

# A novel algorithm for the Bayesian Lasso: A local approximation adjustment approach

Yuhao Li

Supervisor: A/Prof. John Ormerod and

Dr. Mohammad Javad Davouddabadi

A thesis submitted in partial fulfillment of

the requirements for the degree of

Bachelor of Science(Honour)(Data Science)

School of Mathematics and Statistics



June 2023

## **Statement of originality**

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Yuhao Li

## Abstract

Least Absolute Shrinkage and Selection Operator penalized regression, with an abbreviation of Lasso penalized regression, is regarded as a core statistical technique for simultaneous coefficient estimation and model selection. The idea of Lasso is to add the additional  $l_1$  norm penalty function to the objective function that have sum of squared residual only for ordinary regression, generating and eliminating sparse coefficient to achieve efficient and interpretable model selection.

By considering the Lasso problem from a Bayesian perspective, the Bayesian Lasso model uses a double-exponential prior to model the variation of inferential quantity and to obtain interval estimates of coefficients. Moreover, the Bayesian framework offers an automatic tuning process for the tuning parameter  $\lambda$ , which controls the strength of penalization. This process replaces the time-consuming n-fold cross-validation technique used in the ordinary Lasso.

On the other hand, obtaining the posterior of the Bayesian Lasso model involves using Monte Carlo Markov Chain methods, such as Gibbs sampler for exact posterior distribution. Even though MCMC is famous for its oracle property such as generating exact target samples if Markov Chain is run for enough iterations, it is slow and has a high computational cost.

Meanwhile, variational approximation, as a deterministic approximation algorithm for intractable posterior distribution, has been applied prevalently for fast Approximate Bayesian Inference(ABI) among the Bayesian Statistical community. It is also a faster alternative to Monte Carlo methods such as MCMC.

The concept of variational approximation involves proposing a set of known densities, and subsequently identifying the density within that set that best approximates the target. The level of approximation is evaluated using Kullback-Leibler (KL) divergence.

Nevertheless, variational approximation unlike the MCMC methods does not guarantee to approximate the exact target density. It can only find a density close to the target which means the approximation accuracy might be a pivotal concern and it fails to provide

reliable estimates of posterior variances

In order to address the slow speed issue of obtaining posterior distribution of the Bayesian Lasso problem, new alternative ABI methods would be explored, especially for deterministic algorithm such as variational Bayes.

This thesis presents two recently discovered distributions, the univariate and multivariate lasso distributions. These distributions can be used to assist in approximating marginal distributions. Additionally, the thesis introduces two highly efficient and precise variational approximation algorithms and their application in solving the Bayesian Lasso regression problem.

The first method involves matching with a marginal univariate lasso distribution by updating the global parameter for each variable per iteration. Additionally, we propose another algorithm that matches a local bivariate lasso distribution to update the global parameter for each pair of variables per iteration. This algorithm successfully addresses the issue that arise when an initial diagonal covariance matrix is assigned.

To verify the efficiency and accuracy of our algorithm, we conducted numerous experiments using real-world datasets such as the Hitters Dataset and evaluated it using several metrics such as  $l_1$  accuracy and running speed.

## **Acknowledgements**

Thanks to Supervisor and family

Finally, I would like to thank my friends my family: my mother, father, brother, grandparents, and cousins. The last couple years has been full of adversity and I could not have overcome it without your support behind the scenes.

# Contents

<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>1</b>
<b>List of Tables</b>	<b>2</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Contribution . . . . .	8
1.3 Thesis Organization . . . . .	9
<b>2 Literature Review</b>	<b>10</b>
2.1 Bayesian Inference Paradigm . . . . .	10
2.2 Least Absolute Shrinkage and Selection Operator (LASSO) penalized regression . . . . .	11
2.2.1 Lasso penalty formulation . . . . .	11
2.3 Bayesian Lasso . . . . .	12
2.3.1 Bayesian Lasso model . . . . .	12
2.3.2 Bayesian Lasso Gibbs Sampler . . . . .	14
2.4 Expectation Maximization . . . . .	17
2.4.1 Classical Expectation Maximization . . . . .	17
2.4.2 Bayesian Expectation Maximization . . . . .	18
2.4.3 Bayesian EM for Bayesian Lasso model . . . . .	19
2.5 Variational Inference . . . . .	21
2.5.1 Introduction . . . . .	21
2.5.2 KL divergence and Evidence Lower Bound(ELBO) . . . . .	21
2.5.3 Mean-Field Variational Family . . . . .	22
2.5.4 Coordinate Ascent Variational Inference (CAVI) . . . . .	22
2.5.5 MFVB for Bayesian Lasso . . . . .	24

