BOOTSTRAPPED SPARSE BAYESIAN LEARNING FOR SPARSE SIGNAL RECOVERY

Ritwik Giri and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
rgiri@ucsd.edu, brao@ucsd.edu

ABSTRACT

In this article we study the sparse signal recovery problem in a bayesian framework using a novel Bootstrapped Sparse Bayesian Learning method. Sparse Bayesian Learning (SBL) framework is an effective tool for pruning out the irrelevant features and ending up with a sparse representation. In SBL the choice of prior over the variances of the Gaussian Scale mixture has been an interesting area of research for some time now. This motivates us to use a more generalized maximum entropy density as the prior which results in a new variant of SBL. It has been shown to perform better than traditional SBL empirically and it also accelerates the pruning procedure. Because of this advantage, this variant of SBL can be claimed as more robust choice as it is less sensitive to the threshold for pruning. Theoretical justifications have also been provided to show that the proposed model actually promotes sparse point estimates.

Index Terms— Sparse Bayesian Learning, Expectation-Maximization, Bootstrapped, Max-Entropy

1. INTRODUCTION

Sparse Bayesian Learning, using automatic relevance detection was first introduced by Tipping [1] and it has proven to be a very effective and efficient method for a variety of regression and classification problems. SBL can also be viewed as an Empirical Bayes framework, where a type-II likelihood or evidence is maximized to estimate the hyperparameters. It has been shown that the SBL cost function retains a desirable property of the ℓ_0-norm (counting measure) diversity measure (i.e., the global minimum is uniquely achieved at the maximally sparse solution under certain conditions) while often possessing a more limited constellation of local minima than MAP estimation methods [2]. In [3] SBL was first introduced for sparse recovery problem and from then on it has been used as one of the efficient models for this problem because of its huge improvement in performance over a traditional ℓ_1 minimization approaches like LASSO, reweighted ℓ_1 method etc.

Other than SBL, sparse signal recovery problem can also be viewed in a Bayesian setting as a maximum a-posteriori (MAP) solution to a regression problem with the parameters i.e. the regression coefficients having some prior sparse distribution (shown in Figure 1), which promotes sparsity. A Laplacian prior over the coefficients in a MAP setting will lead us to the same cost function as in LASSO [4]. This framework is commonly known as the Type I method. In this method using various sparse distributions over coefficients can lead us to a more sparse solution but the problem of local minima arises, as the resultant sparse penalty function is a concave function. There are some recent works which involves a Bound optimization technique over these concave penalty functions using Majorization-Minimization algorithm. But the exciting result in [5], that most of the sparse priors over the coefficient vector can be represented as a Gaussian Scale Mixture, opens up other options and leads to a Hierarchical or commonly known as Type II framework.

In initial works, it was proposed to use a non informative prior over the scaling hyperparameter in a Gaussian Scale Mixture (GSM). Though this approach performs considerably better than the traditional ℓ_1 norm minimization approach, the question still remains can we use anything better than a non-informative prior? To answer this question in some recent works exponential prior over the scaling parameters were used to connect it back with LASSO, which is named as Bayesian Lasso [6]. Demi-Bayesian Lasso has also been proposed recently, which uses the SBL's Type II maximum likelihood approach. In [7] it has been shown that SBL's Type II maximum likelihood approach is equivalent to MAP estimation where the prior on the parameters is "non-factorial", which leads to a concave penalty function that gives us more sparse solutions and because of the smoothness of the landscape, global minima can be achieved without much hindrance.

In this article we propose a maximum entropy density as the prior over the variances which uses the information learnt from the previous iteration efficiently to generate a weakly informative prior instead of a flat non informative prior. This helps us not only to converge faster but also to obtain more sparse solutions because of an extra shrinkage term in the cost function. Our proposed model is also more robust to the pruning threshold than SBL. We also show that our model is consistent with the analysis from [7], which proves that it will promote exact sparse point estimates.

