathbf{w} \odot \mathbf{h}\|_1 \tag{5.11}$$

where, $\mathbf{w} = [w_1, ...w_L]^T$ is a vector of non negative weights and $\odot$ denotes the element wise product.

Note that, in a weighted LASSO problem, elements of $\hat{\mathbf{h}}_{W\ell_1}$ with higher weights tend to be closer to zero. Hence, to model the expected exponential decay shape of the true IR, the weights are chosen as,

$$w_i = k_1 e^{k_2|i-D|^{k_3}}, \;\; i = 1, ..., L \tag{5.12}$$

where, k_1, k_2, and k_3 are positive constants, chosen heuristically and D is the integer delay. It is evident that the weights are small near the $i = D$ where the direct path peak is expected. As i increases the weights increase and forces the corresponding elements of $\hat{\mathbf{h}}_{W\ell_1}$ to be small.

5.4 Frequency Domain Based Estimators

There have many recent works which focus on a frequency domain based estimator for Relative Transfer Function. In this section we summarize two popular approaches for RTF estimation.

5.4.1 Traditional Frequency Domain Estimation (FD)

In the Short-Time Fourier Transform (STFT) domain, assuming noiseless recordings we can rewrite Equation (5.1), assuming noiseless condition, as:

$$X(\theta,k) = H(\theta)S(\theta,k) \tag{5.13}$$

Where θ denotes the frequency bin and k denotes the frame index. A straightforward estimate of the Transfer Function (TF) can be found using:

$$\hat{H}(\theta) = \frac{\sum_k S^{\star}(\theta,k)X(\theta,k)}{\sum_k |S(\theta,k)|^2} \tag{5.14}$$

The numerator is a sample estimate of the cross Power-Spectral Density (PSD), and the denominator is a sample estimate of the auto PSD. As discussed in [63] this method produces a biased estimate. In future discussions we will refer to this method by FD and include it in our experiments as another baseline method.

5.4.2 Non-Stationarity Based FD Estimation (NSFD) [63]

This method depends on the assumption that the noise signals are stationary, or at least "less dynamic" when compared to the target speech signal. Again in the STFT domain we can represent the model as:

$$X(\theta,k) = H(\theta)S(\theta,k) + E(\theta,k) \tag{5.15}$$

Where E denotes the environmental noise. If we consider that H is static for a specific interval and divide that interval into P frames, then for the p^{th} frame:

$$\Phi^{p}_{XS}(\theta) = H(\theta)\Phi^{p}_{SS}(\theta) + \Phi^{p}_{ES}(\theta) \tag{5.16}$$

Where, $\Phi^{p}_{AB}(\theta)$ denotes the (cross) power spectral density between A and B during the pth frame. Since the noise is stationary, we can write $\Phi^{p}_{ES} = \Phi_{ES}$ and solve the overdetermined set of equations for $p = 1...P$, to estimate H. As in the FD case, in practice the PSDs in the above set of equations are replaced by their sample based estimates.

5.5 Empirical Bayes Estimator with Prior Structure

In this section we present our proposed Empirical Bayes based method to estimate the concerned impulse response in time domain by exploiting the prior structure of Room/ Relative Impulse Response. The main difference of our work from [121] is that we consider both the sparse early reflections and the reverberation tail in a unified Bayesian framework. Unlike [19, 96] we do not need any heuristic choices to control the reverb decay rate, instead our proposed algorithm treats the unknown quantities as hyperparameters and estimates them from the data. Moreover, we do not need any *a*

priori knowledge of SNR since the noise variance is also estimated within the proposed framework.

5.5.1 Model

Consider the model, $\mathbf{x} = \mathbf{Sh} + \varepsilon$, along with the Gaussian Likelihood assumption i.e, $p(\mathbf{x}|\mathbf{h}) \sim N(\mathbf{Sh}, \sigma^2)$.

We assume the following prior distribution over $\mathbf{h}$:

$$p(\mathbf{h}|\gamma_i, c_1, c_2) \sim N(0, \Gamma) \tag{5.17}$$

With:

$$\Gamma = diag\left[\gamma_1, ..., \gamma_P, c_1 e^{-c_2}, ..., c_1 e^{-c_2 m}, ..., c_1 e^{-c_2 M}\right] \tag{5.18}$$

Where:

- γ_p corresponds to p^{th} early reflection
- $c_1 e^{-c_2 m}$ corresponds to m^{th} tap out of the M exponentially decaying reverberation tail components

Note that the proposed approach follows a Relevance Vector Machine /Sparse Bayesian Learning (SBL) [157] framework to incorporate the sparse regularization.

How is Sparsity promoted?

It is not straightforward to see from the above mentioned prior distribution $p(h_i|\gamma_i) = N(h_i; 0, \gamma_i)$ for, $i = 1...P$, how the sparsity is enforced on the initial few taps of the IR, because the hierarchical nature of the prior disguises its character. To expand on this, let's assume that an Inverse Gamma (IG) distribution has been used as the prior over hyperparameters. To find the "true" nature of the prior $p(h_i)$, we integrate out the γ_i and

the marginal is obtained as,

$$\begin{aligned} p(h_i) &= \int p(h_i|\gamma_i)p(\gamma_i)d\gamma_i \\ &= \int N(h_i,0,\gamma_i)IG(\gamma_i;\alpha,\beta)d\gamma_i \\ &= \frac{\beta^{\alpha}\Gamma(\alpha+0.5)}{(2\pi)^{0.5}\Gamma(\alpha)}\left(\beta+\frac{h_i^2}{2}\right)^{-(\alpha+0.5)} \end{aligned} \tag{5.19}$$

This marginal distributions, "true" representation of the behavior of the prior of initial P taps of the IR corresponds to a Student's t-distribution, which is a super Gaussian density (has heavier tails than Gaussian) and has been very popular in sparse recovery literatures because of its ability to promote sparsity. In Figure 5.1 we present the probability density functions (pdf's) of a Student's t distribution with degrees of freedom (ν) = 0.1, and a Gaussian distribution to show why a Student's distribution is suited to promote sparsity. Moreover for our case where a uniform hyperprior $p(\gamma_i)$ has been used (i.e. $\alpha = \beta = 0$), $p(h_i) \propto \frac{1}{|h_i|}$ becomes an improper Jeffrey's prior, which has infinite probability mass at origin.

In the proposed variant of SBL we have also incorporated the reverberation tail regularization by tying the last M diagonal elements of Γ in an exponentially decaying tail. This motivates the name, Structured Sparse Bayesian Learning (S-SBL).

5.5.2 Bayesian Inference

We will follow a Type II likelihood/Evidence maximization [72, 167] procedure to estimate the Impulse Response, $\mathbf{h}$.

As shown in the previous subsection the proposed model has the following hyperparameters, γ_i where, $i = 1..P$, c_1, and c_2, which can be estimated from the data by maximizing the marginal likelihood, i.e., $p(\mathbf{x}|\gamma_i, c_1, c_2)$. The marginal likelihood is

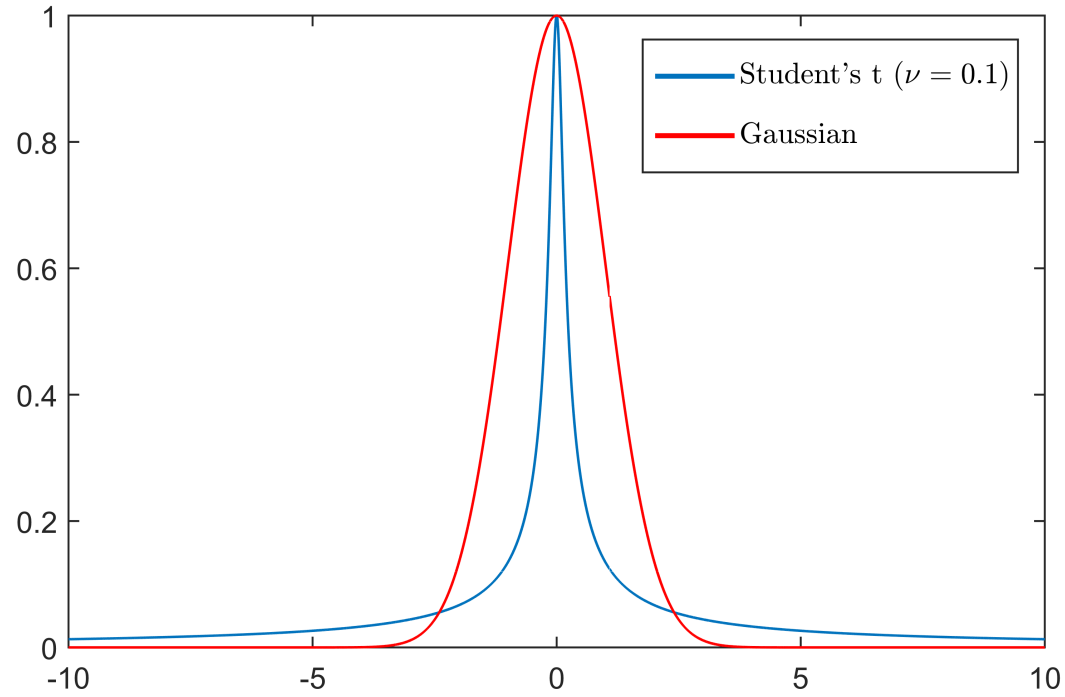

Figure 5.1. Tail Behavior: Student's t vs Gaussian

also referred to as the "evidence for the hyperparameters" by MacKay in [118], and its maximization is known as the evidence maximization or Type II method.

After estimating the hyperparameters, the estimate of the Impulse Response can be computed by,

$$\hat{\mathbf{h}} = \mathbb{E}[\mathbf{h}|\mathbf{x}, \hat{\gamma}_i, \hat{c_1}, \hat{c_2}] \tag{5.20}$$

Because of the Gaussian nature of the chosen prior, $p(\mathbf{h}; \gamma_i, c_1, c_2)$ the concerned posterior of $\mathbf{h}$ can be easily computed as,

$$p(\mathbf{h}|\mathbf{x}; \gamma, c_1, c_2) = N(\mathbf{h}; \mu, \Sigma) \tag{5.21}$$

Where

$$\mu = \sigma^{-2}\Sigma\mathbf{S}^T\mathbf{x} \tag{5.22}$$

$$\Sigma = (\sigma^{-2}\mathbf{S}^T\mathbf{S} + \Gamma^{-1})^{-1} \tag{5.23}$$

Hence, we approximate the true posterior $p(\mathbf{h}|\mathbf{x})$ by $p(\mathbf{h}|\mathbf{x};\gamma,c_1,c_2)$, which follows a Gaussian distribution whose mean and covariance depends on the estimated hyperparameters. Following Equation (5.20), we can use $\hat{\mathbf{h}} = \mu$ as the point estimate of the impulse response.

As discussed above, for the estimation of the hyperparameters we will use an evidence maximization approach, i.e:

$$\begin{aligned}
\hat{\gamma},\hat{c_1},\hat{c_2} &= \arg\max_{\gamma,c_1,c_2} p(\mathbf{x}|\gamma_i,c_1,c_2) \\
&= \arg\min_{\gamma,c_1,c_2} -2\log p(\mathbf{x}|\gamma_i,c_1,c_2) \\
&= \arg\min_{\gamma,c_1,c_2} \log\det\Sigma_x + \mathbf{x}^T\Sigma_x^{-1}\mathbf{x} + constants \\
&= \arg\min_{\gamma,c_1,c_2} J(\gamma,c_1,c_2)
\end{aligned} \tag{5.24}$$

Where, $J(\gamma,c_1,c_2) = \log\det\Sigma_x + \mathbf{x}^T\Sigma_x^{-1}\mathbf{x}$ and $\Sigma_x = \mathbf{S}\Gamma\mathbf{S}^T + \sigma^2 I_N$.

We employ the Expectation-Maximization (EM) algorithm to solve the above optimization because of its monotonic convergence property. To estimate the previously discussed hyperparameters we treat $\mathbf{h}$ as a hidden variable.

In the E step, we compute the Q function:

$$\begin{aligned}
&Q(\gamma,c_1,c_2,\sigma^2) \\
&= \mathbb{E}_{\mathbf{h}|\mathbf{x};\gamma^t,c_1^t,c_2^t,\sigma^2}[\log(p(\mathbf{x}|\mathbf{h};\sigma^2)p(\mathbf{h}|\gamma,c_1,c_2))]
\end{aligned} \tag{5.25}$$

Ignoring the terms that dont involve the concerned hyperparameters,

$$Q(\gamma, c_1, c_2, \sigma^2) = \mathbb{E}_{\mathbf{h}|\mathbf{x};\gamma^t,c_1^t,c_2^t,\sigma^2}[p(\mathbf{h}|\gamma, c_1, c_2))] = -\frac{1}{2}\mathbb{E}\left[\log\det(\Gamma) + \mathrm{tr}(\Gamma^{-1}\mathbf{h}\mathbf{h}^T)\right]$$
$$= -\frac{1}{2}\left[\sum_{i=1}^{P}[\log\gamma_i + \frac{1}{\gamma_i} < h_i^2 >] + \sum_{k=1}^{M}[\frac{e^{c_2 k}}{c_1} < h_{k+P}^2 > -c_2 k] + M\log c_1\right] \tag{5.26}$$

For iteration t we only need to compute the following conditional expectation for all taps $i \in \{1, \ldots, P+M\}$:

$$< h_i^2 >= E_{\mathbf{h}|\mathbf{x};\gamma^t,c_1^t,c_2^t,\sigma^2}[h_i^2] = \Sigma_{(i,i)} + \mu_i^2 \tag{5.27}$$

where $\Sigma_{(i,i)}$ is the i^{th} diagonal element of Σ.

In the M step, maximizing this Q-function with respect to the hyperparameters i.e, γ, c_1, c_2 and σ^2, we get:

$$\gamma_p = \Sigma_{(p,p)} + \mu_p^2 \quad \text{for } p = 1 \ldots P \tag{5.28}$$

$$c_1 = \frac{1}{M}\sum_{m=1}^{M} e^{c_2 m} < h_{m+P}^2 > \tag{5.29}$$

$$\sum_{m=1}^{M} m e^{c_2 m} < h_{m+P}^2 > -c_1\frac{M(M+1)}{2} = 0 \tag{5.30}$$

$$(\sigma^2)^{new} = \frac{\|\mathbf{x} - \mathbf{S}\mathbf{h}\|^2 + (\sigma^2)^{old}\sum_{i=1}^{M+P}(1 - \Sigma_{(i,i)}/\Gamma_i)}{N} \tag{5.31}$$

In Equation 5.29 we will use the estimate of c_2 from the previous iteration. We also need to solve Equation 5.30 to get the closed form update rule of c_2. Representing

it as a polynomial of $v = e^{c_2}$, we can show using Descartes' sign rule that there is only one positive root $\hat{v}$ of 5.30. Therefore we can update c_2 using $c_2 = \log \hat{v}$. Hence, every iteration we will update all the hyperparameters using the update rules shown above, and we can compute the point estimate $\hat{h}$ substituting the updated hyperparameters in Equation (5.22). In the following iteration we will start with the updated μ and Σ, and recompute all the hyperparameters. In practice, few iterations of the above S-SBL procedure yields a converged impulse response estimate $\mathbf{h}$.

Before moving on to experimental validation, in the next subsection we show the connection between S-SBL and the RLS methodology.

5.5.3 Connection between S-SBL and RLS

Simplifying Equation (5.22) we get,

$$\mu = (\mathbf{S}^T\mathbf{S} + \sigma^2\Gamma^{-1})^{-1}\mathbf{S}^T\mathbf{x} \tag{5.32}$$

Comparing this with Solution (5.8) we see that S-SBL can be viewed as an iterative reweighted ridge regression/reweighted ℓ_2 norm minimization algorithm, where the penalty weight factor α is not the same for all taps, and where the penalty weights are estimated every iteration through γ_i, c_1, c_2 and σ^2 which enforces the desired IR structure through regularization.

5.6 Mean Squared Error Properties of S-SBL

In this section we evaluate our proposed Empirical Bayes based S-SBL estimator of impulse response in terms of its Mean Squared Error properties. Let $\overline{\mathbf{h}}$ is the true impulse response. Then the Mean Squared Error (MSE), i.e., the expected quadratic loss

for an estimator $\hat{\mathbf{h}}$ will be,

$$MSE = \text{tr}\left[\mathbb{E}[(\hat{\mathbf{h}}-\mathbf{h})(\hat{\mathbf{h}}-\mathbf{h})^T|\gamma_i, c_1, c_2, \mathbf{h}=\overline{\mathbf{h}}]\right] \tag{5.33}$$

Now we compute the MSE of the Bayes estimator given in Equation 5.20 with fixed value of hyperparameters γ_is, c_1 and c_2, and true impulse response $\overline{\mathbf{h}}$,

$$\begin{aligned} MSE(\gamma_i, c_1, c_2) &= \text{tr}\left[\mathbb{E}[(\hat{\mathbf{h}}-\mathbf{h})(\hat{\mathbf{h}}-\mathbf{h})^T|\gamma_i, c_1, c_2, \mathbf{h}=\overline{\mathbf{h}}]\right] \\ &= \text{tr}\left[\sigma^2 f_1^{-1}(\gamma_i, c_1, c_2) f_0(\gamma_i, c_1, c_2) f_1^{-1}(\gamma_i, c_1, c_2)\right] \end{aligned} \tag{5.34}$$

Where,

$$f_1(\gamma_i, c_1, c_2) = \mathbf{S}^T\mathbf{S} + \sigma^2\Gamma^{-1} \tag{5.35}$$

and,

$$f_0(\gamma_i, c_1, c_2) = \mathbf{S}^T\mathbf{S} + \sigma^2\Gamma^{-1}\overline{\mathbf{h}}\overline{\mathbf{h}}^T\Gamma^{-1} \tag{5.36}$$

The details of the derivation of the MSE expression can be found in [8].

Our goal is to now minimize the given MSE expression with respect to the hyperparameters, i.e., γ_i, c_1 and c_2. For sake of simplicity we will assume that, $\mathbf{S}^T\mathbf{S} = I$. It is interesting to note that for a room impulse response estimation problem, if the training signal i.e, the source is white then we will get the L lag autocorrelation matrix of the source, $R_{SS} = \mathbb{E}[\mathbf{S}^T\mathbf{S}] = I$. If the length of the training sequence increases the finite signal covariance matrix $\mathbf{S}^T\mathbf{S}$ will get closer to an identity matrix.

With the assumption of $\mathbf{S}^T\mathbf{S} = I$ and substituting Γ from Equation (5.18) in the expression of MSE we get,

$$MSE(\gamma_i, c_1, c_2) = MSE_{\gamma}(\gamma_i) + MSE_{c_1,c_2}(c_1, c_2) \tag{5.37}$$

Where,

$$MSE_{\gamma} = \sigma^2 \sum_{i=1}^{P} \frac{\gamma_i^2 + \sigma^2 \overline{h}_i^2}{(\gamma_i + \sigma^2)^2} \tag{5.38}$$

and,

$$MSE_{c_1,c_2} = \sigma^2 \sum_{i=1}^{M} \frac{c_1^2 e^{-2c_2 i} + \sigma^2 \overline{h}_{P+i}^2}{(c_1 e^{-c_2 i} + \sigma^2)^2} \tag{5.39}$$

If we minimize the MSE w.r.t γ_i, it must satisfy the following optimality condition,

$$\gamma_i = \gamma_i^{MSE} = \overline{h}_i^2 \text{ for, } i = 1, ..., P \tag{5.40}$$

Now minimizing MSE_{c_1,c_2} w.r.t c_1 and c_2 by setting partial to zero,

$$\frac{\partial}{\partial c_1} MSE_{c_1,c_2} = 0 \;\; \frac{\partial}{\partial c_2} MSE_{c_1,c_2} = 0 \tag{5.41}$$

After some manipulations we get the following optimality condition for c_1,

$$\sum_{i=1}^{M} \frac{e^{-c_2 i}(c_1 e^{-c_2 i} - \overline{h}_{P+i}^2)}{(c_1 e^{-c_2 i} + \sigma^2)^3} = 0 \tag{5.42}$$

Similarly for c_2 we get,

$$\sum_{i=1}^{M} \frac{i e^{-c_2 i}(c_1 e^{-c_2 i} - \overline{h}_{P+i}^2)}{(c_1 e^{-c_2 i} + \sigma^2)^3} = 0 \tag{5.43}$$

Now considering the no noise assumption and letting $\sigma^2 \to 0$, we get,

$$\sum_{i=1}^{M} \frac{1}{e^{-c_2 i}} \left(1 - \frac{\overline{h}_{P+i}^2}{c_1 e^{-c_2 i}}\right) = 0 \tag{5.44}$$

and,

$$\sum_{i=1}^{M} \frac{i}{e^{-c_2 i}} \left(1 - \frac{\overline{h}_{P+i}^2}{c_1 e^{-c_2 i}}\right) = 0 \tag{5.45}$$

Hence the MSE estimates c_1^{MSE} and c_2^{MSE} will satisfy the above derived optimality conditions given in Equation (5.44) and (5.45).