The rest of the paper is organized in the following way. Section 2 summarizes the Bayesian frameworks for sparse signal recovery problem for a detailed background, Section 3 presents the model and discusses how the bootstrapped prior has been constructed, also presents the inference procedure and Section 4 provides a theoretical justification of the model and discusses why it promotes sparsity. Section 5 summarizes the performance of the proposed model over synthetic data. Finally, Section 6 concludes the paper and talks about some future directions of this work.

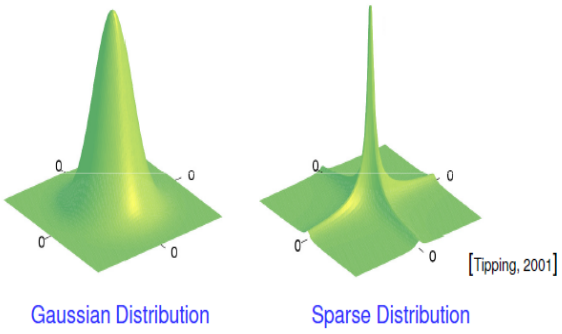

Fig. 1. Example of Sparse Distribution

2. BACKGROUND

Here, we are concerned with the following linear generative model,

$$y = \Phi x + \epsilon \tag{1}$$

where, $\Phi \in \mathbb{R}^{N \times M}$ is a dictionary of unit ℓ_2 norm basis vectors, y is the measurement vector, x is a vector of unknown weights and ϵ is uncorrelated Gaussian noise. For overcomplete dictionaries, i.e., when $M > N$ and rank(Φ)= N, the estimation problem is ill posed and sparsity constraints over the weight vector x is needed. This sparsity constraint motivates the problem to be viewed from a Bayesian point of view, which involves putting a sparse prior over x.

The estimation problem can be easily solved now by obtaining a MAP estimate of x,

$$\hat{x} = \arg\max_{x} p(y|x)p(x) \tag{2}$$

and these methods are referred as Type I methods, such as ℓ_p-quasi-norm approaches [9], FOCUSS algorithm involving Jeffreys prior [10, 11], LASSO involving a Laplacian prior [12, 9] etc.

In recent works, there has been a new approach of using a latent variable structure in a hierarchical bayes framework to represent a more complicated sparse prior over x. The sparse priors on x can be represented as, $p(x) = \int p(x|\gamma)p(\gamma)d\gamma$, allowing the random variable to be viewed in a hierarchy. The framework allows for complicated models in a simple manner and is indispensable as we move towards complex problems with structure. This prior can be written as a Gaussian scale mixture, $p(x_i) = \int N(0, \gamma_i)p(\gamma_i)d\gamma_i$ which includes the popular priors such the Laplacian and Student-t distributions. Also we have the separability constraint, i.e, $p(x) = \prod_{i=1}^{M} p(x_i)$. In estimation stage a MAP estimate of γ is sought and often justified by assuming that a non-informative prior has been employed for $p(\gamma)$. This method is referred as Type II methods which integrate out the unknown x and then solve,

$$\hat{\gamma} = \arg\max_{\gamma} p(\gamma|y) = \arg\max_{\gamma} \int p(y|x)N(0, \gamma)p(\gamma)dx \tag{3}$$

3. PROPOSED MODEL

3.1. Choice of Priors on the variance of Gaussian Scale Mixture

As we have discussed in the introduction, most of the sparse priors over the coefficients can be represented as a Gaussian Scale Mixture (GSM), $p(x_i) = \int N(x_i; 0, \gamma_i)p(\gamma_i)d\gamma_i$, where different choice of $p(\gamma_i)$ will lead to a different sparse distribution $p(x)$ such as, Laplacian, Student-t distribution etc. But the question we are trying to address here is which $p(\gamma_i)$ i.e. the prior over the variance should we choose, and is there a generic choice? In the original work of Tipping, it has been suggested to use a non informative prior over the variances, or treat them as deterministic parameters. But, we believe a better choice of this prior which has at least some information encoded in it, can lead us to much faster convergence along with more sparse coefficients.

3.1.1. How to create this prior?

1. To make this prior informative, the estimated values of hyperparameters from SBL in an empirical bayes framework can be used in an efficient way. This choice leads us to a bootstrapped version.

2. We can use that estimated values of γ as sample mean and generate a maximum entropy density as our prior.

Before we go into more details of this bootstrapped prior, we will discuss about the Maximum Entropy Density framework very briefly.

3.1.2. Maximum Entropy Distribution

A maximum entropy density is obtained by maximizing the the Shanon's Entropy measure given some moment con-

straints.

$$\max H(p(x)) = \max - \int p(x) \ln p(x) dx \tag{4}$$

with constraints, $E[\phi_j(x)] = \int \phi_j(x) p(x) dx = \mu_j$.

The solution distribution of this problem is given as,

$$p(x) \propto exp[-\sum_j \lambda_j \phi_j(x)] \tag{5}$$

3.1.3. Bootstrapped Prior

In our problem we will use the previously obtained estimates of the variances from Empirical Bayes as sample mean of these hyperparameters. We will use this sample mean as the single moment constraint and formulate the maximum entropy prior. As we are using the previously estimations to create this prior, we will name it as a Bootstrapped Prior, which is given as,

$$p(\gamma) = \prod_i \frac{1}{\gamma_i^*} exp(-\frac{\gamma_i}{\gamma_i^*}) \tag{6}$$

where, γ^* are the estimated variances from the empirical bayes framework.

3.2. Bootstrapped SBL

Here we will discuss the different stages sequentially of this Bootstrapped SBL framework:

1. Run SBL with a non informative prior over γ for few initial iterations to obtain the initial estimates γ^*.
2. Use the initial γ^* estimates to create a weakly informative prior using the maximum entropy framework, which leads to an exponential distribution (equation 7).
3. Finally run SBL in a Hierarchical Bayesian framework with the informative bootstrapped prior over γ.

3.3. Inference Procedure

In the inference procedure MAP estimates of both the coefficient vector x and γ are sought. For estimation of γ,

$$\begin{aligned} \hat{\gamma} &= \arg\max_\gamma p(\gamma|y) = \arg\max_\gamma \int p(y|x) p(x|\gamma) p(\gamma) dx \\ &= \arg\min_\gamma y^T \Sigma_y^{-1} y + \ln|\Sigma_y| + \sum_{i=1}^m f(\gamma_i) \end{aligned} \tag{7}$$

Where, $f(\gamma_i) = -2 \ln p(\gamma_i)$ and $\Sigma_y = \lambda I + \Phi \Gamma \Phi^T$, where $\Gamma = \text{diag}(\gamma)$ and λ is the variance of Gaussian noise ϵ. For estimating x we will compute the posterior, $p(x|y;\Gamma) = N(x;\mu,\Sigma)$ Where,

$$\mu = \Gamma \Phi^T (\lambda I + \Phi \Gamma \Phi^T)^{-1} y \tag{8}$$

$$\Sigma = \Gamma - \Gamma \Phi^T (\lambda I + \Phi \Gamma \Phi^T)^{-1} \Phi \Gamma \tag{9}$$

We can use $\hat{x} = \mu$ as the point estimate of the coefficient vector. To estimate γ we have to solve the optimization problem described in equation (7). Because of the space constraint we will not go into the details of the optimization procedure. For details please refer to reference [3]. Like SBL, we also treat our coefficient vector as the hidden data and employ an EM algorithm with the above discussed bootstrapped prior over the variances and the update rule of the variances has the form:

$$\gamma_j = \frac{2(\mu_j^2 + \Sigma_{jj})}{1 + \sqrt{1 + \frac{8}{\gamma_j^*}(\mu_j^2 + \Sigma_{jj})}} \tag{10}$$

4. THEORETICAL JUSTIFICATION

To show that our proposed model promotes exactly sparse point estimates, we will use the approach discussed in [13] to revert back our type II problem in a type I setting and will show that the originated penalty function satisfies the required properties that will promote sparsity.