Now as shown before, in our proposed S-SBL algorithm we are using a Type II inference technique/ evidence inference procedure where the following cost function is being minimized,

$$J(\boldsymbol{\gamma}, c_1, c_2) = \log\det\Sigma_x + \mathbf{x}^T\Sigma_x^{-1}\mathbf{x} \tag{5.46}$$

According to our measurement model if the true impulse response is $\overline{\mathbf{h}}$, then

$$\mathbf{x} = \mathbf{S}\overline{\mathbf{h}} + \boldsymbol{\varepsilon} \tag{5.47}$$

Substituting Equation (5.47) in (5.46) we get,

$$\begin{aligned} J(\boldsymbol{\gamma}, c_1, c_2) = \log\det\Sigma_x + \overline{\mathbf{h}}^T\mathbf{S}^T\Sigma_x^{-1}\mathbf{S}\overline{\mathbf{h}} \\ + \boldsymbol{\varepsilon}^T\Sigma_x^{-1}\boldsymbol{\varepsilon} + 2\boldsymbol{\varepsilon}^T\Sigma_x^{-1}\mathbf{S}\overline{\mathbf{h}} \end{aligned} \tag{5.48}$$

It can be shown that [114], for a long training sequence where the length $N \to \infty$, the minimum of the scaled function $J(\boldsymbol{\gamma}, c_1, c_2)/N$ will be the minimium of its expected value $\mathbb{E}\left[J(\boldsymbol{\gamma}, c_1, c_2)/N\right]$.

$$\begin{aligned} \mathbb{E}\left[J(\boldsymbol{\gamma}, c_1, c_2)/N\right] = \frac{1}{N}\log\det\Sigma_x + \frac{1}{N}\overline{\mathbf{h}}^T\mathbf{S}^T\Sigma_x^{-1}\mathbf{S}\overline{\mathbf{h}} \\ + \frac{1}{N}\text{tr}\left[\sigma^2\Sigma_x^{-1}\right] \end{aligned} \tag{5.49}$$

To simplify this cost function we will use the Sylvester's determinant identity,

$$\begin{aligned} |\Sigma_x| &= |\mathbf{S}\Gamma\mathbf{S}^T + \sigma^2 I_N| = (\sigma^2)^N |\frac{1}{\sigma^2}\mathbf{S}^T\mathbf{S}\Gamma + I_L| \\ &= (\sigma^2)^{N-L} |\mathbf{S}^T\mathbf{S}\Gamma + \sigma^2 I_L| \end{aligned} \tag{5.50}$$

Using this and the assumption $\mathbf{S}^T\mathbf{S} = I$,

$$\log\det\Sigma_x = (N-L)\log\sigma^2 + \log\det(\Gamma + \sigma^2 I_L) \tag{5.51}$$

We will also use the following Woodbury inverse identity,

$$\begin{aligned} \Sigma_x^{-1} &= (\mathbf{S}\Gamma\mathbf{S}^T + \sigma^2 I_N)^{-1} \\ &= \frac{1}{\sigma^2}I - \frac{1}{\sigma^2}\mathbf{S}(\Gamma^{-1} + \frac{1}{\sigma^2}\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\frac{1}{\sigma^2} \end{aligned} \tag{5.52}$$

Using these two identities and the orthonormal assumption in Equation 5.49 and collecting the terms involving γ, c_1 and c_2 we get,

$$Cost = \frac{1}{N}\sum_{k=1}^{L}\left[\log(\sigma^2 + \Gamma_{kk}) - \frac{\Gamma_{kk}(\sigma^2 + \overline{h}_k^2)}{\sigma^2(\sigma^2 + \Gamma_{kk})}\right] \tag{5.53}$$

Using the definition of Γ shown in (5.18) we split up the cost function in two parts: $Cost_\gamma$ (function of γ_i for $i = 1...P$) and $Cost_{c_1,c_2}$ (function of c_1 and c_2).

$$Cost_\gamma = \frac{1}{N}\sum_{k=1}^{P}\left[\log(\sigma^2 + \gamma_k) - \frac{\gamma_k(\sigma^2 + \overline{h}_k^2)}{\sigma^2(\sigma^2 + \gamma_k)}\right] \tag{5.54}$$

$\hat{\gamma}$ that minimizes the above cost function must satisfy,

$$\hat{\gamma}_i = \overline{h}_i^2 \text{ for, } i = 1, ..., P \tag{5.55}$$

Now,

$$Cost_{c_1,c_2} = \frac{1}{N}\sum_{k=1}^{M}\left[\log(\sigma^2 + c_1 e^{-c_2 k}) - \frac{c_1 e^{-c_2 k}(\sigma^2 + \overline{h}_{k+P}^2)}{\sigma^2(\sigma^2 + c_1 e^{-c_2 k})}\right] \tag{5.56}$$

The optimal c_1 and c_2 must satisfy the following conditions (first order),

$$\sum_{k=1}^{M} \frac{e^{-c_2 k}(c_1 e^{-c_2 k} - \overline{h}_{k+P}^2)}{(\sigma^2 + c_1 e^{-c_2 k})^2} = 0 \tag{5.57}$$

and,

$$\sum_{k=1}^{M} \frac{c_1 k e^{-c_2 k}(c_1 e^{-c_2 k} - \overline{h}_{k+P}^2)}{(\sigma^2 + c_1 e^{-c_2 k})^2} = 0 \tag{5.58}$$

Now taking the limit of $\sigma^2 \to 0$ we get the following optimality conditions,

$$\sum_{k=1}^{M} (1 - \frac{\overline{h}_{k+P}^2}{c_1 e^{-c_2 k}}) = 0 \tag{5.59}$$

$$\sum_{k=1}^{M} k(1 - \frac{\overline{h}_{k+P}^2}{c_1 e^{-c_2 k}}) = 0 \tag{5.60}$$

We observe that under the orthonormal assumption and with $\sigma^2 \to 0$, S-SBL estimates of γ_i converges to an optimal estimator in terms of its MSE. Also we can see the strong resemblance of Equation (5.59) and (5.60) to Equation (5.44) and (5.45). This resemblance suggests that the hyperparameters γ, c_1 and c_2 that will maximize the asymptotic ($N \to \infty$) evidence (marginal likelihood) in a Type II inference framework (S-SBL), will also minimize a weighted version of Mean Squared Error,

$$MSE_W = \sigma^2 \sum_{i=1}^{L} w_i \frac{\Gamma_{ii}^2 + \sigma^2 \overline{h}_i^2}{(\Gamma_{ii} + \sigma^2)^2} \tag{5.61}$$

with suitable choice of weights w_i for, $i = 1..L$. This observation has been summarized in the following Theorem.

Theorem 5.6.1 *Hyperparameters $\hat{\gamma}$, $\hat{c_1}$ and $\hat{c_2}$ that maximizes the asymptotic evidence ($N \to \infty$) in proposed empirical bayes framework when $\sigma^2 \to 0$, i.e., satisfies Equation (5.55), (5.59) and (5.60) will also minimize a weighted MSE (MSE_w), shown in Equation (5.61) where the weights are $w_i = 1, for\ i = 1,..,P$ and $w_{i+P} = e^{-\hat{c_2}i}, for\ i = 1,..,M$.*

A similar result on MSE properties of EB estimator can be found in [32], where only an exponentially decaying kernel has been considered. Whereas our work extends that result for a unified framework that incorporates both the sparse early reflections and the exponential decaying tail.

5.7 S-SBL for Echo Cancellation

In this section we present the experimental results in an Echo Cancellation (EC) task where S-SBL is used to estimate the impulse response between a speaker and a microphone present in a room. EC is usually done by adaptive systems in an online fashion to capture any changes of the concerned Room Impulse Response (RIR). But in this section, we will assume that the environment along with the positions of the speaker and microphone are not changing over time, hence we can estimate the RIR offline. This is a feasible assumption in several systems, such as in gaming consoles, where after deployment the acoustic environment is not changing over time. As pointed out in [61], in these scenarios a calibration phase could be used to estimate the RIR by playing a disguised calibration sound. Even in Adaptive Echo Cancellation (AEC) systems, an initial estimate of the channel/ RIR is required and our proposed method could be used for a better starting point.

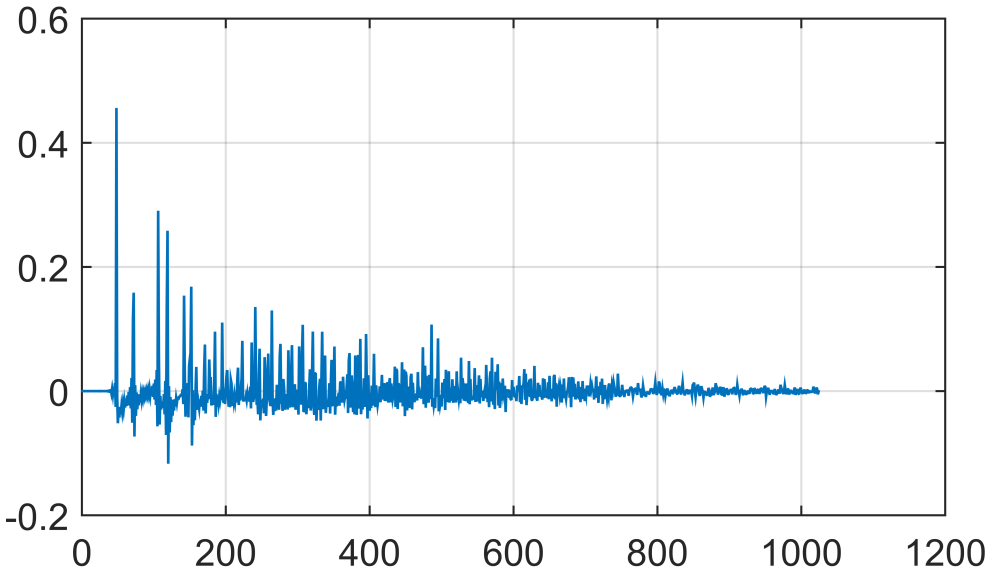

Figure 5.2. Room Impulse Response generated from Image model

5.7.1 Experimental Settings

For this echo cancellation experiment, we will use white training sequence of length 2048 samples as the source. An acoustic room impulse response (h_{true}) was calculated using the theoretical image model of a room with the parameters shown in Table 5.1. Training signal is then convolved with the true RIR to generate the measurement recording. The measurement is then corrupted using additive Gaussian noise and presented along with the source signal, to the competing algorithms to produce an estimate of the RIR ($\hat{h}$). Variance of the additive noise is controlled to present two SNR cases. Performances of the algorithms have been measure using the normalized misalignment measure, given by:

$$MSE = 10\log_{10}\frac{||h_{true} - \hat{h}||^2}{||h_{true}||^2} \tag{5.62}$$

Table 5.1. Experimental Settings

Parameters	Values
Sampling Frequency	8 kHz
Room Dimension (m)	(5, 4, 6)
Source Position (m)	(2, 3.5, 2)
Receiver Position (m)	(2, 1.5, 2)
Reverberation Time (Sec)	0.3
Number of samples	1024

Table 5.2. Misalignment Measure in Echo Cancellation

Algorithms	SNR= 0 dB	SNR= 10 dB
	MSE (dB)	MSE (dB)
RLS	+1.41	-8.16
ℓ_1 ($\lambda = 0.01$)	-2.75	-10.89
ℓ_1 ($\lambda = 0.05$)	-3.67	-5.95
ℓ_1 ($\lambda = 0.1$)	-1.74	-2.83
Weighted ℓ_1 ($\lambda = 0.01$)	+0.25	-9.16
Weighted ℓ_1 ($\lambda = 0.05$)	-2.84	-10.92
Weighted ℓ_1 ($\lambda = 0.1$)	-4.71	-10.65
S-SBL (proposed)	**-4.87**	**-11.03**

5.7.2 Competing Algorithms

- ℓ_2 regularized Least Square (RLS)
- ℓ_1 regularized Least Square (with different λ)
- Weighted ℓ_1 regularized Least Square (with different λ)
- Empirical bayes based estimator: S-SBL (Proposed)

5.7.3 Results

In Table 5.2 we tabulate the performance of all the competing algorithms for Echo Cancellation task with experimental setting discussed above with two SNR conditions at 0 dB and 10 dB. It is evident that performance of both the ℓ_1 regularization based methods depend heavily on the choice of the regularization parameter λ. Choice of λ depends on the additive noise energy. As we find out for 0 dB case $\ell_1(\lambda = 0.05)$ performs the best among other choices of λ, but for 10 dB case $\lambda = 0.01$ choice produces the lowest MSE. Similar behavior can be observed for Weighted ℓ_1 based estimator. Hence some prior knowledge on the input SNR condition is needed to choose optimal λ to achieve the best performance of ℓ_1 based methods. Whereas, for our proposed Empirical Bayes based estimator, S-SBL does not need any prior knowledge regarding the input SNR and still produces the lowest misalignment measure for both the cases, which makes our proposed estimator a robust choice compared to other competing algorithms.

5.8 S-SBL for Blocking Matrix Construction

In this section we present the detailed experimental results to evaluate several competing algorithms for relative impulse response estimation task, in terms of their target signal blocking ability.

5.8.1 Experimental Settings

We follow the experimental setting described in [96] and use the publicly available database of measured impulse responses [79] to generate the reverberant recordings. The signal for the target source, a female utterance, has been taken from the task of the online Signal Separation Campaign (SISEC) 2013 [131]. All other details are summarized below in Table 5.3.

The testing utterance (female talker) is 10 s long, which we divide into intervals

Table 5.3. Experimental Settings

Parameters	Values
Sampling Frequency	8 kHz
SNR_{in}	0 dB
Target Angle	0°
Directional Noise Angle	-60°
Microphone Pair	[3 4] (3 cm)
Distance between source and mic	2 m
T_{60}	360 ms

of 1024 samples, i.e., 128 ms at 8 kHz . Experiments are conducted on each interval independently. The average Attenuation Rate (described in the next subsection) is been reported over the intervals where speech is present. For all our experiments we use $P = 30$ for S-SBL, although we have found out that our algorithm is not very sensitive to different choices for P. We use $L = 512$ for the length of the concerned Relative Impulse Response (ReIR). Least square method was used to estimate the true ReIR using long noise free recording following [96] and the true ReIR is shown in Figure 5.3.

5.8.2 Performance Metric

To quantitatively evaluate the competing algorithms, we use a well-known and widely used performance metric called the Attenuation Rate [96].

The Attenuation Rate (ATR) can be evaluated as the ratio between SNR_{out} and SNR_{in} in dB scale, where:

$$SNR_{in} = \frac{\sum_{i=L,R}\sum_{n}[(h_i \star s)(n)]^2}{\sum_{i=L,R}\sum_{n}[\varepsilon_i(n)]^2} \tag{5.63}$$

and,

$$SNR_{out} = \frac{\sum_n [(\hat{h}_{rel} \star s_L)(n) - s_R(n)]^2}{\sum_n [(\hat{h}_{rel} \star \varepsilon_L)(n) - \varepsilon_R(n)]^2} \tag{5.64}$$

The numerator of SNR_{out} measures the leakage of the target signal whereas the denominator measures the attenuation of the noise signal. Overall, the more negative the value of ATR is, the better is the blocking performance. A low ATR indicates a good noise reference signal for further processing (such as single-channel postfiltering).

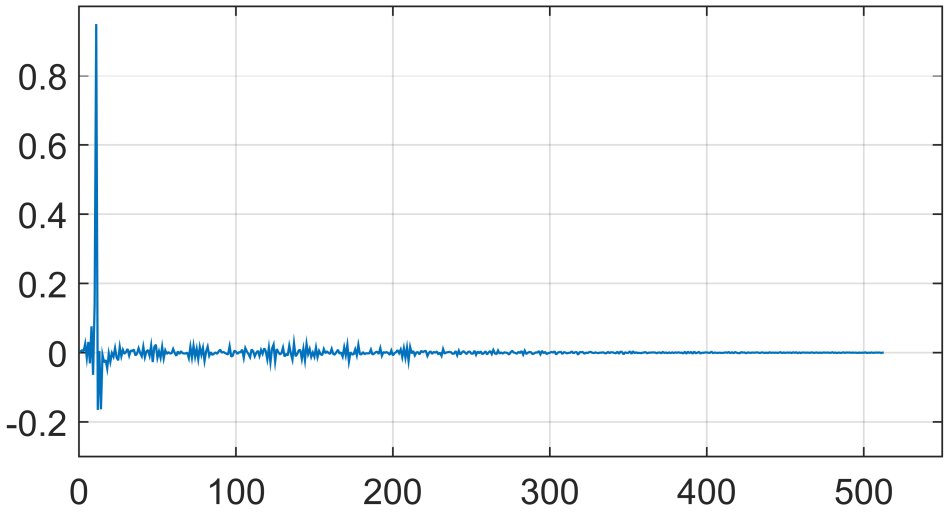

Figure 5.3. True Relative Impulse Response (ReIR)

5.8.3 Results

In this section, we present results for the diffuse noise case (white and babble) and directional noise case (white and interfering talker).

Diffuse noise

In Table 5.4 we show the average ATR obtained using all competing algorithms for two diffuse noise cases. In first case the target speech is contaminated by stationary

Gaussian noise which has been generated independently for each channel and in the second case we have used omnidirectional babble noise to contaminate the target signal. As expected, all the algorithms perform better in presence of white noise compared to the babble noise case. Next, the proposed S-SBL approach achieves the best attenuation rate for both cases, most significantly so in the babble noise case. Informal subjective listening exercises to the output of the blocking matrix also consistently show noticeable differences. Also note that the performance of the ℓ_1 based methods is very sensitive to the choice of the regularization parameter λ. We only report the ATR of the best case, i.e. using the optimum choice of λ.