Now using the following relationship in (7),

$$y^T \Sigma_y^{-1} y = \min_x \frac{1}{\lambda} ||y - \Phi x||_2^2 + x^T \Gamma^{-1} x \tag{11}$$

as in [13], we can show that the Type II coefficients can be obtained by solving the following problem,

$$x_{II} = \arg\min_x L_{II}(x) \tag{12}$$

where,

$$L_{II}(x) = ||y - \Phi x||_2^2 + \lambda g_{II}(x) \tag{13}$$

and,

$$g_{II}(x) = \min_\gamma \sum_i \frac{x_i^2}{\gamma_i} + \ln|\Sigma_y| + \sum_i f(\gamma_i) \tag{14}$$

with, $f(\gamma_i) = -2 \ln P(\gamma_i)$.

So using bootstrapped prior we get,

$$f(\gamma_i) = 2 \ln \gamma_i^* + 2 \frac{\gamma_i}{\gamma_i^*} \tag{15}$$

which is a concave and non decreasing function. This is a sufficient condition as shown in [13] for $g_{II}(x)$ to be a concave and non-decreasing function of $|x|$. Hence it will lead to a point sparse estimate of x, i.e the coefficients.

Now in a Type II framework after using this bootstrapped prior the cost function in γ space becomes,

$$L_{II}(\gamma) = \ln|\Sigma_y| + y^T \Sigma_y^{-1} y + \sum_j \frac{\gamma_j}{\gamma_j^*} \tag{16}$$

The key difference of this cost function from SBL is the last term, which is a result of the bootstrapped prior over γ. It can be thought of an extra shrinkage term which facilitates the pruning process to obtain sparse estimate.

5. SIMULATION RESULTS

To validate our model, we will use some synthetically generated data and we will compare the recovery performance with the traditional Sparse Bayesian Learning algorithm. Comparison of SBL with other well known Sparse signal recovery algorithms such as LASSO, reweighted ℓ_1 minimization or reweighted ℓ_2 minimization can be found in recent literatures [2, 3].

5.1. Problem Specification

We will generate the measurement vector y using a $N \times M = 25 \times 100$ dictionary Φ, whose elements are generated from a normal distribution with mean=0 and variance=1. Hence we can say that Spark measure of the dictionary matrix will be $(N + 1) = 26$. In our coefficient vector x of dimension 100 we will have randomly placed $k = 8$ non zero elements. We will present these generated measurements and the dictionary to our algorithm. The estimated coefficients will then be compared with the original x_{gen} that has been used to generate the measurement. Now as $k < \frac{N+1}{2}$ we can say that there will be a unique sparse coefficient vector x. For noisy cases we will also present the SNR to our algorithm and like SBL we will use a fixed noise variance value during the estimation stage.

5.2. Estimation Performance

Following the previously discussed experimental setting we perform these experiments for different SNR and present the averaged performance over 1000 instances for SBL and ME-SBL (Proposed method: Max Entropy SBL). Figure 2 shows the average number of non zero coefficients for both the models. Its evident that for noisy environment ME-SBL outperforms SBL as it gives a more sparse estimate by pruning the coefficient vector more efficiently, which could be because of the extra shrinkage term that has been shown in the type II cost function. Figure 3 shows the normalized mean square error for both the models and again we can see that ME-SBL outperforms SBL in noisy cases when SNR is low. This result proves that ME-SBL is actually pruning out the unnecessary coefficients which results in less number of non zero elements and also reduced normalized mean square error. We believe that for more large scale problem, that we have been working on, this performance difference will be more noticeable.

Another major advantage of the proposed model is its less sensitivity to the pruning threshold, which makes it a more robust choice.

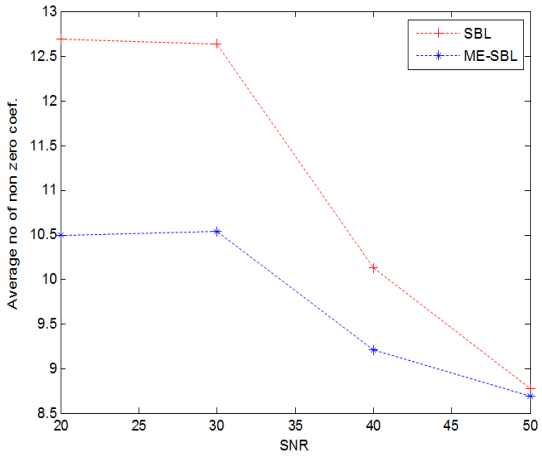

Fig. 2. Average number of non zero coefficients in the estimate

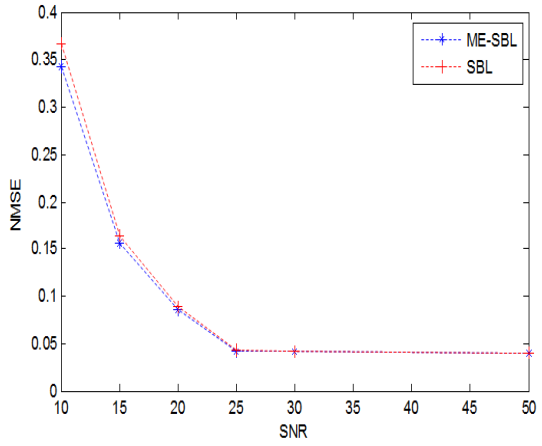

Fig. 3. Normalized mean square error plot for different SNR

6. CONCLUSION

In this paper we proposed a new variant of Sparse Bayesian Learning and addressed an important question of the choice of the prior over the variance in SBL. We have also shown theoretically, borrowing some analysis from relevant work that this bootstrapped prior promotes point sparse estimates. Experimentally also it performs better than SBL in noisy environments, as shown in our simulation results. We can also establish a connection of the proposed model with a deterministic weighted ℓ_1 norm minimization approach. A detailed analysis of this and efficient optimization algorithms will be a topic of our future works.

Acknowledgment

This research was supported by Qualcomm and National Science Foundation grant CCF-1144258.

7. REFERENCES

[1] Michael E Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.

[2] David Paul Wipf, *Bayesian methods for finding sparse representations*, ProQuest, 2006.

[3] David P Wipf and Bhaskar D Rao, "Sparse bayesian learning for basis selection," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153–2164, 2004.

[4] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[5] Jason Palmer, David Wipf, Kenneth Kreutz-Delgado, and Bhaskar Rao, "Variational em algorithms for non-gaussian latent variable models," *Advances in neural information processing systems*, vol. 18, pp. 1059, 2006.

[6] Trevor Park and George Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[7] David P Wipf, Bhaskar D Rao, and Srikantan Nagarajan, "Latent variable bayesian models for promoting sparsity," *Information Theory, IEEE Transactions on*, vol. 57, no. 9, pp. 6236–6255, 2011.

[8] 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*, pp. 346–359, 2009.

[9] 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*, vol. 51, no. 3, pp. 760–770, 2003.

[10] Cédric Févotte and Simon J Godsill, "Blind separation of sparse sources using jeffreys inverse prior and the em algorithm," in *Independent Component Analysis and Blind Signal Separation*, pp. 593–600. Springer, 2006.

[11] 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*, vol. 45, no. 3, pp. 600–616, 1997.

[12] Yuanqing Lin and Daniel D Lee, "Bayesian ℓ_1-norm sparse learning," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 5, pp. V–V.

[13] David P Wipf and Yi Wu, "Dual-space analysis of the sparse linear model.," in *NIPS*, 2012, pp. 1754–1762.