Table 5.4. ATR measure in diffuse noise scenario

Algorithms	White Noise	Omni Babble Noise
	ATR (dB)	ATR (dB)
FD	-6.18	-3.68
NSFD	-11.24	-5.18
RLS	-7.36	-4.35
ℓ_1 ($\lambda = 0.05$)	-8.30	-5.59
Weighted ℓ_1 ($\lambda = 0.1$)	-11.01	-6.35
S-SBL (proposed)	**-12.05**	**-7.49**

Directional Noise

In Table 5.5 we present the average ATR obtained using all competing algorithms in directional noise. Specifically, in the first case the target speech is contaminated by directional Gaussian noise generated following the experimental setting discussed above, and in the second case we have used a male speaking interferer. This situation is more challenging compared to diffuse noise, even more so when the directional noise is a speech interferer. The performance of all the algorithms is reduced in directional white noise when compared with diffuse white noise. In Figure 5.4, 5.5 and 5.6 we show the

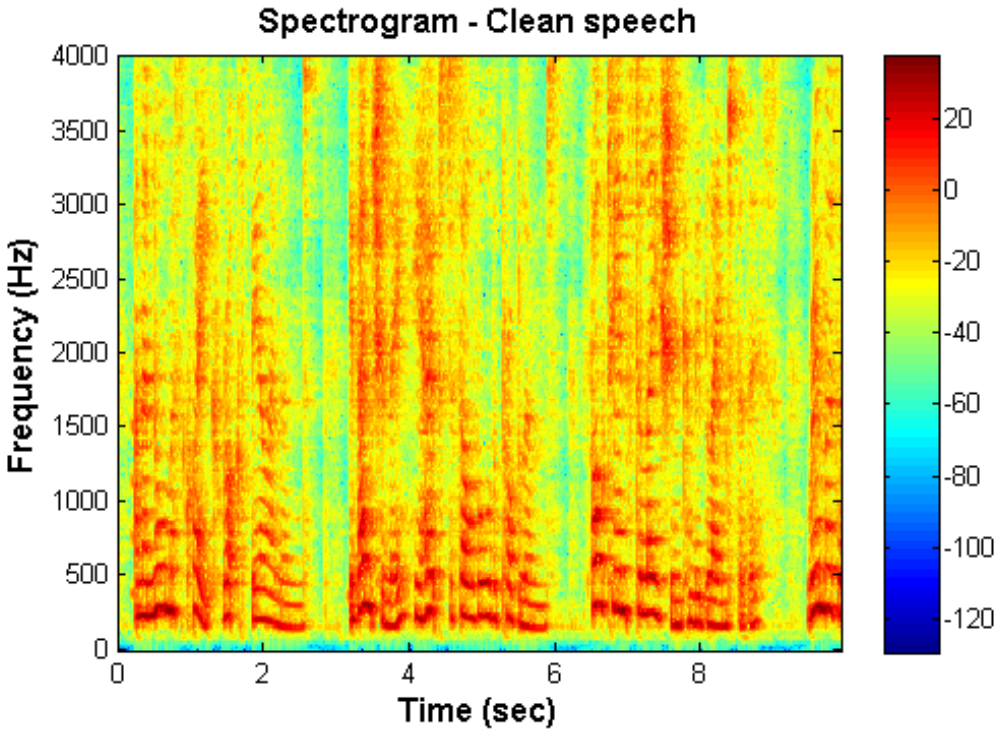

Figure 5.4. Spectrogram of clean utterance recorded at left mic

spectrograms of the clean speech and the noise reference signal obtained using S-SBL and NSFD, respectively, in the case of directional white noise. It is evident from Figure 5.6 that dominant low-frequency speech harmonic structure is still present in the NSFD noise reference estimate. For a speech interferer, when there is no Voice Activity Detection (VAD) all algorithms struggle and often produces positive ATR. The main reason behind this result is that the RTF estimate could be that of the speech interferer, since there is no way to distinguish who is the desired target. Hence, we present results assuming that an Oracle VAD is available for both target and the interferer. Hence, the ReIR is only computed when the target is present but not the interferer. In real life scenarios, an oracle VAD can be substituted by a VAD operating on a close talk microphone recording, or a phone microphone recording. We have conducted such experiments using the database presented in [170] and the results are encouraging. It iss evident from Table 5.5 that

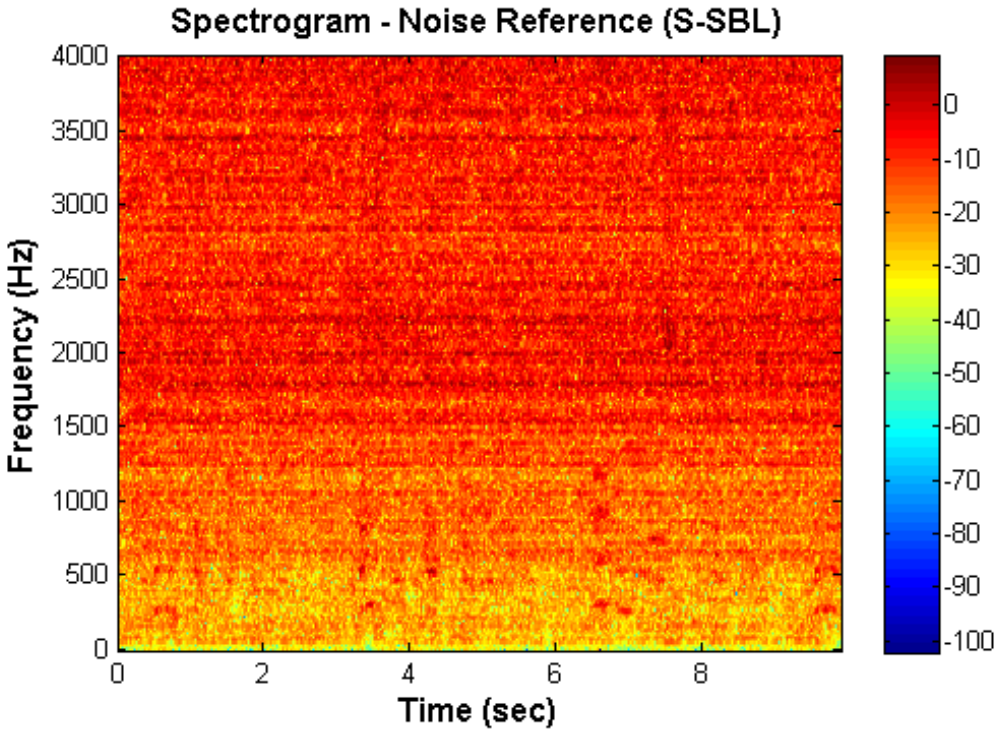

Figure 5.5. Spectrogram of the noise reference signal obtained using S-SBL (Directional white noise)

even in presence of directional noise sources, S-SBL cancels out the target efficiently compared to other competing algorithms.

Table 5.5. ATR measure in presence of directional noise

Algorithms	White	Talker (with VAD)
	ATR (dB)	ATR (dB)
FD	-3.98	-0.86
NSFD	-10.37	-9.63
RLS	-7.25	-11.40
ℓ_1 ($\lambda = 0.05$)	-8.66	-6.76
Weighted ℓ_1 ($\lambda = 0.1$)	-10.39	-11.22
S-SBL (proposed)	**-10.79**	**-15.72**

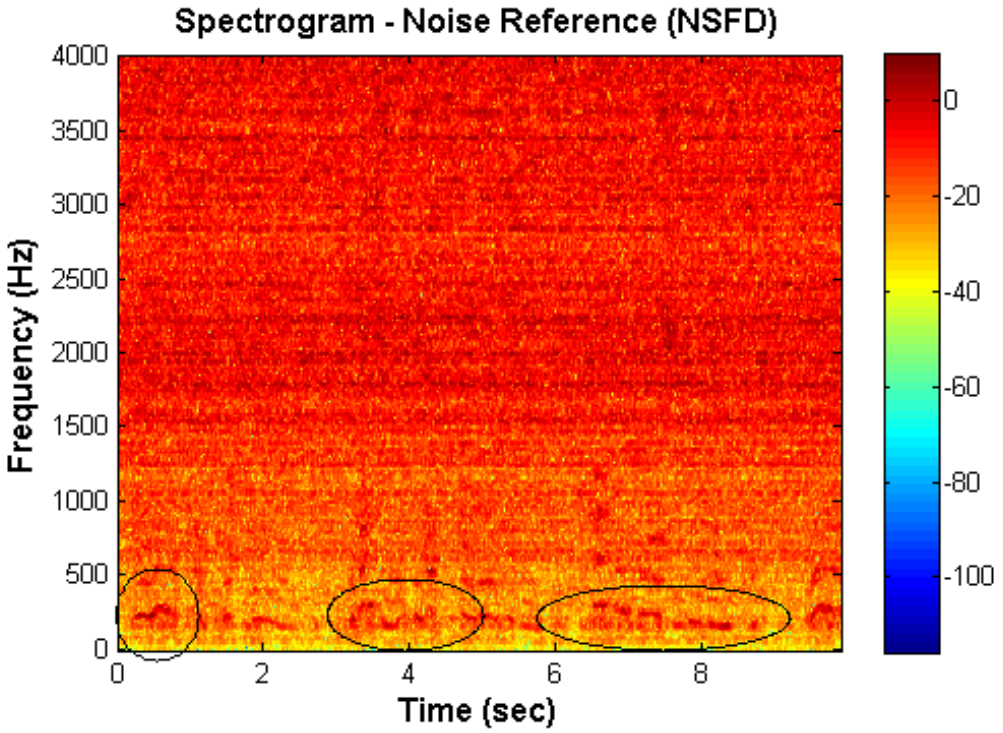

Figure 5.6. Spectrogram of the noise reference signal obtained using NSFD (Directional white noise)

5.8.4 Effect of Recording Length

In Figure 5.7 we present how the increase of recording length affects the performance of all the competing algorithms for a diffused noise case (omni babble). As expected performance of all the algorithms slightly improves with a growing recording length. Same experiment is repeated for directional white noise case with varying recording lengths, and the results are presented in Figure 5.8. Similar behavior, i.e. better performance with growing length of the recording is noticed in this case too.

Though the longer recordings improve the ReIR estimation performance, in real life the dynamic nature of ReIR may prove to be a hindrance. Because the surrounding acoustic environment along with the positions of the target and the microphones may not remain same during the recording length and may change, which results in change in

ReIR. Hence the ability of our proposed approach, S-SBL, to estimate ReIR using very short and noisy recordings makes it a useful choice.

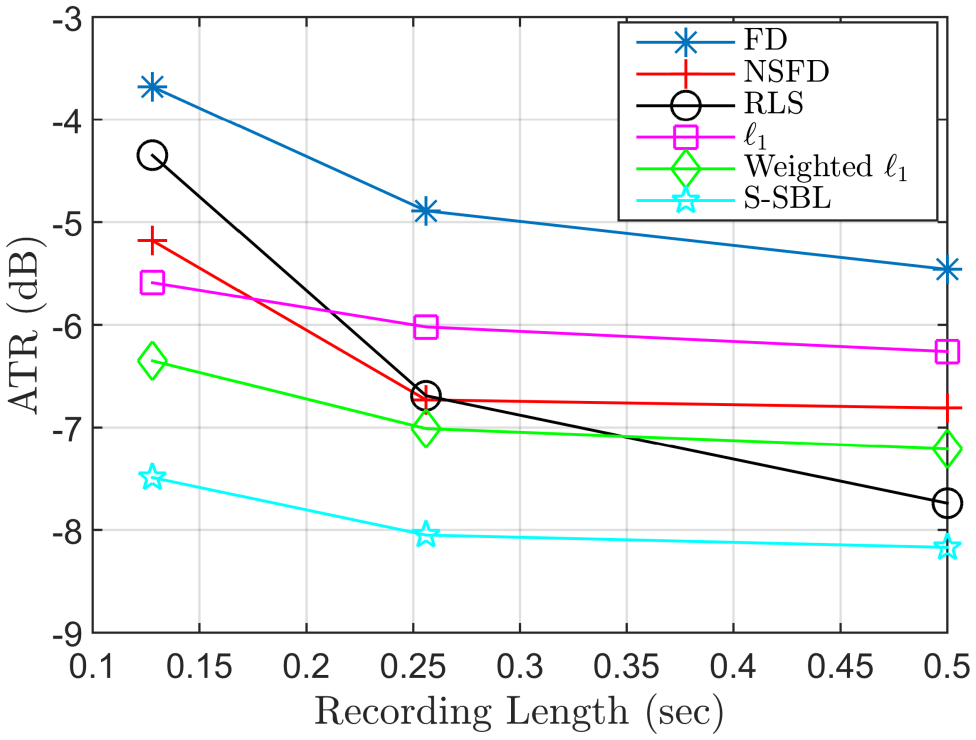

Figure 5.7. Attenuation Rate vs Length of the recording in presence of omnidirectional babble noise

5.9 Conclusion

We proposed a novel Bayesian approach of estimating room/ relative impulse response using short, noisy, reverberant recordings. Our proposed time domain solution benefits from exploiting the prior IR structure by employing both sparsity inducing prior for early reflection and exponentially decaying kernel for reverberation tail, during estimation. We also analyze the MSE properties of our estimator and show that the evidence maximization procedure can also be interpreted as a weighted MSE minimization problem. Detailed experimental results also show consistent improvement of our pro-

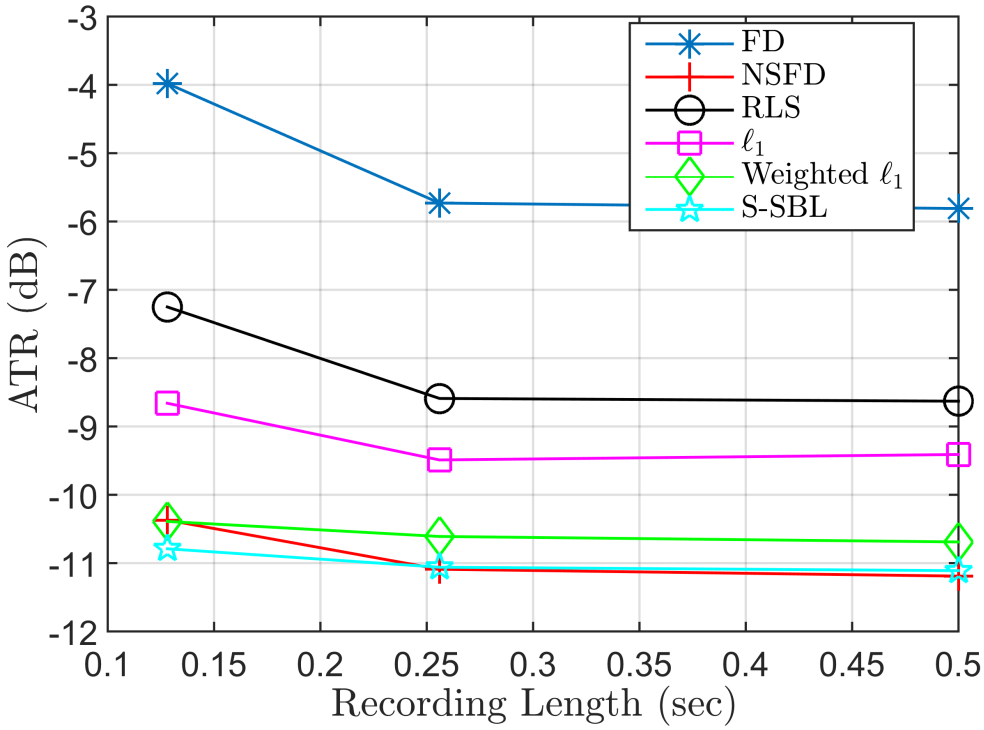

Figure 5.8. Attenuation Rate vs Length of the recording in presence of directional white noise

posed approach over competing algorithms. Incorporating this relative impulse response estimation technique in a generalized sidelobe canceller structure to improve the binaural noise suppression performance will be considered in our future works.

5.10 Acknowledgment

The material in this chapter is, in part, a reprint of material is in preparation for submission under the title, ”Empirical Bayes based Relative/ Room Impulse Response Estimation” and also based on the material as it appears in, ”Dynamic Relative Impulse Response Estimation using Structured Sparse Bayesian Learning”, 41st IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2016. In both cases, dissertation author was the primary author, Fred Mustiere contributed to the

research, and Tao Zhang and B.D. Rao supervised the research.

Chapter 6

Reweighted Algorithms for Independent Vector Analysis

6.1 Introduction

Blind Source Separation (BSS) addresses the problem of recovering original sources from the mixture with the assumption that the mixing process is linear [86, 104, 152] . A generalization of this problem, Joint Blind Source Separation (JBSS) has gained lot of interest in recent times because of its wide range of engineering applications in different fields [94, 103, 107]. For example JBSS has proven to be useful in speech separation [80, 94], functional magnetic resonance (fMRI) [103, 107] and electroencephalograph (EEG) data analysis [106, 117] etc. In these problems, the key assumption is that, each latent source has dependencies across the datasets but independent of all other sources within a dataset [1]. JBSS solutions exploit this higher order dependencies of sources across the datasets to achieve reliable source separation and also solves the permutation ambiguity that arises from using BSS algorithms individually [94, 103].

One particular formulation of JBSS is known as Independent Vector Analysis (IVA) [94], which is an extension of the well studied BSS algorithm, Independent Component Analysis (ICA). IVA exploits the higher order dependencies of a SCV across datasets by employing a dependent multivariate source prior distribution instead of independent univariate distributions, which is the case for ICA [5]. By this modeling approach, it imposes inter-vector source independence and also preserves the higher order intra-vector source dependencies across datasets. It also removes any permutation ambiguity during the learning process and does not need any pre or post processing step [5, 103].

IVA was originally introduced for the speech separation task in [94], where it assumed a multivariate Laplace distribution as the Source Component Vector (SCV) distribution and no correlation within the SCVs. We will denote this implementation of

IVA as IVA-L. In some applications the assumption of no second order dependencies within the SCVs can seriously degrade the source separation ability. For example in group fMRI studies, SCVs are expected to have significant correlation as shown in [107, 148]. This acted as a motivation behind employing a Multivariate Gaussian distribution instead of an isotropic uncorrelated multivariate Laplace distribution as source prior and source separation performance improvement was noticed for the correlated SCV case [3, 6]. This implementation of IVA will be denoted as IVA-G. But this approach is not robust enough as it fails to separate sources if the source correlation is not significant, because of the nonidentifiability condition discussed in [3].

Recently, selecting the appropriate multivariate source prior to improve the separation performance has become a research focus. In [109, 143] authors have proposed to use a multivariate student's t distribution as a source prior and have shown performance improvement in a speech separation task over the original IVA implementation (IVA-L). Whereas, in [4, 108] authors have proposed a new variant of IVA using a multivariate generalized Gaussian distribution as source prior with a specific value for the shape parameter. In [133] authors have employed a Complex Gaussian Scale mixture as the source prior for this task.

In this work we propose a generalized scale mixture distribution, Multivariate Power Exponential Scale Mixture (M-PESM) as the source prior for IVA. Recently both Multivariate and univariate Gaussian scale mixtures (GSM) [14, 101, 132, 157] and Laplacian scale mixtures (LSM) [65] have gained lot of interest because of their ability to represent heavytailed distributions. The M-PESM representation includes the popular M-GSM and M-LSM as special cases and provides a mechanism to present a unified view. This class of distributions also helps us to exploit both the higher order (greater than second order) dependencies within a SCV (unlike IVA-G) and also any intra-source correlation (second order dependency), present across the datasets (unlike IVA-L). We

also show that both IVA-L and IVA-G are special cases of the unified framework. By employing a specific member (Multivariate Generalized t distribution) of M-PESM as the source prior, our unified framework leads to two Reweighted algorithms for IVA.

The rest of the chapter is organized as follows. In Section 6.2, we review the IVA framework formulation for JBSS. In Section 6.3, a generalized scale mixture representation, the Multivariate Power Exponential Scale Mixtures (M-PESM) family, is presented. In Section 6.4, we derive a unified Maximum Likelihood based inference framework for IVA by employing a multivariate source prior from the family of M-PESM. We also discuss some special cases of the unified framework and establish connections with current algorithms in the literature. We present experimental results of the proposed algorithms in Section 6.5, in different settings and finally conclusions and some future directions of this work are presented in Section 6.6.

6.2 Independent Vector Analysis (IVA): Problem Formulation

In this section we formulate the Joint Blind Source Separation (JBSS) framework of interest, i.e., Independent Vector Analysis. Let there be K datasets, with each having been formed from a distinct linear mixture of N independent sources. Mixing model for the t^{th} observation $\mathbf{x}_t^{[k]}$ among T iid observations is given by,

$$\mathbf{x}_t^{[k]} = \mathbf{A}^{[k]}\mathbf{s}_t^{[k]} \;\; 1 \leq k \leq K,\; 1 \leq t \leq T \tag{6.1}$$

Where, $\mathbf{s}_t^{[k]} = \left[s_{1,t}^{[k]}, ..., s_{N,t}^{[k]}\right]^T$ is a zero mean source vector, and $\mathbf{A}^{[k]} \in \mathbb{R}^{N\times N}$ is the invertible mixing matrix. IVA assumes that the sources are mutually independent and the n^{th} Source Component Vector (SCV) can be written as, $\mathbf{s}_{n,t} = \left[s_{n,t}^{[1]}, ..., s_{n,t}^{[K]}\right]^T$. Due to the independent assumption, the joint distribution of all the sources reduces to the product

of the N (number of sources) SCV distributions, i.e, $p(\mathbf{s}_{1,t},...,\mathbf{s}_{N,t}) = \prod_{n=1}^{N} p(\mathbf{s}_{n,t})$. IVA solves the blind source separation problem by finding the K demixing matrices $\mathbf{W}^{[k]}$ and the source vector estimates for each datasets,

$$\mathbf{y}_t^{[k]} = \mathbf{W}^{[k]}\mathbf{x}_t^{[k]} \;\; 1 \leq k \leq K,\; 1 \leq t \leq T \tag{6.2}$$

and the n^{th} SCV estimate can be written as, $\mathbf{y}_{n,t} = \left[y_{n,t}^{[1]},...,y_{n,t}^{[K]}\right]^T$. It is important to note that the mixing matrixes, $\mathbf{A}^{[k]}$ for each dataset (each k) are distinct and not related to each other. Another key assumption of IVA is that a source within one dataset is dependent on at most one source in another dataset which enables to solve the permutation ambiguity of BSS. But the scaling ambiguity still remains, which can be removed by assuming that the sources have unit variance and scaling the demixing vectors, $\mathbf{w}_n^{[k]}$, to estimate unit variance sources. Assumption of higher order dependencies can be realized by modeling each SCV with a multivariate dependent probability distribution, also known as source prior distribution. In the seminal work of IVA [94], a multivariate Laplace distribution was used as the source prior to capture the dependencies within a SCV and across datasets. In this article we propose a class of multivariate generalized scale mixture distribution family as source prior which significantly enriches the type of sources that can be dealt with as well as leads to the development of a rich and general classs of algorithms for source separation.

6.3 Source Prior: Multivariate Scale Mixtures

Scale mixture distributions namely Gaussian Scale mixtures (GSM) and Laplacian Scale mixtures (LSM) have gained lot of attention in recent years because of their ability to represent complex heavy tailed super gaussian distributions in a simple hierarchical manner [65, 101, 132]. In recent works of compressed sensing and sparse recovery, scale

mixtures have been used as prior distribution to model sparsity. In [14] Multivariate GSM (M-GSM) has been used to model group sparsity in a sparse recovery problem. For JBSS, in [94] authors have shown that Multivariate Laplace distribution is a special case of M-GSM. In this work, a more general scale mixture family, Multivariate Power Exponential Scale Mixture (M-PESM), which is a generalization of M-GSM and M-LSM, is presented and has been used as a source prior for JBSS.

6.3.1 Multivariate Power Exponential (M-PE)

Power exponential (PE) distributions were first introduced by Box and Tiao (1962) in the context of robust regression to deal with non-normality. In this work we are concerned with the Multivariate PE (M-PE) distribution, which is also known as Generalized Gaussian Distribution (GGD) and has received lot of attention in the literature. The probability density function of a multivariate PE (M-PE) is defined by [138],

$$p(\mathbf{x}|\mathbf{M},\beta,z) = \frac{1}{|\mathbf{M}|^{1/2}} h_{\beta,z}(\mathbf{x}^T\mathbf{M}^{-1}\mathbf{x}) \tag{6.3}$$

for any $\mathbf{x} \in \mathbb{R}^{L\times 1}$, where $\mathbf{M}$ is a $L \times L$ symmetric real correlation matrix, and $h()$ is known as the density generator defined by,

$$h_{\beta,z}(y) = \frac{\beta\Gamma(\frac{L}{2})}{\pi^{\frac{L}{2}}\Gamma(\frac{L}{2\beta})z^{\frac{L}{2\beta}}} \exp\left(-\frac{y^\beta}{z}\right) \tag{6.4}$$

Where, $z > 0$ is the scale parameter and $\beta > 0$ is the shape parameter of the Multivariate PE.

It is evident from the above given form, that $\beta = 1$ results in the Multivariate Gaussian distribution, whereas $\beta = 1/2$ connects to the well known Multivariate Double exponential or Laplace distribution. $\beta < 1$ leads to distribution with heavier tails than the Gaussian distribution (super Gaussian density). It is interesting to note that, when the

correlation matrix, $\mathbf{M} = I$, M-PE (joint pdf) becomes a function of the ℓ_2 norm of the multidimensional random variable, hence a spherically symmetric distribution.

6.3.2 Multivariate PESM (M-PESM)

Multivariate PESM family of distributions refer to distributions that can be represented as follows:

$$p_{\mathbf{X}}(\mathbf{x}) = \int p_{\mathbf{X}|z}(\mathbf{x}|z)p_z(z)dz = \int \text{M-PE}(\mathbf{x};\mathbf{M},\beta,z)p_z(z)dz \tag{6.5}$$

Choice of distributional parameter β along with different suitable mixing densities, i.e., $p_z(z)$, will lead to different marginalized prior distributions including the super Gaussian distributions.

Some special cases of M-PESM includes Multivariate Gaussian Scale Mixtures (M-GSM) when shape parameter $\beta = 1$, Multivariate Laplace Scale Mixtures (M-LSM) when shape parameter $\beta = 1/2$, Multivariate Uniform Scale Mixtures (M-USM) when $\beta \to \infty$.

6.3.3 When M-PESM?

Here we will try to address the question as to when a multivariate spherically symmetric distribution can be represented as M-PESM. We use the result dealing with integral representations, discussed in [162] to answer this question. This result can also be viewed as an extension of the result provided for M-GSM in [101] to the general M-PESM case.

Theorem 6.3.1 *A distribution $p_{\mathbf{X}}(\mathbf{x}) = f(||\mathbf{x}||_2)$, which is spherically symmetric about the origin has a Multivariate Power Exponential Scale Mixture (M-PESM) representation with shape parameter β and scatter matrix $\mathbf{M} = I$, if and only if $g(r) = f(r^{\frac{1}{2\beta}})$, where $r =$*

$||\mathbf{x}||_2$, *is completely monotone.*

Proof:

To prove our result we will use the following definition of completely monotone [23],

Lemma 6.3.2 *A function $f(x)$ is completely monotonic on $(0,\infty)$ if, $(-1)^n f^{(n)}(x) \geq 0,\ \ n = 0,1,..$ for every $x \in (0,\infty)$.*

Using this definition the following result on monotonicity, also known as Bernstein's theorem, was established in [162],

Theorem 6.3.3 *A necessary and sufficient condition that $f(x)$ should be completely monotonic on $(0,\infty)$ is that, $f(x) = \int_0^\infty e^{-zx} d\alpha(z)$, where $\alpha(z)$ is non-decreasing on $(0,\infty)$.*

Now to prove the first part of Theorem 6.3.1, lets assume that X is a spherically symmetric random vector of dimension L with a distribution $p_{\mathbf{X}}(\mathbf{x}) = f(||\mathbf{x}||_2)$ which has a M-PESM representation. Hence,

$$\begin{aligned} f(||\mathbf{x}||_2) &= \int_0^\infty \text{M-PE}(\mathbf{x}; I, \boldsymbol{\beta}, z) d\alpha(z) \\ &= \int_0^\infty \frac{\beta \Gamma(\frac{L}{2})}{\pi^{\frac{L}{2}} \Gamma(\frac{L}{2\beta}) z^{\frac{L}{2\beta}}} \exp\left(-\frac{r^{2\beta}}{z}\right) d\alpha(z) \end{aligned} \tag{6.6}$$

Where, $r = ||\mathbf{x}||_2$ and $\alpha(z)$ could be interpreted as the cumulative distribution function (CDF) of the scale mixing density. Let,

$$g(r) = f(r^{1/2\beta}) = \int_0^\infty \frac{\beta \Gamma(\frac{L}{2})}{\pi^{\frac{L}{2}} \Gamma(\frac{L}{2\beta}) z^{\frac{L}{2\beta}}} \exp\left(-\frac{r}{z}\right) d\alpha(z) \tag{6.7}$$

Hence from the definition of completely monotone, its straightforward to see that $g(r)$ is completely monotonic on $(0,\infty)$.

Conversely, suppose $g(r)$ is completely monotone on $(0,\infty)$. Hence from Bernstein's theorem,

$$g(r) = \int_0^\infty \exp\big(-zr) d\alpha(z) \tag{6.8}$$

for some non decreasing $\alpha(z)$ on $(0,\infty)$. Hence, we get a M-PESM representation,

$$p_{\mathbf{X}}(\mathbf{x}) = f(||\mathbf{x}||_2) = g(r^{2\beta}) = \int_0^\infty \exp(-zr^{2\beta}) d\alpha(z) \tag{6.9}$$

This completes the proof.

6.3.4 Example of M-PESM: Multivariate Generalized t Distribution (M-GT)

In this example, we will consider an inverse gamma (IG) distribution as our mixing density $p_z(z) = IG(q,q)$, where $IG(x;a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\big(-\frac{b}{x}\big)$ in the hierarchical representation (6.5) for the M-PESM family. It leads to a multivariate generalized t distribution [11] which is a superset of all the multivariate source priors that have been used in practice in several recent works, e.g. Multivariate Gaussian, Multivariate Laplace and Multivariate Student's t distributions, among others.

The Multivariate Generalized t Distribution has the form:

$$p_{\text{M-GT}}(\mathbf{x};q,\beta,\mathbf{M}) = \frac{\eta}{(q+s^\beta)^{q+\frac{L}{2\beta}}} \tag{6.10}$$

Where $s = \mathbf{x}^T\mathbf{M}^{-1}\mathbf{x}$, η is the normalization constant. Interestingly, β and q provide the flexibility to represent different tail behavior using this distribution. Larger values of β and q correspond to thin tailed distributions whereas smaller values of β and q are associated with heavy tailed distributions.

Some special case of note is $\beta = 1$, which leads to a Multivariate student's t

Table 6.1. Variants of Multivariate GT distribution

q	β	Source Prior	IVA Algorithm
$q \to \infty$	1	M-Gaussian	IVA-G [3, 6]
$q \to \infty$	1/2	M-Laplace	IVA-L [94]
$q \geq 0$ (degrees of freedom)	1	M-Student's t distribution	IVA-Reweighted ℓ_2
$q \geq 0$ (shape parameter)	1/2	M-Generalized Double Pareto	IVA-Reweighted ℓ_1

distribution, a prior that has been used in MMV version of the popular Sparse Bayesian Learning (SBL)/Relevance Vector Machine (RVM) work [157] and can be decomposed as a Multivariate Gaussian Scale mixture with inverse Gamma as the mixing density. Employing $\beta = 1/2$ leads to a Multivariate Generalized Double Pareto distribution (GDP) which can be represented as a scale mixture of Multivariate Laplace following equation (6.5). Univariate version of GDP has been discussed in [10] in details. In Table 6.1, we summarize some special cases of Multivariate GT that have been used in literature as multivariate source prior for IVA that arise by different choices of the shape parameters of M-GT, i.e. β and q (With $\mathbf{M} = I$).

Among Scale Mixtures, both univariate and multivariate GSM in particular have gained a lot of interest over the years in the literature and the proposed M-PESM framework is an interesting generalization as, M-GSM, M-LSM and M-USM are subsets of the proposed M-PESM. As shown in [132], M-GSM can only be used to represent supergaussian densities, i.e. distributions with positive kurtosis whereas M-PESM representation can also be used for subgaussian densities along with supergaussian densities. One example is the previously discussed M-GT distribution, which becomes a thin tailed subgaussian distribution for $\beta > 1$ and $q = 1$.

6.4 Maximum Likelihood: IVA inference using EM

In this section, we derive the inference procedure by maximizing the likelihood of the observations using an EM algorithm, where a member of M-PESM family has been employed as the multivariate source prior, $p(\mathbf{s}_{n,t}) = \int \text{M-PE}(\mathbf{s}_{n,t}; \mathbf{M}_n, \beta, z_{n,t}) p(z_{n,t}) dz_{n,t}$ is a M-PESM.

A well known result regarding the density of linear transformation will be used in the derivation which says, if $\mathbf{s}$ is a random vector and has probability distribution $p_{\mathbf{S}}(\mathbf{s})$, then the density of $\mathbf{x} = \mathbf{A}\mathbf{s}$ is,

$$p_{\mathbf{X}}(\mathbf{x}) = |\det(\mathbf{W})| p_{\mathbf{S}}(\mathbf{W}\mathbf{x}) \tag{6.11}$$

Where $\mathbf{W} = \mathbf{A}^{-1}$.

We will employ an EM algorithm while treating $z_{n,t}$ as the hidden variable. Considering T i.i.d observations, the complete data log likelihood becomes,

$$\begin{aligned} L(\mathbf{W}) &= \sum_{t=1}^{T} \log p(\mathbf{x}_t^{[1]}, ..., \mathbf{x}_t^{[K]}, \mathbf{z}_t), \text{ Where, } \mathbf{z}_t = [z_{1,t}, ..., z_{N,t}] \\ &= \sum_{t} [\sum_{k=1}^{K} \log |\det(\mathbf{W}^{[k]})| + \sum_{n=1}^{N} (\log p_{\mathbf{s}_{n,t}|z_{n,t}}(\mathbf{y}_{n,t}|z_{n,t}) \\ &+ \log p(z_{n,t}))] \end{aligned} \tag{6.12}$$

To compute the Q function we need the conditional expectation of the log likelihood with respect to the posterior of the hidden variables. Since in the M step we will maximize the Q function w.r.t $\mathbf{W}^{[k]}$, we can ignore the last term $\log p(z_{n,t})$. Only the second term has dependencies on the hidden variable, hence in E step we are only concerned with this

term, i.e.,

$$\sum_{n=1}^{N} \log p_{\mathbf{s}_{n,t}|z_{n,t}}(\mathbf{y}_{n,t}|z_{n,t}) = -\sum_{n=1}^{N} \frac{(\mathbf{y}_{n,t}^T \mathbf{M}_n^{-1} \mathbf{y}_{n,t})^{\beta}}{z_{n,t}} + \text{Constants} \tag{6.13}$$

E step essentially becomes computation of $\mathbb{E}_{z_{n,t}|\mathbf{s}_{n,t}}\left[\frac{1}{z_{n,t}}\right]$. The derivation of the concerned conditional expectation where a M-GT (a member of M-PESM) has been employed as the source prior uses a similar trick that has been used in [132]. Details of the derivation are given in the Appendix (Supplemental material). The weights are found as,

$$w_{n,t} = \mathbb{E}_{z_{n,t}|\mathbf{s}_{n,t}}\left[\frac{1}{z_{n,t}}\right]_{\mathbf{s}_{n,t}=\mathbf{y}_{n,t}} = \frac{q + \frac{K}{2\beta}}{q + E_{n,t}^{\beta}} \tag{6.14}$$

Where, $E_{n,t} = \mathbf{y}_{n,t}^T \mathbf{M}_n^{-1} \mathbf{y}_{n,t}$.

In the M step we will employ a gradient based method to maximize the Q function. After E step we have the following normalized (by T) cost function,

$$\begin{aligned} C_{IVA} = &-\sum_{k=1}^{K} \log|\det(\mathbf{W}^{[k]})| \\ &-\frac{1}{T}\sum_{t}\sum_{n} \mathbb{E}_{z_{n,t}|\mathbf{s}_{n,t}}\left[\log p_{\mathbf{s}_{n,t}|z_{n,t}}(\mathbf{y}_{n,t})\right] \end{aligned} \tag{6.15}$$

The derivative of the cost function with respect to each demixing matrix gives,

$$\frac{\partial C_{IVA}}{\partial \mathbf{W}^{[k]}} = -(\mathbf{W}^{[k]})^{-T} - \frac{1}{T}\sum_{t}\sum_{n} \mathbb{E}_{z_{n,t}|\mathbf{s}_{n,t}}\left[\frac{\partial \log p(\mathbf{y}_{n,t})}{\partial y_{n,t}^{[k]}}\right] \frac{\partial y_{n,t}^{[k]}}{\partial \mathbf{W}^{[k]}} \tag{6.16}$$

Following the derivation shown in [3] and using the notation of score function of multivariate random vector, $\phi^{[k]}(\mathbf{y}_{n,t})$ we get,

$$\frac{\partial C_{IVA}}{\partial \mathbf{W}^{[k]}} = -(\mathbf{W}^{[k]})^{-T} + \frac{1}{T}\sum_{t} \phi^{[k]}(\mathbf{y}_{:,t})(\mathbf{x}_t^{[k]})^T \tag{6.17}$$

Where,$\boldsymbol{\phi}^{[k]}(\mathbf{y}_{:,t}) = \left[\boldsymbol{\phi}^{[k]}(\mathbf{y}_{1,t}), ..., \boldsymbol{\phi}^{[k]}(\mathbf{y}_{N,t})\right]^T$ is formed by selecting the k^{th} entries from each of the N multivariate score functions for sample t,

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{y}_{n,t}) &= -\mathbb{E}_{z_{n,t}|\mathbf{s}_{n,t}}\left[\frac{\partial \log p(\mathbf{y}_{n,t})}{\partial \mathbf{y}_{n,t}}\right] \\ &= \mathbb{E}_{z_{n,t}|\mathbf{s}_{n,t}}\left[\frac{1}{z_{n,t}}\right]\frac{\partial(\mathbf{y}_{n,t}^T\mathbf{M}_n^{-1}\mathbf{y}_{n,t})^{\beta}}{\partial \mathbf{y}_{n,t}} \\ &= 2w_{n,t}\beta E_{n,t}^{\beta-1}\mathbf{M}_n^{-1}\mathbf{y}_{n,t}\end{aligned} \tag{6.18}$$

Each of the K demixing matrices is updated sequentially using gradient descent method,

$$\mathbf{W}^{[k]} \leftarrow \mathbf{W}^{[k]} - \mu\frac{\partial C_{IVA}}{\partial \mathbf{W}^{[k]}} \tag{6.19}$$

Where, μ is positive scalar step size. According to recent works in IVA, it is suggested to use natural gradient for a faster convergence, which can be obtained by postmultiplying (6.17) by $(\mathbf{W}^{[k]})^T\mathbf{W}^{[k]}$. Hence, the natural gradient update rule of the demixing matrix becomes,

$$\mathbf{W}^{[k]} \leftarrow \mathbf{W}^{[k]} - \mu\left(\frac{1}{T}\sum_t \boldsymbol{\phi}^{[k]}(\mathbf{y}_{:,t})(\mathbf{y}_t^{[k]})^T - I\right)\mathbf{W}^{[k]} \tag{6.20}$$

It is interesting to note that the weights $(w_{n,t})$ in the expression of the multivariate score function (6.18) is a function of the SCV estimates of the previous iteration, which leads to this realm of reweighted algorithms for IVA.

6.4.1 Learning Intra-source Second order Dependencies

In the seminal work of IVA a Multivariate Laplace distribution along with the assumption that there is no correlation within the SCV, has been used as the source prior. Because of this strong assumption it may limit the performance of this IVA implementation in tasks where the degree of second order dependencies are expected to be significant. Our proposed unified framework enables us to capture any present

correlation structure by learning the Intra-source correlation matrix $\mathbf{M}_n$.

Revisiting the M step from previous subsection and collecting the terms with $\mathbf{M}_n$, we get,

$$Cost_{\mathbf{M}_n} = \sum_t \left[w_{n,t} (\mathbf{y}_{n,t}^T \mathbf{M}_n^{-1} \mathbf{y}_{n,t})^{\beta} + \frac{1}{2} \log |\mathbf{M}_n| \right] \tag{6.21}$$

Taking derivative w.r.t $\mathbf{M}_n$ and equating it to zero, we get the Maximum Likelihood estimate [138] as,

$$\mathbf{M}_n = \frac{2\beta}{T} \sum_t w_{n,t} (\mathbf{y}_{n,t}^T \mathbf{M}_n^{-1} \mathbf{y}_{n,t})^{\beta-1} \mathbf{y}_{n,t} \mathbf{y}_{n,t}^T \tag{6.22}$$

We will also add a regularization term to the update of $\mathbf{M}_n$ to make it robust to the estimation error of $\mathbf{y}_{n,t}$ over the iterations. Hence,

$$\mathbf{M}_n \leftarrow \frac{2\beta}{T} \sum_t w_{n,t} (\mathbf{y}_{n,t}^T \mathbf{M}_n^{-1} \mathbf{y}_{n,t})^{\beta-1} \mathbf{y}_{n,t} \mathbf{y}_{n,t}^T + \alpha I \tag{6.23}$$

Where, α is a positive scalar to maintain the positive definite property of $\mathbf{M}_n$. We will also normalize $\mathbf{M}_n$ after every iteration, i.e., $\mathbf{M}_n \leftarrow \mathbf{M}_n / ||\mathbf{M}_n||_F$.

6.4.2 Special Cases of Source Prior

In Table 6.1 we have listed how by choosing distributional parameters of a M-GT distribution we can represent several multivariate source priors that have been used in previous works for IVA. Here we will show how by choosing the specific distributional parameters in the unified inference framework, it leads to well known IVA implementations.

Multivariate Laplace Distribution : IVA-L

From Table 6.1 we see for specific values of distributional parameters ($q \to \infty, \beta = 1/2$), M-GT can be used to represent M-Laplace source prior. Now to relate with the unified IVA framework taking the limit as $q \to \infty$ in Equation (6.14) we get, $w_{n,t} = 1$ for all the sources. Hence the score function in Equation (6.18) becomes same as shown in [94] (with the choice of scatter matrix $\mathbf{M}_n = I$).

Multivariate Gaussian Distribution : IVA-G

Similarly with $q \to \infty, \beta = 1$, M-GT can be used to represent M-Gaussian source prior. Again the weights, $w_{n,t} = 1$ for all the sources. M-Gaussian has been used as a source prior for IVA in [6] with the choice of exploiting second order dependencies by learning the correlation matrix of each SCV ($\mathbf{M}_n$). By choosing aforementioned specific distributional parameters, score function (6.18) becomes same as in [6] and we can also exploit the correlation structure by using (6.23). We will denote this implementation as IVA-G.

Multivariate Generalized Double Pareto Distribution : IVA-Re ℓ_1

With the choice of $q = \varepsilon, \beta = 1/2$ we get our first proposed reweighted algorithm: IVA-Re ℓ_1. Its evident that with this choice of distributional parameters, weights given in (6.14) capture the relative energy differences between sources and reweights the score function (6.18) based on the SCV estimates of the previous iteration. Similar reweighted approach, iterative reweighted ℓ_1 norm minimization [30], has been explored for sparse recovery task in past. Also to note that, our proposed reweighted approach can also capture any correlation present within SCV by learning $\mathbf{M}_n$ (6.23). In our simulations we will use $q = \varepsilon = 0.1$.

Multivariate Student's t Distribution : IVA-Re ℓ_2

By choosing $q = \varepsilon, \beta = 1$ we get second of our proposed reweighted algorithms: IVA-Re ℓ_2. Weights, which are computed in (6.14), becomes function of 2 norm of the estimated SCV from previous iteration and are used to reweight the score function (6.18) . Hence we name this algorithm as IVA-Re ℓ_2. ε can be interpreted as the degrees of freedom of the Student's t distribution and it controls the tail nature of the source prior. Lower values of ε increases kurtosis or in other words makes the prior more heavytailed, whereas by increasing ε the tail nature approaches Gaussian and for $\varepsilon \to \infty$, Student's t distribution becomes a M-Gaussian distribution. In our simulations we will use $q = \varepsilon = 0.1$.

6.5 Simulations

Performance of the proposed reweighted algorithms in a source separation task has been shown in this section via simulations, following the standard set up, widely used in IVA literatures [3,4,6,108]. In our experiments we generate, $N = 3$ SCVs from a multivariate Laplace distribution. The k^{th} entry of each SCV is used a latent source for the k^{th} dataset. We also generate random mixing matrices, whose entries are generated from a normal distribution with mean 0 and variance 1. $T = 1000$ observations are generated following the mixing model given in (6.1) and presented to all the competing algorithms to estimate the SCVs, i.e. $\mathbf{y}_t^{[k]}$ and the demixing matrices, i.e. $\mathbf{W}^{[k]}$. The performance of the JBSS algorithms are evaluated using an extension of the normalized inter-symbol-interference (ISI), known as joint ISI which has been used in recent IVA literatures [4, 108]. Joint ISI penalizes SCV estimates that are not consistently aligned across datasets and is also normalized such that, $0 \leq$ Joint ISI ≤ 1, where 0 means ideal separation performance.

6.5.1 Uncorrelated Sources

During the first set of experiments we consider that, there is no correlation within an SCV across datasets. SCVs are generated randomly from a multivariate Laplace distribution with correlation matrix $\Gamma_n = I$. In Figure 6.1 we present the separation performances in terms of mean Joint ISI over 100 trials for all the competing algorithms for $K = 6, 9$ and,12 (number of datasets). As expected in this case the performance of IVA-G is the worst among all the competing algorithms, and the reason being the true SCVs posses no correlation, hence its impossible to achieve JBSS with only second order statistics in this case. Its also evident that both the Reweighted algorithms do better than IVA-L, even though there is a model mismatch (since sources are multivariate Laplace). As expected when the number of datasets increases, source separation performance also improves for both the reweighted algorithms and also IVA-L.

6.5.2 Correlated Sources

As discussed before, in many applications sources have second order (linear) dependencies across the datasets. Hence, here we consider the SCVs generated from a multivariate Laplace distribution following [57], with a randomly generated positive definite correlation matrix $\mathbf{\Gamma_n}$ for n^{th} source. In Fig. 6.2 we present the mean Joint ISI over 100 trials for all the competing algorithms for $K = 6, 9$ and,12 (number of datasets). Presence of correlation improves the performance of IVA-G and comparable with IVA-L. Even in this case Reweighted algorithms perform better than both IVA-L and IVA-G and the reason could be the flexibility of exploiting both higher order and second order statistics.

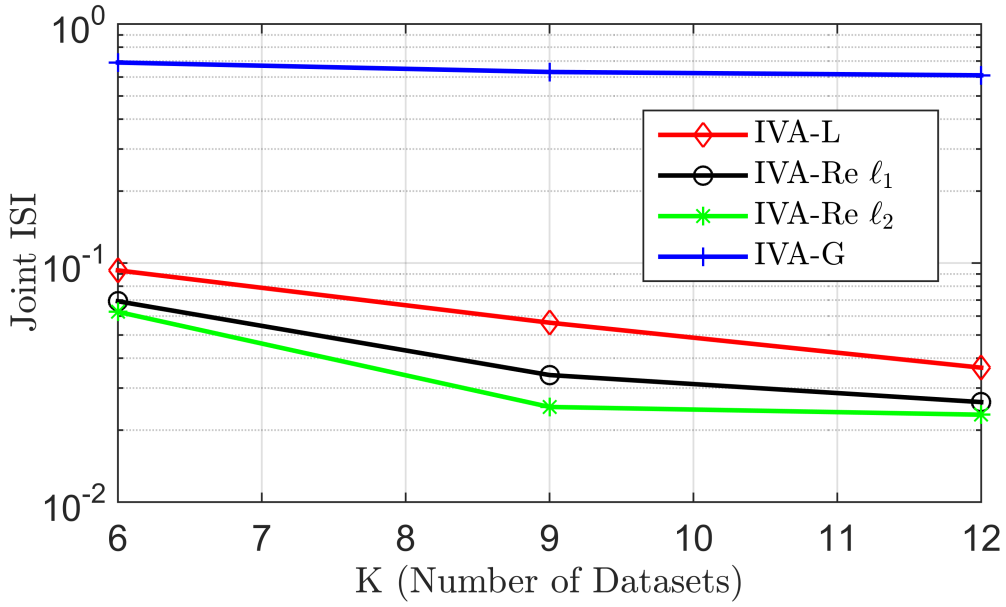

Figure 6.1. Joint ISI measure for Uncorrelated Sources using different competing algorithms

6.5.3 Convergence Issues

As discussed in [5] for non-Gaussian sources the cost function has local minimas at permutation ambiguities, which could be detrimental for these iterative algorithms, specially when the number of datasets (K) is less than the number of sources (N). Whereas, IVA-G has very desirable convergence properties as shown in [3] even for $K < N$ case. This acts as a motivation to a popular practice of using the IVA-G solution for initialization for other implementations of IVA. Exploration of this approach also for reweighted algorithms for the case of $K < N$ will be done in future.

6.6 Conclusion

In this article we have introduced a new class of generalized scale mixture distribution family: M-PESM as the multivariate source prior for IVA. Following a maximum likelihood inference procedure, choice of a specific member (M-GT) of M-

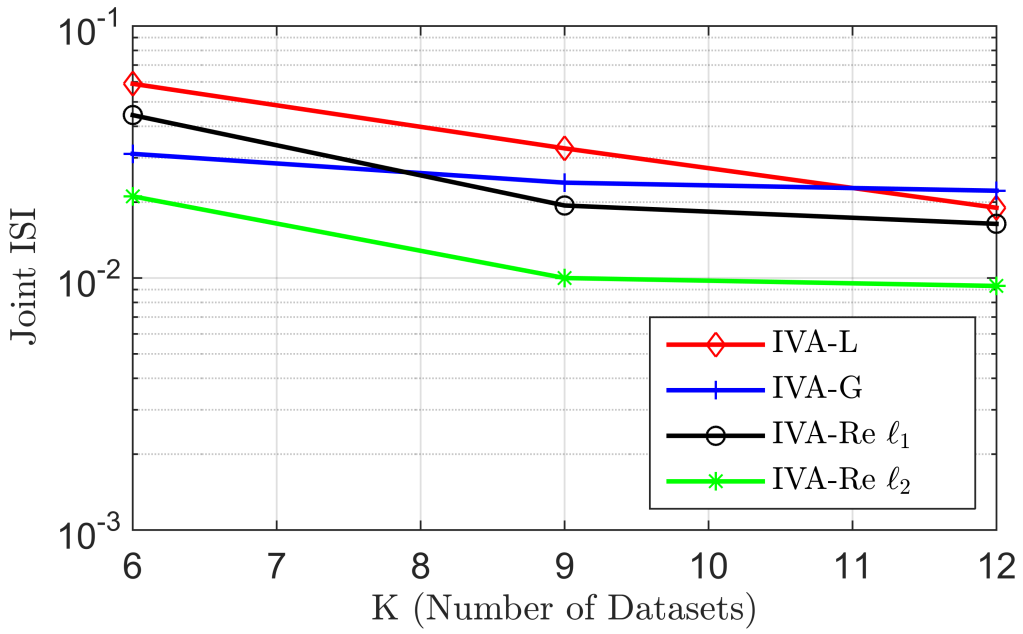

Figure 6.2. Joint ISI measure for Correlated Sources using different competing algorithms

PESM leads to two novel reweighted algorithms for IVA with the ability to exploit both second order and higher order dependencies within a SCV. This unified framework also includes two popular IVA implementations (IVA-L and IVA-G), that have been used in the literature. Simulation results show the superior performance of the reweighted algorithms over both IVA implementations (IVA-L and IVA-G).

As a future direction of this work, we intend to improve the convergence speed of our proposed reweighted algorithms by developing Newton's method based optimization technique with M-GT as the source prior.

6.7 Appendix

6.7.1 Derivation of Equation (6.14)

To compute the concerned expectation we will employ the following trick. Differentiating inside the integral of the marginalized $p(\mathbf{s}_{n,t})$ we get,

$$\begin{aligned} p'(\mathbf{s}_{n,t}) &= \frac{d}{d\mathbf{s}_{n,t}} \int_0^\infty p(\mathbf{s}_{n,t}|z_{n,t}) p(z_{n,t}) dz_{n,t} \\ &= -2\beta \times \lambda_{n,t}^{\beta-1} \mathbf{M}_n^{-1} \mathbf{s}_{n,t} \int_0^\infty \frac{1}{z_{n,t}} p(\mathbf{s}_{n,t}, z_{n,t}) dz_{n,t} \end{aligned} \tag{6.24}$$

Where, $\lambda_{n,t} = \mathbf{s}_{n,t}^T \mathbf{M}_n^{-1} \mathbf{s}_{n,t}$.

Now employing the product rule of probability $p(\mathbf{s}_{n,t}, z_{n,t}) = p(\mathbf{s}_{n,t}) p(z_{n,t}|\mathbf{s}_{n,t})$ and taking $p(\mathbf{s}_{n,t})$ outside the integral we get,

$$\begin{aligned} p'(\mathbf{s}_{n,t}) &= -2\beta \times \lambda_{n,t}^{\beta-1} \mathbf{M}_n^{-1} \mathbf{s}_{n,t} p(\mathbf{s}_{n,t}) \int_0^\infty \frac{1}{z_{n,t}} p(z_{n,t}|\mathbf{s_n,t}) dz_{n,t} \\ &= -2\beta \times \lambda_{n,t}^{\beta-1} \mathbf{M}_n^{-1} \mathbf{s}_{n,t} p(\mathbf{s}_{n,t}) \mathbb{E}_{z_{n,t}|\mathbf{s}_{n,t}} \left[\frac{1}{z_{n,t}} \right] \end{aligned} \tag{6.25}$$

Now lets consider a special case where a Multivariate GT has been employed as a prior, $p(\mathbf{s}_{n,t})$. We can write, $p(\mathbf{s}_{n,t}) = \eta \exp(-f(\mathbf{s}_{n,t}))$, where,

$$f(\mathbf{s}_{n,t}) = (q + \frac{K}{2\beta}) \log \left(q + \lambda_{n,t}^{\beta} \right) \tag{6.26}$$

So,

$$\begin{aligned} p'(\mathbf{s}_{n,t}) &= -p(\mathbf{s}_{n,t}) f'(\mathbf{s}_{n,t}) \\ &= -2\beta \times \lambda_{n,t}^{\beta-1} \mathbf{M}_n^{-1} \mathbf{s}_{n,t} p(\mathbf{s}_{n,t}) \frac{q + \frac{K}{2\beta}}{q + \lambda_{n,t}^{\beta}} \end{aligned} \tag{6.27}$$

Comparing Equation 6.25 and Equation 6.27 we get,

$$E_{z_{n,t}|\mathbf{s}_{n,t}}\left[\frac{1}{z_{n,t}}\right]_{\mathbf{s}_{n,t}=\mathbf{y}_{n,t}} = \frac{q+\frac{K}{2\beta}}{q+E_{n,t}^{\beta}} \tag{6.28}$$

Where, $E_{n,t} = \mathbf{y}_{n,t}^{T}\mathbf{M}_{n}^{-1}\mathbf{y}_{n,t}$.

6.8 Acknowledgment

The text of this chapter is based on the material as it appears in: ”Reweighted Algorithms for Independent Vector Analysis (IVA)”, submitted to IEEE Signal Processing Letters. Dissertation author was the primary author, while B.D. Rao and H. Garudadri supervised the research.

Chapter 7

Multi Task Learning

7.1 Introduction

Consider a linear regression problem, where there are L set of tasks (or measurement vectors) denoted as $\{\mathbf{y}_i\}_{1\ldots L}$ where, $\mathbf{y}_i \in \mathbb{R}^{n_i \times 1}$.

$$\mathbf{y_i} = \mathbf{X}_i \mathbf{w_i} + \boldsymbol{\varepsilon}_i \tag{7.1}$$

Where, $\mathbf{X}_i \in \mathbb{R}^{n_i \times m}$ is the data matrix constructed using training data, $\mathbf{w}_i \in \mathbb{R}^{m \times 1}$ is the coefficient vector and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{n_i \times 1}$ could be interpreted as measurement noise. Assuming that the measurement noise is zero mean Gaussian with unknown variance λ, the likelihood function for the coefficient vector $\mathbf{w}_i$ based on the ith task output/target $\mathbf{y}_i$ can be expressed as,

$$p(\mathbf{y}_i|\mathbf{w_i}, \lambda) = (2\pi\lambda)^{-n_i/2} \exp\left(-\frac{||\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i||_2^2}{2\lambda} \right) \tag{7.2}$$

When the number of features (m) is greater than the number of data points (n_i) in model (7.1) the problem becomes under-determined [53]. That means there could be infinite number of solutions for the regression coefficients that perfectly explain the data. To obtain a unique solution of regression coefficients we often employ a sparsity promoting regularization, which means only few relevant features will be selected [91]. There has been a lot of interest and work on promoting sparsity using ℓ_1 norm regularization [47, 48, 156]. From a Bayesian perspective supergaussian (i.e. priors with heavier tails than gaussian) distributions have been employed as prior to promote sparsity in the coefficient vector with reasonable success [132, 134]. For Multitask Learning (MTL) or a Multiple Measurement Vector (MMV) sparse recovery problem, notion of joint sparsity has been introduced [9, 168, 174]. Key assumption behind this is that all the tasks will share the same set of relevant features. Joint sparse regularization approach has been used, where we seek row sparsity in the regression coefficient matrix by employing a

multivariate supergaussian prior distributions to model joint sparsity, which encourages the entire rows of the coefficient matrix to have zero elements [66, 83, 176]. Joint regularization using ℓ_{2-1} mixed norm is a straightforward extension of LASSO (single task/measurement case), which has been used extensively to solve this problem [130]. In real life applications we often see that all the tasks may not always share the same set of features and some of the tasks could be outliers or could be negatively correlated with other tasks. To model the outlier tasks, recently a Dirty model for MTL has been introduced which uses a combined regularization of ℓ_1/ℓ_∞ to model the joint sparsity and ℓ_1 to model outliers [87]. A probabilistic interpretation of this dirty model has also been proposed in [84]. It has also been discussed in recent literatures [83, 176] that if the model is able to capture the task relatedness, i.e. any present correlation structure, the generalization capability of the model increases significantly. Recently some works [41, 142] have also proposed using Iterative Reweighted Least Square (IRLS) approaches to model joint sparsity from a MTL point of view. In [161] authors have extended the reweighted ℓ_1 minimization [30] approach to model the joint sparsity for MMV recovery problem. In Bayesian based approaches, Multivariate Gaussian Scale mixtures (M-GSM) and Multivariate Laplacian Scale Mixtures (M-LSM) have been used as prior distributions to promote joint sparsity, because of their supergaussian nature. In [160] authors have proposed a new sparse Bayesian multitask learning method based on a GSM prior which also models the correlation structure within tasks.

In the previous chapter, we have introduced a multivariate extension of our recently proposed generalized Scale Mixture framework [72], namely Multivariate Power Exponential Scale Mixtures (M-PESM) as a source prior for a joint blind source separation task. In this paper we present the usefulness of M-PESM to model the joint sparsity and show its application in a multi-task learning framework. This work will primarily focus on the Multivariate Generalized t distribution (M-GT) family of priors, a member

of M-PESM, since it has a wide range of tail shapes and includes heavy tailed super gaussian distributions. We also derive a unified MAP estimation framework using M-GT as sparsity inducing prior and show that many of the popular regularization based MTL algorithms falls under our proposed unified framework. Our model also has the flexibility of learning any correlation structure present between tasks which will help us to model any outlier task or task with negative correlation.

The rest of the chapter is organized as follows. In Section 7.2, a generalized scale mixture representation, the Multivariate Power Exponential Scale Mixtures (M-PESM) family, is presented. In Section 7.3, we derive a unified MAP based inference procedure by employing a joint sparsity promoting prior distribution from the family of M-PESM. In Section 7.4, we discuss some special cases of the unified framework and show connections with current algorithms in the literature. We present experimental results of the proposed algorithms using both synthetic data and real data in Section 7.5, in different settings and finally conclusions and some future directions of this work are presented in Section 7.6.

7.2 Sparsity Inducing Prior: Scale Mixtures

For joint sparse regularization from a MMV or MTL point of view, multivariate Gaussian scale mixtures and Laplace scale mixtures have been used as sparsity promoting prior. In this section, we discuss a recently proposed [70], more general Multivariate Power Exponential Scale Mixture (M-PESM) distribution, which is a generalization of M-GSM and M-LSM.

7.2.1 Multivariate Power Exponential (M-PE)

In this work we are concerned with the M-PE distribution, which is also known as Generalized Gaussian Distribution (GGD) and has received lot of attention in the

literature. The probability density function of a M-PE is defined by [138],

$$p_{\text{M-PE}}(\mathbf{x}|\mathbf{M},\beta,z) = \frac{1}{|\mathbf{M}|^{1/2}} h_{\beta,z}(\mathbf{x}^T\mathbf{M}^{-1}\mathbf{x}) \tag{7.3}$$

for any $\mathbf{x} \in \mathbb{R}^{L\times 1}$, where $\mathbf{M}$ is a $L \times L$ symmetric real correlation matrix, and $h()$ is known as the density generator defined by,

$$h_{\beta,z}(y) = \frac{\beta\Gamma(\frac{L}{2})}{\pi^{\frac{L}{2}}\Gamma(\frac{L}{2\beta})z^{\frac{L}{2\beta}}} \exp\left(-\frac{y^\beta}{z}\right) \tag{7.4}$$

Where, $z > 0$ is the scale parameter and $\beta > 0$ is the shape parameter of the M-PE. It is evident from the above given form, that $\beta = 1$ results in the Multivariate Gaussian distribution, whereas $\beta = 1/2$ connects to the well known Multivariate Double exponential or Laplace distribution.

7.2.2 Multivariate PESM (M-PESM)

Multivariate PESM family of distributions refer to distributions that can be represented as follows:

$$p_{\mathbf{X}}(\mathbf{x}) = \int p_{\text{M-PE}}(\mathbf{x};\mathbf{M},\beta,z)p_z(z)dz \tag{7.5}$$

Some special cases of M-PESM includes Multivariate Gaussian Scale Mixtures (M-GSM) when shape parameter $\beta = 1$, Multivariate Laplace Scale Mixtures (M-LSM) when shape parameter $\beta = 1/2$, Multivariate Uniform Scale Mixtures (M-USM) when $\beta \to \infty$. More theoretical details and the properties of M-PESM can be found in [70].

7.2.3 Multivariate Generalized t Distribution (M-GT)

In this example, we will consider an inverse gamma (IG) distribution as our mixing density $p_z(z) = IG(q,q)$, where $IG(x;a,b) = \frac{b^a}{\Gamma(a)}x^{-a-1}\exp\big(-\frac{b}{x}\big)u(x)$ in the hierarchical representation (7.5) for the M-PESM family. It leads to a multivariate generalized t distribution [11] which also includes well known supergaussian densities,

Table 7.1. Variants of Multivariate GT distribution

q	β	Prior Distribution	Penalty Function	SSR Algorithm
$q \to \infty$	1	M-Normal	$\|\|\mathbf{W}\|\|_F$	M-Ridge Regression
$q \to \infty$	1/2	M-Laplacian	$\|\|\mathbf{W}\|\|_{2,1}$	M-LASSO
$q \geq 0$ (degrees of freedom)	1	M-Student t distribution	$\sum_i \log(\varepsilon + \|\|\mathbf{w}_{i,:}\|\|_2^2)$	Iterative Reweighted Least Squares
$q \geq 0$ (shape parameter)	1/2	M-Generalized Double Pareto	$\sum_i \log(\varepsilon + \|\|\mathbf{w}_{i,:}\|\|_2)$	Reweighted ℓ_1

useful to promote joint sparsity e.g. Multivariate Laplace, Multivariate Student's t distributions, among others. The Multivariate Generalized t Distribution has the form:

$$p_{\text{M-GT}}(\mathbf{x}; q, \beta, \mathbf{M}) = \frac{\eta}{(q + s^{\beta})^{q + \frac{L}{2\beta}}} \tag{7.6}$$

Where $s = \mathbf{x}^T \mathbf{M}^{-1} \mathbf{x}$, η is the normalization constant. Interestingly, β and q provide the flexibility to represent different tail behavior using this distribution. In Table 7.1, we summarize some special cases of Multivariate GT that have been used in literature to promote joint sparsity that arise by different choices of the shape parameters of M-GT, i.e. β and q (With $\mathbf{M} = I$).

7.3 Bayesian Inference

In this section we derive a unified estimation algorithm using M-PESM as the sparse prior. Then we specialize the result using the M-GT as the sparse prior and also show that the generalized algorithm reduces to well known Multi task learning algorithms.

7.3.1 Unified MAP Estimation

Because of the independence between rows of the coefficient matrix $\mathbf{W}$, every $p(\mathbf{w}_{i,:})$ has an independent scale mixture representation, i.e,

$$p(\mathbf{w}_{i,:}) = \int_0^{\infty} p(\mathbf{w}_{i,:}|z_i) p(z_i) dz_i \tag{7.7}$$

For EM algorithm we will treat scale parameters z_i as hidden variables. Hence the complete data log-likelihood can be written as,

$$\log p(\mathbf{Y},\mathbf{W},\mathbf{z}) = \log p(\mathbf{Y}|\mathbf{W}) + \sum_{i=1}^{m} \log p(\mathbf{w}_{i,:}|z_i) + \sum_{i=1}^{m} \log p(z_i) \tag{7.8}$$

To compute the Q function we need the conditional expectation of the complete data log likelihood with respect to the conditional posterior of the hidden variables, i.e, $p(\mathbf{z}|\mathbf{W},\mathbf{Y})$ which reduces to $p(\mathbf{z}|\mathbf{W})$ by virtue of the Markovian property. Now in the M step we will maximize the Q function with respect to $\mathbf{W}$, so we are only interested in the first two terms of the Equation (7.8). Since only the second has dependencies on the hidden variable $\mathbf{z}$, in the E step we are only concerned with this term, i.e,

$$\begin{aligned} \sum_{i=1}^{m} \log p(\mathbf{w}_{i,:}|z_i) &= \sum_{i=1}^{m} \log p_{\text{M-PE}}(\mathbf{w}_{i,:};\mathbf{M}_i,\beta,z_i) \\ &= -\sum_{i=1}^{m} \frac{(\mathbf{w}_{i,:}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T)^{\beta}}{z_i} + \text{constants} \end{aligned} \tag{7.9}$$

Hence, the E step essentially becomes computation of the following conditional expectation, $E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right]$.

The derivation of the concerned conditional expectation where a M-GT has been employed as the sparsity inducing prior, is given in Appendix of the last chapter, which has been found as,

$$E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right] = \frac{q+\frac{L}{2\beta}}{q+E_i^{\beta}} \tag{7.10}$$

Where, $E_i = \mathbf{w}_{i,:}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T$. Lets define the weights as,

$$v_i = E_{z_i|\mathbf{w}_{i,:}}\left[\frac{1}{z_i}\right] = \frac{q+\frac{L}{2\beta}}{q+E_i^{\beta}} \tag{7.11}$$

Hence the M step becomes,

$$\mathbf{W}^{(k+1)} = \arg\min_{\mathbf{W}} \sum_{i=1}^{L} \frac{1}{2\lambda}\|\mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i\|_2^2 + \sum_{i=1}^{m} v_i^{(k+1)}(\mathbf{w}_{i,:}\mathbf{M}_i^{-1}\mathbf{w}_{i,:}^T)^{\beta} \tag{7.12}$$

It's evident from the M step that our proposed unified framework falls under the reweighted schemes where weights of $(k+1)^{th}$ iteration, i.e, $v_i^{(k+1)}$ depend on the coefficients from previous iteration.

7.3.2 Learning Task Correlation

By incorporating a data adaptive correlation matrix $\mathbf{M}_i$ in our algorithm, we can capture any outlier tasks. It will also help to exploit any present correlation structure in $\mathbf{w}_{i,:}$ through learning $\mathbf{M}_i$ adaptively. In our algorithm we will constrain all the $\mathbf{M}_i = \mathbf{M}$, to prevent overfitting because of the large number of parameters.

Revisiting the M step and taking derivative with respect to $\mathbf{M}$ and equating it to zero we get,

$$\mathbf{M}^{(k+1)} = \frac{2\beta}{m}\sum_{i=1}^{m} v_i^{(k+1)}(\mathbf{w}_{i,:}\mathbf{M}^{(\mathbf{k})^{-1}}\mathbf{w}_{i,:}^T)^{\beta-1}\mathbf{w}_{i,:}^T\mathbf{w}_{i,:} \tag{7.13}$$

In real applications we will also add a regularization term to the update of $\mathbf{M}$ to make it robust to the estimation error of $\mathbf{W}$ over the iterations.

$$\mathbf{M}^{(k+1)} \leftarrow \frac{2\beta}{m}\sum_{i=1}^{m} v_i^{(k+1)}(\mathbf{w}_{i,:}\mathbf{M}^{(\mathbf{k})^{-1}}\mathbf{w}_{i,:}^T)^{\beta-1}\mathbf{w}_{i,:}^T\mathbf{w}_{i,:} + \alpha I \tag{7.14}$$

Where, α is a small positive scalar, to maintain the positive definite property of $\mathbf{M}$. We will also normalize $\mathbf{M}$ after every update, i.e, $\hat{\mathbf{M}}^{(k+1)} \leftarrow \mathbf{M}^{(k+1)}/\|\mathbf{M}^{(k+1)}\|_F$. This data adaptive correlation matrix $\mathbf{M}$ can also be interpreted as data adaptive kernel which helps

to exploit any structure present among the tasks which is a significant advantage over algorithms that are blind to any correlation structure.

7.4 Special Cases of Unified Framework

In this section by choosing specific distributional parameters we will show how our proposed unified framework leads to well known Multitask Learning algorithms.

7.4.1 ℓ_{2-1} Minimization: Joint Feature Selection

ℓ_{2-1} norm minimization based joint feature selection approach [130] is one of the earliest multitask learning algorithm employing joint sparse regularization. From a Bayesian point of view employing a M-Laplace distribution as the joint sparsity inducing prior over the rows of the coefficient matrix and seeking a MAP estimate will lead to this algorithm. Interestingly we see from Table 7.1 that for specific values of the shape parameters ($q \to \infty, \beta = 1/2$), a Multivariate GT distribution can be used to represent M-Laplace. Now to relate with the unified MAP estimation framework taking the limit as $q \to \infty$ in Equation (7.11) we get $v_i = 1$. Hence in the M step we are solving a ℓ_{2-1} norm penalized regression problem where weights are not changing over iteration, showing that ℓ_{2-1} Minimization is a special case of our unified framework.

7.4.2 Iterative Reweighted ℓ_1 minimization (IRL-1)

In [112,161] an iterative reweighted ℓ_1 minimization algorithm has been discussed to promote joint sparsity. From a Bayesian point of view, MAP estimation of the coefficient matrix with a M-Generalized double pareto distribution as a prior will lead to the same cost function. Now, substituting the distributional parameters ($q = \varepsilon, \beta = 1/2$) from Table 7.1 in Equation (7.11) we get weights as, $v_i = \frac{\varepsilon + L}{\varepsilon + \sqrt{\mathbf{w}_{i,:}\mathbf{w}_{i,:}^T}} = \frac{\varepsilon + L}{\varepsilon + ||\mathbf{w}_{i,:}||_2}$, same as shown in [112] using MM algorithm. It's evident that this algorithm also falls under our

proposed unified framework. On the other hand our framework also allows learning the correlation structure between tasks and leads to correlation aware regularization penalty unlike the algorithm discussed in [112]. We will refer to the context aware version of this algorithm as **C-IRL-1** which involves computing the weights v_i following Equation (7.11) with $q = \varepsilon, \beta = 1/2$, updating the correlation matrix $\mathbf{M}$ using Equation (7.14) with $\beta = 1/2$ and then solving a weighted ℓ_{2-1} mixed norm minimization problem shown in Equation (7.12).

7.4.3 Iterative Reweighted Least Squares (IRLS)

Iterative Reweighted Least Square (IRLS) was first proposed from a single measurement sparse recovery perspective. In recent works [41, 142] it has been extended for joint sparse regularization both from a MMV recovery and Multitask learning point of view. As shown in Table 7.1, employing a M-student t distribution as a prior and following the MAP estimation route will lead to the same cost function as discussed in [41]. By choosing the specific distributional parameters (from Table 7.1) and substituting in Equation (7.11) we get, $v_i = \frac{\varepsilon + L/2}{\varepsilon + \mathbf{w}_{i,:}\mathbf{w}_{i,:}^T} = \frac{\varepsilon + L/2}{\varepsilon + ||\mathbf{w}_{i,:}||_2^2}$, which is a straightforward extension of Reweighted ℓ_2 minimization algorithm [33] for MMV case. Since our unified framework allows us to learn the correlation structure, in our proposed correlation aware IRLS (**C-IRLS**) the weights will be computed as, $v_i = \frac{\varepsilon + L/2}{\varepsilon + \mathbf{w}_{i,:}\mathbf{M}^{-1}\mathbf{w}_{i,:}^T}$ We will also learn the correlation matrix using Equation (7.14) and then we just need to solve a weighted least squares problem following Equation (7.12) with $\beta = 1$.

Updating the shape parameter ε

In our proposed C-IRLS we will also employ a useful update strategy of the shape parameter ε of multivariate GT distribution. Choice of ε controls the kurtosis of the prior distribution, where a low value of ε corresponds to a higher kurtosis. Starting with a very

low value of ε may hurt the performance of the algorithm as it can get stuck to a local optima since the regularization term is concave. Whereas starting with a comparatively higher value of ε and slowly decreasing it will help our proposed C-IRLS to converge to the global optima. This can be also viewed as adapting the tail nature of the prior distribution over the iterations as it approaches global optima. Similar discussion can also be found in [33] where the motivation is that higher value of ε will result in the undesirable local minimas being "filled in".

The ε update rule that has been used in this work is as follows: If the relative Frobenius norm of the coefficient matrix $\mathbf{W}$, from previous iteration is less than $\sqrt{\varepsilon}/100$ we decrease the value of ε by a factor of 10. The algorithm is run till $\varepsilon < 1e-6$ or the maximum number of iterations which is 100 in all our experiments.

7.5 Experiments

In this section we carry out experiments using both synthetic data and real data to evaluate the empirical performances of the above discussed models.

7.5.1 Competing Algorithms

All the competing algorithms have been summarized below with a brief summary.

1. ℓ_{2-1} mixed norm minimization based MTL. [130]
2. M-FOCUSS: MMV based FOCal Underdetermined System Solver (M-FOCUSS) with $p = 0.8$. [39]
3. IRL-1: Iterative Reweighted ℓ_1 minimization for joint sparsity. [161]
4. C-IRL-1: Correlation aware Reweighted ℓ_1 minimization. (proposed)
5. TMSBL: Temporal MMV Sparse Bayesian Learning. [160]

6. IRLS: Iterative Reweighted Least Squares. [41]

6. C-IRLS: Correlation aware Iterative Reweighted Least Squares. (proposed)

7. DM: Dirty model with combined regularization of ℓ_1 and ℓ_1/ℓ_∞ to model outliers. [87]

7.5.2 Experiments with Synthetic Data

In this case we will assume that same data matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ (where, $n = 50, m = 100$) has been used for all the tasks. The entries of the data matrix $\mathbf{X}$ have been sampled from a standard Gaussian distribution with mean zero and standard deviation 1. Lets assume that there are $L = 10$ tasks and all the task share the same set of $K = 22$ relevant features. We will also assume that the first two tasks and the last eight tasks are positively correlated but the two groups are negatively correlated. Thus the nonzero rows of the coefficient matrix $\mathbf{W}_{gen}$ have been sampled from a multivariate Gaussian with mean zero vector and covariance matrix with 1's on the diagonals and either $+\beta$ or $-\beta$ on the off diagonal elements, depending on the locations. Now the target matrix $\mathbf{Y}$ is obtained following $\mathbf{Y} = \mathbf{X}\mathbf{W}_{gen} + \varepsilon$. Where the additive noise is gaussian and the variance is chosen such that SNR is 10 dB. The target matrix $\mathbf{Y}$ and data matrix $\mathbf{X}$ are shown to all the competing algorithms and the reconstruction error of model coefficients are measured as: Error $= \frac{\|\hat{\mathbf{W}}-\mathbf{W}_{gen}\|_F}{\|\mathbf{W}_{gen}\|_F}$. The same experiment has been repeated 50 times and the averaged error has been reported in Table 7.2. We run the experiments for two values of $\beta = 0.9$ and, 0. In the first case there is a significant correlation structure between tasks, so we hope to see a significant improvement for our proposed correlation aware algorithms. Whereas in the second case there is no correlation structure so we expect to see similar performance of both Correlation aware and correlation unaware algorithms. In Table 7.2 for $\beta = 0.9$ we see that C-IRLS performs significantly better compared to

IRLS whereas C-IRL-1 also shows little improvement over IRL-1.

Table 7.2. Averaged Reconstruction Error using Synthetic Data

Methods	Error	
	$\beta = 0.9$	$\beta = 0$
ℓ_{2-1}	0.6007	0.4800
M-FOCUSS	0.6321	0.4559
IRL-1	0.3768	0.2712
C-IRL-1 (Proposed)	0.3679	**0.2710**
TMSBL	0.4325	0.3168
IRLS	0.4795	0.3056
C-IRLS (Proposed)	**0.3633**	0.3030
DM	0.6489	0.5629

7.5.3 Experiments with Real Data

In this section we consider the reconstruction of images of hand written digits taken from the popular MNIST dataset. Since for these handwritten digits the background pixels are always zero and most of them share same locations across all the images, joint sparsity could be used here. We downsample the images to 14×14 pixels and vectorize them, where each image is represented using a 196 dimensional vector. We randomly choose 8 images of digit '0' and two randomly chosen images of digit '1' and digit '9'. Last two digits i.e, '1' and '9' can be interpreted as outlier tasks. Now in MTL setup, model coefficients $\mathbf{w}_l$ are the vectorized pixel values. Again we will choose the same data matrix $\mathbf{X} \in \mathbb{R}^{120\times 196}$ for all the tasks and the entries of $\mathbf{X}$ are sampled from a standard Gaussian distribution with mean zero and standard deviation 1. Following the previous section we will generate the target matrix $\mathbf{Y}$ with some additive noise where the SNR is 20 dB. We compare the reconstruction error by several competing algorithms in Table 7.3. We again see the improvement of performance by correlation aware algorithms, where

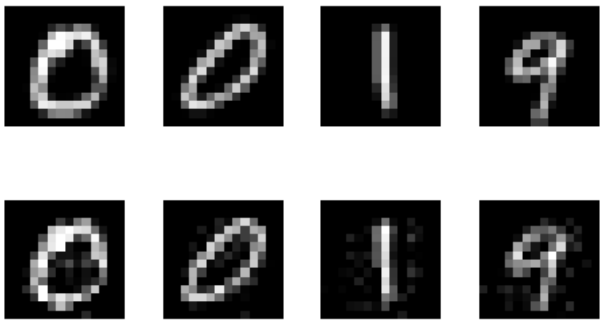

Figure 7.1. (Top) True Images, (Bottom) Recon. images using C-IRLS

C-IRLS produces the best reconstruction error.

Table 7.3. Averaged Reconstruction Error using MNIST

Methods	Error
ℓ_{2-1}	0.3879
M-FOCUSS	0.3218
IRL-1	0.2965
C-IRL-1 (Proposed)	0.2834
TMSBL	0.3039
IRLS	0.3056
C-IRLS (Proposed)	**0.2426**
DM	0.4212

In Figure 7.2(a) we show true images of two ’0’s (7th and 8th task) and the outliers ’1’ and ’9’ (9th and 10th task) and also the corresponding reconstructed images using C-IRLS. In Figure 7.2(b) we show the correlation matrix that has been learned by C-IRLS (White corresponds to 1 and black corresponds to 0). Interestingly we find out that our model has been able to learn high correlation between the first 8 tasks (images of ’0’) and also a very low correlation between a true task and last two outlier tasks. Another interesting observation is the correlation learned between 7th and 8th task in Figure 7.2(b) (Red circled), which is also low, though they belong to the same digit. For sanity check, we can verify from Figure 7.2(a) that the 7th task and 8th task, i.e., two

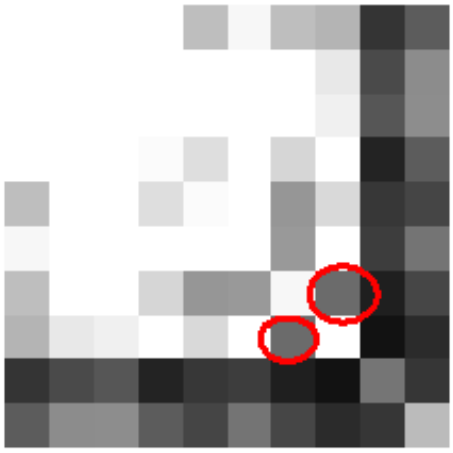

Figure 7.2. Correlation between tasks learned by C-IRLS for MNIST

true images of handwritten '0' are significantly different which leads to a low correlation value captured by **C-IRLS**.

7.6 Conclusion

In this chapter we have introduced a new class of multivariate scale mixture prior distribution to model joint sparsity and derived a unified inference framework which covers many of the popular Multitask learning algorithms. Our proposed correlation aware algorithms provide the flexibility of exploiting any present correlation structure between tasks. Our experimental results over both synthetic data and real data shows improvements of the proposed correlation aware approaches over other competing algorithms.

7.7 Acknowledgment

The text in this chapter, in full, is a reprint of material as in "Multivariate Scale Mixtures for Joint Sparse Regularization in Multi-Task Learning", submitted to IEEE

International conference on Acoustics, Speech and Signal Processing (ICASSP) 2017. The dissertation author was the primary researcher and B.D. Rao supervised the research.

Chapter 8

Block Sparse Excitation based Speech Modeling

8.1 Introduction

In speech modeling, an all pole model is most commonly used to model the vocal tract. Depending on the nature of the utterance, voiced, unvoiced or mixed, the input to the all-pole filter is either a glottal pulse train, white noise, or a combination of glottal pulses and white noise respectively. Estimation of the model parameters has a long history and a popular approach is the linear prediction (LP) based all pole model parameters estimation which involves minimizing the 2-norm of the residual, the difference between the observed signal and the predicted signal. The residual signal in all pole modeling is the input excitation sequence. Because of the 2-norm minimization approach, such estimation methods work well for unvoiced speech where the input to the filter is white noise. The 2-norm minimization based linear prediction approach suffers from some well known problems [120] in the case of voiced speech. The spectrum of the resulting model tends to overestimate the spectral powers at the formant frequencies, providing a sharper contour than the original vocal tract response. Several different methods have been proposed to alleviate these effects. Some of the proposed techniques involve a general rethinking of the spectral modeling problem [52, 128] while some others are based on changing the statistical assumptions made on the prediction error in the minimization process [44, 149]. Recently, instead of minimizing the 2-norm of the residual, methods based on minimizing the one norm of the residual, to accommodate the spike train nature of the input sequence, have been suggested with some success for voiced speech [69]. Interesting algorithms [31, 69] based on reweighted ℓ_1 approaches have been employed to exploit the sparsity assumption on the input process.

In case of voiced speech, the excitation can be considered to be a sparse excitation of a quasi-periodic nature [42]. The excitation component of the voiced speech production model is known as the glottal excitation. The structure of this glottal excitation has been

an interesting topic of research for several years. From Figure 8.1 the temporal extent of the glottal pulses show that a block sparse structure is more appropriate. Thus to make the voiced speech modeling task more robust and efficient we propose a framework where the excitation has a prior block sparse quasi-periodic structure. It is useful to note that block sparsity has been studied before in the context of sparse signal recovery, but they are usually for under-determined problems and the block sparsity is imposed on the solution vector [180], not on the residual as discussed here. The model is then generalized to deal with the broad spectrum of speech signals. In our proposed model the residual is modeled as being a linear combination of two components: a block sparse component and a Gaussian i.i.d white noise component. By appropriately weighting the components, this model for the input can deal with all speech utterances; voiced, unvoiced speech and mixed excitation speech.

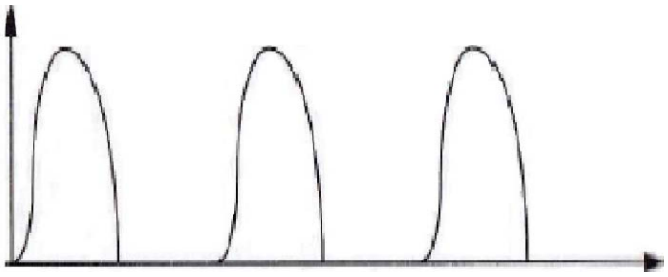

Figure 8.1. Shape of Glottal Excitation

The rest of the chapter is organized in the following way. Section 8.2 presents the model and discusses its advantages and disadvantages and Section 8.3 provides a detailed description of the estimation procedure of the parameters. Section 8.4 summarizes the performance of the proposed model over synthetic data, and Section 8.5 presents the results of the speech modeling problem over the Vowel dataset and finally Section 8.6 concludes the chapter.

8.2 Proposed Model

Since we are modeling the vocal tract using all-pole models, we will consider the signal to have been generated by an all-pole filter excited by an appropriate input, either block sparse, white noise or a combination. The all-pole model parameters and the nature of excitation input sequence are not known before hand. For instance, in speech this depends on the utterance. This production model can be described by the following difference equation,

$$x(n) = \sum_{k=1}^{M} a_k x(n-k) + w(n) + e(n) \tag{8.1}$$

Thus $x(n)$ is written as a linear combination of past M samples. Here a_k are the model parameters and $w(n)$ is the block sparse excitation sequence, whereas $e(n)$ is the non sparse white noise component. Now considering this production model for a segment of sample length N, for n=1 to N, we can represent this model in matrix form as,

$$\mathbf{Y} = X\mathbf{a} + \mathbf{w} + \mathbf{e} \tag{8.2}$$

Where, $\mathbf{Y} = [x(M+1), x(M+2) \cdots x(N)]^T$, X is the known data matrix which is constructed from the known time series data. A pictorial representation of this model is shown in Figure 8.2. The main idea behind this model is that **w** will capture the (block) sparse excitation and **e** will capture the standard non-sparse Gaussian excitation and provide a richer class of excitation sequences and richer class of models. In the context of speech, by appropriate weighting of these components we have the ingredients to deal with all types of speech signals. For voiced speech, **w** will dominate the residual. For unvoiced speech, **e** will dominate the residual. For mixed speech both components would be present at appropriate levels. For the block sparse structure of **w**, we assume that the

all the block sizes are equal and equal to d, and that the blocks are non-overlapping and contiguous, i.e. block boundaries known. Though a more general block structure can be imposed, our experiments indicate that the methods developed work reasonably with a properly chosen block size d.

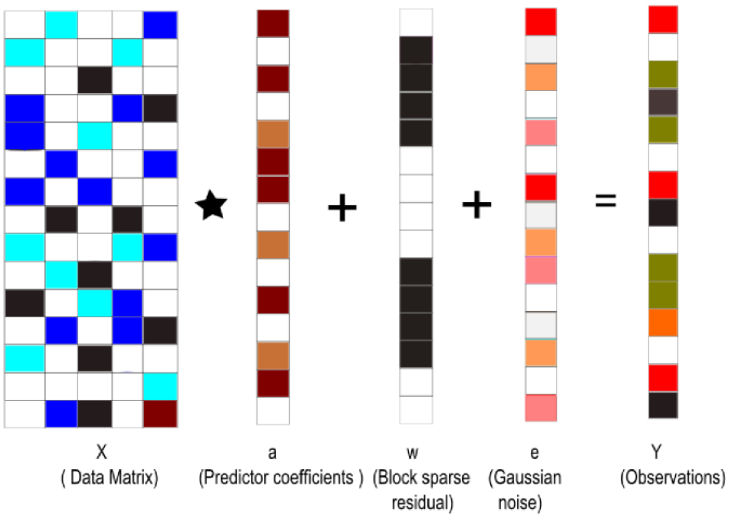

Figure 8.2. Pictorial Representation of the proposed model

8.3 Parameter Estimation

To estimate the parameters of our model, we can proceed in two ways. First is a deterministic setting where an extension of the ℓ_1 norm is considered such as a mixed norm ℓ_1/ℓ_2 norm, i.e. minimizing the ℓ_1 norm of the ℓ_2 norm of the blocks. In our work, we have chosen a probabilistic setting by adopting the empirical Bayes approach because of its flexibility and it also readily allows this type of two component noise modeling technique [89]. In particular, we utilize the Sparse Bayesian learning (SBL) [157] methodology. Detailed analysis of the original SBL for sparse signal recovery have been extensively discussed in several literatures [165] [167]. Interested readers are referred to these references for more details. We will use a standard EM algorithm to estimate the

parameters of our model. It is assumes that

$$p(\mathbf{e}) = N(0, \sigma^2 I) \tag{8.3}$$

Thus,

$$p(\mathbf{Y} - X\mathbf{a}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} exp - \frac{|\mathbf{Y} - X\mathbf{a} - \mathbf{w}|^2}{2\sigma^2} \tag{8.4}$$

For this model framework we will assume that the error $\mathbf{w}$ has a normal distribution with mean zero and a block structure of block size d. Under the SBL formulation, the covariance matrix of these error blocks is modeled as $\gamma_i I, i = 1, .., L$. Hence the covariance matrix of the complete error sequence is

$$\Gamma = \text{diag}(\gamma_1 I, ..., \gamma_L I) \tag{8.5}$$

Here γ_i is the hyperparameter which controls the variance of the i^{th} block and have to be learnt. If $\gamma_i = 0$, it means that the corresponding block will also be zero.

To estimate the values of the parameters $\mathbf{a}, \sigma^2$ and γ_is we will use the EM algorithm and will consider $\mathbf{w}$ as the latent variable. The complete loglikelihood can be written as,

$$L = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}|\mathbf{Y} - X\mathbf{a} - \mathbf{w}|^2 - \frac{N}{2}\log 2\pi - \frac{1}{2}\log(\det(\Gamma)) - \frac{1}{2}\mathbf{w}^\intercal\Gamma^{-1}\mathbf{w} \tag{8.6}$$

The Q function is defined as,

$$Q = E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{t-1}}[L] \tag{8.7}$$

Thus we need to know, $E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{t-1}}[\mathbf{w}]$ and $E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{t-1}}[\mathbf{w}^\intercal\mathbf{w}]$

After some simple manipulations we obtain,

$$\hat{W}_1 = E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{t-1}}[\mathbf{w}] = (I+\sigma^2_{t-1}\Gamma^{-1}_{t-1})^{-1}(\mathbf{Y}-X\mathbf{a_{t-1}}) \tag{8.8}$$

and,

$$\begin{aligned}\hat{W}_2 &= E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{t-1}}[\mathbf{w}^\intercal\mathbf{w}] \\ &= (I+\sigma^2_{t-1}\Gamma^{-1}_{t-1})^{-1}(\mathbf{Y}-X\mathbf{a_{t-1}})(\mathbf{Y}-X\mathbf{a_{t-1}})^\intercal(I+\sigma^2_{t-1}\Gamma^{-1}_{t-1})^{-1}+(\sigma^{-2}_{t-1}I+\Gamma^{-1}_{t-1})^{-1}\end{aligned} \tag{8.9}$$

In the M-step we will maximize the Q function with respect to our model parameters. So after taking derivative with respect to the parameters and setting them to zero we get,

$$\gamma_i = \frac{1}{d}\sum_{j=(i-1)d+1}^{id}\hat{w_j^2} \text{ where}, \hat{w_j^2} = [\hat{W}_2]_{j,j} \tag{8.10}$$

$$\sigma^2 = \frac{1}{N}[|\mathbf{Y}-X\mathbf{a}|^2 - 2(\mathbf{Y}-X\mathbf{a})^\intercal\hat{W}_1 + tr(\hat{W}_2)] \tag{8.11}$$

$$\mathbf{a} = (X^\intercal X)^{-1}X^\intercal(\mathbf{Y}-\hat{W}_1) \tag{8.12}$$

Hence by using these update rules the parameters of the model can be estimated in each iteration.

8.4 Experiments on Synthetic data

In this section we will discuss the experiments over the synthetic data to validate our above mentioned models. Here, we will use an all pole model that has been obtained

after modeling a speech segment using LPC technique, to produce the synthetic speech signal by passing three different types of excitations through it. As we are dealing with block sparse excitations, the period of these block excitation becomes an important factor and this can be viewed as the pitch period. Thus, in the language of speech domain all the experiments have been performed using two pitch frequencies, 100 Hz and 200 Hz. Now as this pitch frequency changes with time in case of speech signals, a little randomization has also been introduced when using this pitch frequency. We did the experiments for two cases where case 1 is $f_1 = 100 + N(0,9)$ and case 2 is $f_2 = 200 + N(0,9)$ where $N(0,9)$ is normal random variable with mean 0 and variance 9. For all these experiments we have used block size= 6 (empirically chosen).

The performance of a spectral envelope estimation method can be measured in many ways. An often used criterion for measuring quality is the spectral distortion between estimated all pole model $S'(\omega, \mathbf{a})$ and the true all pole model $S(\omega)$ which is the ground truth where, $S(\omega) = \frac{1}{|A(e^{j\omega})|^2}$ and $A(e^{j\omega})$ is defined by the filter coefficient vector $(a_0 ..... a_M)$.

This Spectral Distortion measure is defined as,

$$SD = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} [10 \log_{10} S(\omega) - 10 \log_{10} S'(\omega, \mathbf{a})]^2 d\omega} \tag{8.13}$$

For a pair of spectra $S(\omega)$ and $S'(\omega, \mathbf{a})$, by applying Parseval's Theorem we can relate the l_2 cepstral distance of the spectra to the previously defined log spectral distortion,

$$SD^2 = \sum_{n=-\infty}^{\infty} (c_n - c'_n)^2 \tag{8.14}$$

For these experiments over synthetic data, cepstral coefficients are determined

from the all pole model coefficients using the recursive relation [124] and the spectral distortion is measured using the above mentioned cepstral distance of the spectra. For three different types of input signals these experiments are performed. (Input 1: Block sparse signal, Input 2: Block sparse signal plus additive white Gaussian noise, Input 3: white Gaussian noise)

In Table 8.1 the spectral distortion measures are tabulated, using the mean of 200 frames of these three input signals.

Table 8.1. Spectral Distortion Measure over synthetic data

Inputs	Frequency	Std of noise	Spectral Distortion	
			Proposed Model	LPC
Input1	100 Hz		**1.0484**	1.0651
	200 Hz		**1.0279**	1.0660
Input2	100 Hz	0.1	**0.6155**	0.7010
		0.4	**0.2814**	0.3541
		0.6	**0.2776**	0.2989
	200 Hz	0.1	**0.6562**	0.8363
		0.4	**0.3639**	0.4320
		0.6	**0.3019**	0.3069
Input3		0.2	0.2683	**0.2432**

From the results shown in Table 8.1 it is evident that our proposed modeling method is very effective for voiced and mixed excitation signals.

8.5 Experiments over Vowel dataset

As discussed before, our proposed model can deal with all the aspects of speech: voiced, unvoiced and also the mixed excitations. The experiments in the previous section using synthetic data also endorses our claim. So in this section we will continue our experiments over the vowel dataset using the proposed model and we will compare the performance of our model with widely used LPC speech modeling technique. This

dataset has audio recording of 12 Vowels i.e /i/, /I/, /ɛ/, /æ/, /ʌ/, /a/, /ɔ/, /U/, /u/, /ɜ/, /e/, /o/ spoken by a male speaker. The sampling frequency is 16 Khz.

Table 8.2. Spectral Distortion Measure over Vowel data

Vowels	Models	
	Proposed Model	LPC
/i/	**4.1492**	4.6053
/I/	4.0753	**4.0511**
/ɛ/	4.0985	**3.8473**
/æ/	**3.7462**	3.8677
/ʌ/	**4.4092**	4.4179
/a/	**3.2895**	3.4036
/ɔ/	5.2601	**5.2598**
/U/	**4.6470**	4.8754
/u/	5.7985	**5.6795**
/ɜ/	**4.8576**	5.0481
/e/	**3.6325**	3.6431
/o/	**5.0795**	5.1003

Speech signals are quasi-stationary, so they are divided into segments within which the signal can be regarded as stationary. We will use a 20 ms window as each segment, hence it will consist of 320 samples. All pole model of order (M)=20 has been used to model each of these segments. The spectral distortion measure for each vowel is computed as the mean over all the speech segments of that vowel. For both the models, the spectral distortion measure for each vowel is tabulated in Table 8.2. For 8 cases out of 12 vowels, our model performs better than well known LPC technique in terms of spectral distortion measure. Figure 8.3 shows the estimated envelopes using both the models along with the periodogram of a speech segment of vowel /a/. One can observe that the modeling technique results in formants that do not have the peaky behavior, LPC techniques are known to suffer from.

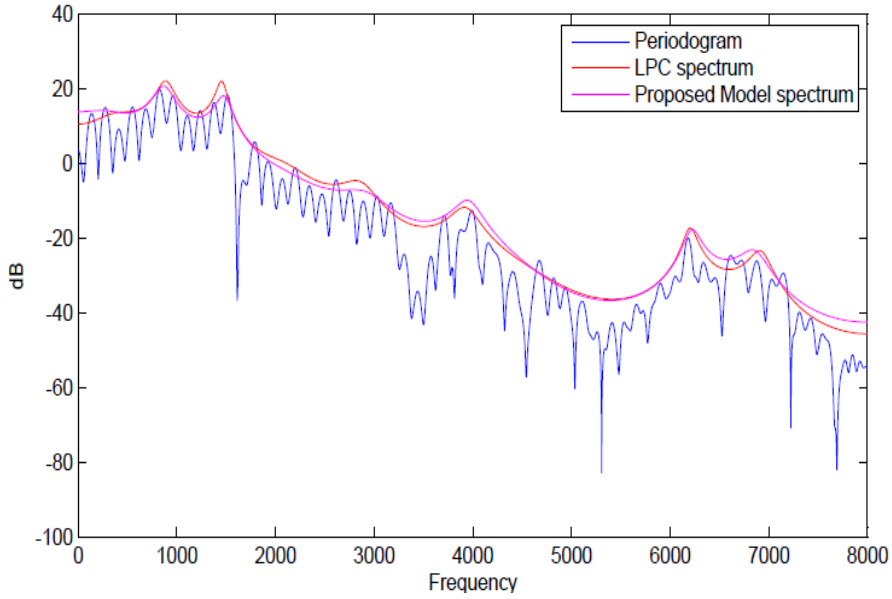

Figure 8.3. Spectrum of a segment of vowel /a/

8.6 Conclusion

In this chapter, we have proposed a novel model to reconstruct block sparse excitation from the output of an all pole filter. We have used our model for the speech modeling task and the spectral distortion measure of the estimated envelope establishes our claim, that this is a more generalized and efficient modeling approach than linear prediction. As this problem is closely related to a more general deconvolution problem, applying these models in several other applications along with theoretically establishing the optimality of this model will be the direction of the future works.

8.7 Acknowledgment

The material in this chapter, in full, is a reprint of material published as ”Block Sparse excitation based All-Pole modeling with applications to Speech”, in 39th IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP) 2014. The dissertation author was the primary researcher and B.D. Rao supervised the research

Bibliography

[1] Tulay Adali, Matthew Anderson, and Geng-Shen Fu. Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. *Signal Processing Magazine, IEEE*, 31(3):18–33, 2014.

[2] Rizwan Ahmad and Philip Schniter. Iteratively reweighted ℓ_1 approaches to sparse composite regularization. *arXiv preprint arXiv:1504.05110*, 2015.

[3] Matthew Anderson, Tülayı Adali, and Xi-Lin Li. Joint blind source separation with multivariate gaussian model: Algorithms and performance analysis. *Signal Processing, IEEE Transactions on*, 60(4):1672–1683, 2012.

[4] Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tulay Adali. Independent vector analysis, the kotz distribution, and performance bounds. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3243–3247. IEEE, 2013.

[5] Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tulay Adali. Independent vector analysis: Identification conditions and performance bounds. *Signal Processing, IEEE Transactions on*, 62(17):4399–4410, 2014.

[6] Matthew Anderson, Xi-Lin Li, and Tülay Adalı. Nonorthogonal independent vector analysis using multivariate gaussian model. In *Latent Variable Analysis and Signal Separation*, pages 354–361. Springer, 2010.

[7] Aleksandr Aravkin, James V Burke, Alessandro Chiuso, and Gianluigi Pillonetto. On the estimation of hyperparameters for empirical bayes estimators: Maximum marginal likelihood vs minimum mse. *IFAC Proceedings Volumes*, 45(16):125–130, 2012.

[8] Aleksandr Y Aravkin, James V Burke, Alessandro Chiuso, and Gianluigi Pillonetto. Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ard and glasso. *Journal of Machine Learning Research*, 15(1):217–252, 2014.

[9] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[10] Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013.

[11] Olcay Arslan. Family of multivariate generalized t distributions. *Journal of Multivariate Analysis*, 89(2):329–337, 2004.

[12] Siu-Kui Au. Connecting bayesian and frequentist quantification of parameter uncertainty in system identification. *Mechanical Systems and Signal Processing*, 29:328–342, 2012.

[13] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. Bayesian compressive sensing using laplace priors. *Image Processing, IEEE Transactions on*, 19(1):53–63, 2010.

[14] S Derin Babacan, Shigeru Nakajima, and Minh N Do. Bayesian group-sparse modeling and variational inference. *Signal Processing, IEEE Transactions on*, 62(11):2906–2921, 2014.

[15] Suhrid Balakrishnan and David Madigan. Priors on the variance in sparse bayesian learning: the demi-bayesian lasso. *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, pages 346–359, 2009.

[16] Dror Baron, Shriram Sarvotham, and Richard G Baraniuk. Bayesian compressive sensing via belief propagation. *Signal Processing, IEEE Transactions on*, 58(1):269–280, 2010.

[17] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[18] James L Beck. Bayesian system identification based on probability logic. *Structural Control and Health Monitoring*, 17(7):825–847, 2010.

[19] Alexis Benichoux, Laurent SR Simon, Emmanuel Vincent, and Rémi Gribonval. Convex regularizations for the simultaneous recording of room impulse responses. *IEEE Transactions on Signal Processing*, 62(8):1976–1986, 2014.

[20] Christian R Berger, Shengli Zhou, James C Preisig, and Peter Willett. Sparse channel estimation for multicarrier underwater acoustic communication: From subspace methods to compressed sensing. *IEEE Transactions on Signal Processing*, 58(3):1708–1721, 2010.

[21] Augustinus J Berkhout, Diemer de Vries, and Peter Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

[22] S. Bernstein. Sur les fonctions absolutement monotones. *Acta Mathematica*, 52(1), 1928.

[23] Serge Bernstein. Sur les fonctions absolument monotones. *Acta Mathematica*, 52(1):1–66, 1929.

[24] Giulio Bottegal, Aleksandr Y Aravkin, Håkan Hjalmarsson, and Gianluigi Pillonetto. Outlier robust system identification: a bayesian kernel-based approach. *IFAC Proceedings Volumes*, 47(3):1073–1078, 2014.

[25] Sébastien Bourguignon, Hervé Carfantan, and Jérôme Idier. A sparsity-based method for the estimation of spectral lines from irregularly sampled data. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):575, 2007.

[26] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[27] Richard J Butler, James B McDonald, Ray D Nelson, and Steven B White. Robust and partially adaptive estimation of regression models. *The review of economics and statistics*, pages 321–327, 1990.

[28] Emmanuel J Candès. Compressive sampling. In *Proceedings of the international congress of mathematicians*, volume 3, pages 1433–1452. Madrid, Spain, 2006.

[29] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[30] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[31] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

[32] F Carli, Tianshi Chen, Alessandro Chiuso, Lennart Ljung, and Gianluigi Pillonetto. On the estimation of hyperparameters for bayesian system identification with exponentially decaying kernels. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5260–5265. IEEE, 2012.

[33] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*, pages 3869–3872. IEEE, 2008.

[34] Tianshi Chen, Henrik Ohlsson, and Lennart Ljung. On the estimation of transfer functions, regularizations and gaussian processesrevisited. *Automatica*, 48(8):1525–1535, 2012.

[35] Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

[36] Israel Cohen, Sharon Gannot, and Baruch Berdugo. Real-time tf-gsc in nonstationary noise environments. *Israel Institute of Technology*, pages 1–4, 2003.

[37] Etienne Corteel. Equalization in an extended area using multichannel inversion and wave field synthesis. *Journal of the audio Engineering Society*, 54(12):1140–1161, 2006.

[38] Shane F Cotter and Bhaskar D Rao. Sparse channel estimation via matching pursuit with application to equalization. *IEEE Transactions on Communications*, 50(3):374–377, 2002.

[39] Shane F Cotter, Bhaskar D Rao, Kjersti Engan, and Kenneth Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *Signal Processing, IEEE Transactions on*, 53(7):2477–2488, 2005.

[40] Michelle L Daniels and Bhaskar D Rao. Compressed sensing based scalable speech coders. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 92–96. IEEE, 2012.

[41] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[42] John R Deller, John G Proakis, and John HL Hansen. *Discrete-time processing of speech signals*. Ieee New York, NY, USA:, 2000.

[43] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[44] Etienne Denoël and J-P Solvay. Linear prediction of speech with a least absolute error criterion. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(6):1397–1403, 1985.

[45] Persi Diaconis, Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

[46] David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[47] David L Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

[48] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[49] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2006.

[50] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[51] Marco F Duarte and Richard G Baraniuk. Spectral compressive sensing. *Applied and Computational Harmonic Analysis*, 35(1):111–129, 2013.

[52] L Anders Ekman, W Bastiaan Kleijn, and Manohar N Murthi. Regularized linear prediction of speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):65–73, 2008.

[53] Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[54] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

[55] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[56] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009.

[57] Torbjørn Eltoft, Taesu Kim, and Te-Won Lee. On the multivariate laplace distribution. *Signal Processing Letters, IEEE*, 13(5):300–303, 2006.

[58] Jianqing Fan, Jinchi Lv, and Lei Qi. Sparse high dimensional models in economics. *Annual review of economics*, 3:291, 2011.

[59] Mário AT Figueiredo. Adaptive sparseness for supervised learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1150–1159, 2003.

[60] Mário AT Figueiredo, José M Bioucas-Dias, and Robert D Nowak. Majorization–minimization algorithms for wavelet-based image restoration. *Image Processing, IEEE Transactions on*, 16(12):2980–2991, 2007.

[61] Dinei Florencio and Zhengyou Zhang. Maximum a posteriori estimation of room impulse responses. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 728–732. IEEE, 2015.

[62] Urs Gamper, Peter Boesiger, and Sebastian Kozerke. Compressed sensing in dynamic mri. *Magnetic resonance in medicine*, 59(2):365–373, 2008.

[63] Sharon Gannot, David Burshtein, and Ehud Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *Signal Processing, IEEE Transactions on*, 49(8):1614–1626, 2001.

[64] Sharon Gannot and Israel Cohen. Speech enhancement based on the general transfer function gsc and postfiltering. *Speech and Audio Processing, IEEE Transactions on*, 12(6):561–571, 2004.

[65] Pierre Garrigues and Bruno A Olshausen. Group sparse coding with a laplacian scale mixture prior. In *Advances in neural information processing systems*, pages 676–684, 2010.

[66] Marcel V Gerven, Botond Cseke, Robert Oostenveld, and Tom Heskes. Bayesian source localization with the multivariate laplace prior. In *Advances in Neural Information Processing Systems*, pages 1901–1909, 2009.

[67] Daniele Giacobello, Mads Græsbøll Christensen, Manohar N Murthi, Søren Holdt Jensen, and Marc Moonen. Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction. *IEEE Signal processing letters*, 17(1):103–106, 2010.

[68] Daniele Giacobello, Mads Græsbøll Christensen, Manohar N Murthi, Søren Holdt Jensen, and Marc Moonen. Sparse linear prediction and its applications to speech processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1644–1657, 2012.

[69] Daniele Giacobello, Mads Græsbøll Christensen, Manohar N. Murthi, Søren Holdt Jensen, and Marc Moonen. Sparse linear prediction and its applications to speech processing. *IEEE Transactions on Audio, Speech & Language Processing*, 20(5):1644–1657, 2012.

[70] Ritwik Giri, Bhaskar Rao, and Harinath Garudadri. Reweighted algorithms for independent vector analysis. *IEEE Signal Processing Letters*, In review, 2016.

[71] Ritwik Giri and Bhaskar D. Rao. Bootstrapped sparse bayesian learning for sparse signal recovery. In *48th Asilomar Conference on Signals, Systems and Computers, ACSSC 2014, Pacific Grove, CA, USA, November 2-5, 2014*, pages 1657–1661, 2014.

[72] Ritwik Giri and Bhaskar D Rao. Type i and type ii bayesian methods for sparse signal recovery using scale mixtures. *arXiv preprint arXiv:1507.05087*, 2015.

[73] Ritwik Giri, Bhaskar D Rao, Fred Mustiere, and Tao Zhang. Dynamic relative impulse response estimation using structured sparse bayesian learning. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 514–518. IEEE, 2016.

[74] G-O Glentis, Kostas Berberidis, and Sergios Theodoridis. Efficient least squares adaptive algorithms for fir transversal filtering. *IEEE signal processing magazine*, 16(4):13–41, 1999.

[75] Prem K Goel and Morris H Degroot. Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association*, 76(373):140–147, 1981.

[76] Irina F Gorodnitsky, John S George, and Bhaskar D Rao. Neuromagnetic source imaging with focuss: a recursive weighted minimum norm algorithm. *Electroencephalography and clinical Neurophysiology*, 95(4):231–251, 1995.

[77] Irina F Gorodnitsky and Bhaskar D Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3):600–616, 1997.

[78] Paul Gustafson. Aspects of bayesian robustness in hierarchical models. In *Bayesian robustness*, pages 63–80. Institute of Mathematical Statistics, 1996.

[79] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, pages 313–317. IEEE, 2014.

[80] Jiucang Hao, Intae Lee, Te-Won Lee, and Terrence J Sejnowski. Independent vector analysis for source separation using a mixture of gaussians prior. *Neural computation*, 22(6):1646–1673, 2010.

[81] Stefan Haufe, Vadim V Nikulin, Andreas Ziehe, Klaus-Robert Müller, and Guido Nolte. Combining sparsity and rotational invariance in eeg/meg source reconstruction. *NeuroImage*, 42(2):726–738, 2008.

[82] Lihan He and Lawrence Carin. Exploiting structure in wavelet-based bayesian compressive sensing. *Signal Processing, IEEE Transactions on*, 57(9):3488–3497, 2009.

[83] Daniel Hernández-Lobato and José Miguel Hernández-Lobato. Learning feature selection dependencies in multi-task learning. In *Advances in Neural Information Processing Systems*, pages 746–754, 2013.

[84] Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Zoubin Ghahramani. A probabilistic model for dirty multi-task feature selection.

[85] Yiteng Huang, Jingdong Chen, and Jacob Benesty. Immersive audio schemes. *IEEE Signal Processing Magazine*, 28(1):20–32, 2011.

[86] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[87] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.

[88] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *Signal Processing, IEEE Transactions on*, 56(6):2346–2356, 2008.

[89] Yuzhe Jin and Bhaskar D Rao. Algorithms for robust linear regression by exploiting the connection to sparse signal recovery. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3830–3833. IEEE, 2010.

[90] Yuzhe Jin and Bhaskar D Rao. Support recovery of sparse signals in the presence of multiple measurement vectors. *IEEE Transactions on Information Theory*, 59(5):3139–3157, 2013.

[91] Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253, 2009.

[92] Julian Keilson and FW Steutel. Mixtures of distributions, moment inequalities and measures of exponentiality and normality. *The Annals of Probability*, pages 112–130, 1974.

[93] Sean C Kerman and James B McDonald. Skewness–kurtosis bounds for the skewed generalized t and related distributions. *Statistics & Probability Letters*, 83(9):2129–2134, 2013.

[94] Taesu Kim, Hagai T Attias, Soo-Young Lee, and Te-Won Lee. Blind source separation exploiting higher-order frequency dependencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):70–79, 2007.

[95] Johannes Klein, Martin Pollow, Pascal Dietrich, and Michael Vorländer. Room impulse response measurements with arbitrary source directivity. In *40th Italian (AIA) Annual Conference on Acoustics*, 2013.

[96] Zbynek Koldovsky, Jiri Malek, and Sharon Gannot. Spatial source subtraction based on incomplete measurements of relative transfer function.

[97] Zbynek Koldovsky, Petr Tichavsky, and David Botka. Noise reduction in dual-microphone mobile phones using a bank of pre-measured target-cancellation filters. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 679–683. IEEE, 2013.

[98] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal processing magazine*, 13(4):67–94, 1996.

[99] Alexander Krueger, Ernst Warsitz, and Reinhold Haeb-Umbach. Speech enhancement with a gsc-like structure employing eigenvector-based transfer function ratios estimation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):206–219, 2011.

[100] Kai Labusch, Erhardt Barth, and Thomas Martinetz. Simple method for high-performance digit recognition based on sparse coding. *Neural Networks, IEEE Transactions on*, 19(11):1985–1989, 2008.

[101] Kenneth Lange and Janet S Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2(2):175–198, 1993.

[102] Bracha Laufer, Ronen Talmon, and Sharon Gannot. Relative transfer function modeling for supervised source localization. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 1–4. IEEE, 2013.

[103] Jong-Hwan Lee, Te-Won Lee, Ferenc A Jolesz, and Seung-Schik Yoo. Independent vector analysis (iva): multivariate approach for fmri group study. *Neuroimage*, 40(1):86–109, 2008.

[104] Te-Won Lee. *Independent component analysis*. Springer, 1998.

[105] Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer, 1998.

[106] Xi-Lin Li, Tülay Adalı, and Matthew Anderson. Joint blind source separation by generalized joint diagonalization of cumulant matrices. *Signal Processing*, 91(10):2314–2322, 2011.

[107] Yi-Ou Li, Tülay Adali, Wei Wang, and Vince D Calhoun. Joint blind source separation by multiset canonical correlation analysis. *Signal Processing, IEEE Transactions on*, 57(10):3918–3929, 2009.

[108] Yanfeng Liang, Jack Harris, Syed Mohsen Naqvi, Gaojie Chen, and Jonathon A Chambers. Independent vector analysis with a generalized multivariate gaussian source prior for frequency domain blind source separation. *Signal Processing*, 105:175–184, 2014.

[109] Yun Liang, Gang Chen, SMR Naqvi, and Jonathon A Chambers. Independent vector analysis with multivariate student's t-distribution source prior for speech separation. *Electronics Letters*, 49(16):1035–1036, 2013.

[110] Yuanqing Lin, Jingdong Chen, Youngmoo Kim, and Daniel D Lee. Blind channel identification for speech dereverberation using l1-norm sparse learning. In *Advances in Neural Information Processing Systems*, pages 921–928, 2007.

[111] Yuanqing Lin and Daniel D Lee. Bayesian regularization and nonnegative deconvolution for room impulse response estimation. *Signal Processing, IEEE Transactions on*, 54(3):839–847, 2006.

[112] Qing Ling, Zaiwen Wen, and Wotao Yin. Decentralized jointly sparse optimization by reweighted minimization. *Signal Processing, IEEE Transactions on*, 61(5):1165–1170, 2013.

[113] Chuanhai Liu and Donald B Rubin. Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, 5(1):19–39, 1995.

[114] Lennart Ljung. System identification. In *Signal Analysis and Prediction*, pages 163–173. Springer, 1998.

[115] Zhaosong Lu. Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming. *Mathematical Programming*, 147(1-2): 277–307, 2014.

[116] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic resonance in medicine*, 58(6):1182–1195, 2007.

[117] Sai Ma, Vince D Calhoun, Ronald Phlypo, and Tülay Adalı. Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis. *NeuroImage*, 90:196–206, 2014.

[118] David JC MacKay. Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11(5):1035–1068, 1999.

[119] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *Image Processing, IEEE Transactions on*, 17(1):53–69, 2008.

[120] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[121] Jiri Malek and Zbynek Koldovsky. Sparse target cancellation filters with application to semi-blind noise extraction. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2109–2113. IEEE, 2014.

[122] Dmitry Malioutov, Müjdat Çetin, and Alan S Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Transactions on Signal Processing*, 53(8):3010–3022, 2005.

[123] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[124] Richard J Mammone, Xiaoyu Zhang, and Ravi P Ramachandran. Robust speaker recognition: A feature-based approach. *Signal Processing Magazine, IEEE*, 13(5):58, 1996.

[125] Shmulik Markovich, Sharon Gannot, and Israel Cohen. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1071–1086, 2009.

[126] Donald W Marquardt and Ronald D Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

[127] James B McDonald and Whitney K Newey. Partially adaptive estimation of regression models via the generalized t distribution. *Econometric theory*, 4(03):428–457, 1988.

[128] Manohar N Murthi and Bhaskar D Rao. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *Speech and Audio Processing, IEEE Transactions on*, 8(3):221–239, 2000.

[129] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[130] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[131] Nobutaka Ono, Zbynek Koldovsky, Shigeki Miyabe, and Noboru Ito. The 2013 signal separation evaluation campaign. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6. IEEE, 2013.

[132] Jason Palmer, Kenneth Kreutz-Delgado, Bhaskar D Rao, and David P Wipf. Variational em algorithms for non-gaussian latent variable models. In *Advances in neural information processing systems*, pages 1059–1066, 2005.

[133] Jason A Palmer, Ken Kreutz-Delgado, and Scott Makeig. Probabilistic formulation of independent vector analysis using complex gaussian scale mixtures. In *Independent Component Analysis and Signal Separation*, pages 90–97. Springer, 2009.

[134] Jason A Palmer, Ken Kreutz-Delgado, and Scott Makeig. Strong sub-and super-gaussianity. In *Latent Variable Analysis and Signal Separation*, pages 303–310. Springer, 2010.

[135] Jason Allan Palmer. Variational and scale mixture representations of non-gaussian densities for estimation in the bayesian linear model: Sparse coding, independent component analysis, and minimum entropy segmentation. 2006.

[136] Raffaele Parisi, Riccardo Russo, Michele Scarpiniti, and Aurelio Uncini. Localization of audio sources by multiple binaural sensors. In *Digital Signal Processing (DSP), 2013 18th International Conference on*, pages 1–5. IEEE, 2013.

[137] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[138] Frédéric Pascal, Lionel Bombrun, Jean-Yves Tourneret, and Yannick Berthoumieu. Parameter estimation for multivariate generalized gaussian distributions. *Signal Processing, IEEE Transactions on*, 61(23):5960–5971, 2013.

[139] Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

[140] Gianluigi Pillonetto, Alessandro Chiuso, and Giuseppe De Nicolao. Prediction error identification of linear systems: a nonparametric gaussian regression approach. *Automatica*, 47(2):291–305, 2011.

[141] Gianluigi Pillonetto and Giuseppe De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

[142] Jian Pu, Yu-Gang Jiang, Jun Wang, and Xiangyang Xue. Multiple task learning using iteratively reweighted least square. In *Proceedings of the Twenty-Third*

international joint conference on Artificial Intelligence, pages 1607–1613. AAAI Press, 2013.

[143] Waqas Rafique, Syed Mohsen Naqvi, Philip JB Jackson, and Jonathon A Chambers. Iva algorithms using a multivariate student's t source prior for speech source separation in real room environments. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 474–478. IEEE, 2015.

[144] Bhaskar D Rao. Signal processing with the sparseness constraint. In *ICASSP*, volume 98, page 1, 1998.

[145] Bhaskar D Rao, Kjersti Engan, Shane F Cotter, Jason Palmer, and Kenneth Kreutz-Delgado. Subset selection in noise based on diversity measure minimization. *Signal Processing, IEEE Transactions on*, 51(3):760–770, 2003.

[146] Bhaskar D Rao and Kenneth Kreutz-Delgado. An affine scaling methodology for best basis selection. *Signal Processing, IEEE Transactions on*, 47(1):187–200, 1999.

[147] David Ruppert. What is kurtosis? an influence function approach. *The American Statistician*, 41(1):1–5, 1987.

[148] Indrayana Rustandi, Marcel Adam Just, and Tom Mitchell. Integrating multiple-study multiple-subject fmri datasets using canonical correlation analysis. In *Proceedings of the MICCAI 2009 Workshop: Statistical modeling and detection issues in intra-and inter-subject functional MRI data analysis*, 2009.

[149] J Schroeder and R Yarlagadda. Linear predictive spectral estimation via the l_1 norm. *Signal processing*, 17(1):19–29, 1989.

[150] M Schwab, P Noll, and T Sikora. Noise robust relative transfer function estimation. In *Signal Processing Conference, 2006 14th European*, pages 1–5. IEEE, 2006.

[151] Petre Stoica, Prabhu Babu, and Jian Li. Spice: A sparse covariance-based estimation method for array processing. *IEEE Transactions on Signal Processing*, 59(2):629–638, 2011.

[152] James V Stone. *Independent component analysis*. Wiley Online Library, 2004.

[153] Ronen Talmon, Israel Cohen, and Sharon Gannot. Relative transfer function identification using convolutive transfer function approximation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):546–555, 2009.

[154] Sakari Tervo, Jukka Pätynen, and Tapio Lokki. Acoustic reflection localization from room impulse responses. *Acta Acustica united with Acustica*, 98(3):418–440, 2012.

[155] Sakari Tervo and Timo Tossavainen. 3d room geometry estimation from measured impulse responses. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 513–516. IEEE, 2012.

[156] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[157] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244, 2001.

[158] Jeremy P Vila and Philip Schniter. Expectation-maximization gaussian-mixture approximate message passing. *Signal Processing, IEEE Transactions on*, 61(19):4658–4672, 2013.

[159] Jeremy P Vila and Philip Schniter. An empirical-bayes approach to recovering linearly constrained non-negative sparse signals. *Signal Processing, IEEE Transactions on*, 62(18):4689–4703, 2014.

[160] Jing Wan, Zhilin Zhang, Jingwen Yan, Taiyong Li, Bhaskar D Rao, Shiaofen Fang, Sungeun Kim, Shannon L Risacher, Andrew J Saykin, and Li Shen. Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 940–947. IEEE, 2012.

[161] Mu-Hsin Wei, Waymond R Scott Jr, and James H McClellan. Jointly sparse vector recovery via reweighted 1 minimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3929–3932. IEEE, 2012.

[162] David Vernon Widder. The laplace transform. 1946. *Zbl0139*, 29504, 1959.

[163] David Wipf and Srikantan Nagarajan. A unified bayesian framework for meg/eeg source imaging. *NeuroImage*, 44(3):947–966, 2009.

[164] David Wipf and Srikantan Nagarajan. Iterative reweighted and methods for finding sparse solutions. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):317–329, 2010.

[165] David Wipf, Jason Palmer, and Bhaskar Rao. Perspectives on sparse bayesian learning. *Computer Engineering*, 16(1):249, 2004.

[166] David P Wipf and Srikantan S Nagarajan. A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632, 2008.

[167] David P Wipf and Bhaskar D Rao. Sparse bayesian learning for basis selection. *Signal Processing, IEEE Transactions on*, 52(8):2153–2164, 2004.

[168] David P Wipf and Bhaskar D Rao. An empirical bayesian strategy for solving the simultaneous sparse approximation problem. *Signal Processing, IEEE Transactions on*, 55(7):3704–3716, 2007.

[169] David P Wipf, Bhaskar D Rao, and Srikantan Nagarajan. Latent variable bayesian models for promoting sparsity. *Information Theory, IEEE Transactions on*, 57(9):6236–6255, 2011.

[170] William S. Woods, Elior Hadad, Ivo Merks, Buye Xu, Sharon Gannot, and Tao Zhang. A real-world recording database for ad hoc microphone arrays. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015.

[171] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[172] Yi Wu and David P Wipf. Dual-space analysis of the sparse linear model. In *Advances in Neural Information Processing Systems*, pages 1745–1753, 2012.

[173] Allen Y Yang, Sudarshan Iyengar, Shankar Sastry, Ruzena Bajcsy, Philip Kuryloski, and Roozbeh Jafari. Distributed segmentation and classification of human actions using a wearable motion sensor network. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[174] Yang Yang, Yi Yang, Zi Huang, Heng Tao Shen, and Feiping Nie. Tag localization with spatial correlations and joint group sparsity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 881–888. IEEE, 2011.

[175] Atulya Yellepeddi and Dinei Florencio. Sparse array-based room transfer function estimation for echo cancellation. *IEEE Signal Processing Letters*, 21(2):230–234, 2014.

[176] Yi Zhang and Jeff G Schneider. Learning multiple tasks with a sparse matrix-normal penalty. In *Advances in Neural Information Processing Systems*, pages 2550–2558, 2010.

[177] Zhengyou Zhang, Qin Cai, and J Stokes. Multichannel acoustic echo cancelation in multiparty spatial audio conferencing with constrained kalman filtering. In *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[178] Zhilin Zhang and Bhaskar D Rao. Sparse signal recovery in the presence of correlated multiple measurement vectors. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3986–3989. IEEE, 2010.

[179] Zhilin Zhang and Bhaskar D Rao. Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning. *Selected Topics in Signal Processing, IEEE Journal of*, 5(5):912–926, 2011.

[180] Zhilin Zhang and Bhaskar D. Rao. Extension of sbl algorithms for the recovery of block sparse signals with intra-block correlation. *IEEE Transactions on Signal Processing*, 61(8):2009–2015, 2013.