<b>3</b>	<b>Methodology</b>	<b>27</b>
3.1	Introduction . . . . .	27
3.2	Basic Setting for the Bayesian Lasso Problem . . . . .	27
3.3	Lasso distribution . . . . .	28
3.3.1	Univariate Lasso Distribution . . . . .	29
3.3.2	Bivariate Lasso Distribution . . . . .	31
3.4	Local-Global Algorithm . . . . .	36
3.4.1	Univariate local global algorithm . . . . .	36
3.4.2	Bivariate local global algorithm . . . . .	38
<b>4</b>	<b>Experiment Result and Analysis</b>	<b>40</b>
4.1	Experimental Setting . . . . .	40
4.1.1	Parameter selection . . . . .	40
4.1.2	Evaluation metric . . . . .	40
4.1.3	Experimental datasets . . . . .	40
4.2	Experimental Result . . . . .	42
4.2.1	Approximation Accuracy Result . . . . .	42
4.2.2	Approximation Density Visualization . . . . .	42
<b>5</b>	<b>Discussion and Conclusion</b>	<b>46</b>
5.1	Limitation . . . . .	46
5.2	Future Work . . . . .	46
5.3	Conclusion . . . . .	46
	<b>Bibliography</b>	<b>47</b>

# List of Figures

1.1	Variational Inference intuition, where $X$ is data $\mathcal{D}$ , $\mathcal{D}$ is equivalent to $Q$ defined above . . . . .	5
1.2	Visualization of Mean-Field Variational Approximation compared with exact posterior when correlation is large . . . . .	7
2.1	Graphical comparison between lasso regression and ridge regression . . . .	12
3.1	Visualization of Lasso Distribution PDF for different parameter setting . .	29



# List of Tables

4.1	Experiment Result on Hitters dataset . . . . .	42
4.2	Experiment Result on Kakadu dataset . . . . .	43
4.3	Experiment Result on bodyfat dataset . . . . .	43
4.4	Experiment Result on Prostate dataset . . . . .	44
4.5	Experiment Result on Credit dataset . . . . .	44
4.6	Experiment Result on Eyedata dataset . . . . .	45

# Chapter 1

## Introduction

### 1.1 Background and Motivation

#### Introduction of Lasso Problem

The Least Absolute Selection and Shrinkage Operator (Lasso) regression proposed by [Tibshirani \(1996\)](#) belongs to one of the shrinkage methods. As one of the most traditional shrinkage methods, Lasso regression has been proven in Statistical Community over the years.

The Lasso serves two purposes. One is the estimation of regression parameter and the other is to the effective shrinking of the coefficients to achieve variable selection purpose, which is also the fundamental difference of Lasso with other methods. The Lasso regression is helpful particularly for high-dimensional because of its sparsity nature.

**Explain purpose of Lasso** The linear regression model is

$$y = \mu 1_n + X\beta + \epsilon. \quad (1.1)$$

where  $\beta$  is a  $p * 1$  vector of regression coefficient,  $y$  is a  $n * 1$  vector of response variable,  $X$  is a  $n * p$  matrix of covariates,  $\mu$  is a  $n * 1$  vector of population mean, and  $\epsilon$  is the model uncertainty which is distributed normally with mean 0 and variance  $\sigma^2$

The Least square estimator suggest the sum of square of the difference between estimated response variable and true response variable should be used as loss function as described in Equation (1.2)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} (\tilde{y} - X\beta)^T (\tilde{y} - X\beta). \quad (1.2)$$

The tuning parameter  $\lambda$  controls the strength of the penalty.

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (\tilde{y} - X\beta)^T (\tilde{y} - X\beta) + \lambda \|\beta\|_1, \quad (1.3)$$

where  $\lambda \geq 0$ .  $\tilde{y} = y - \bar{y}\mathbf{1}_n$ . Larger penalization leads to more sparse solution of regression coefficient due to a square constraint set results from  $L_1$  penalty function, resulting in implicit variable selection of parameter. This can generate a higher prediction accuracy as well as an more interpretable model since we could drop the estimated regression coefficient that has 0 and state they have weak effect for prediction according to Tibshirani (1996). However, due to non-existence of derivative of absolute value of regression coefficient  $\beta$ , alternative improved algorithm have been purposed and deployed such as least angle regression (LARS), iterative soft-thresholding, subgradient method, and iteratively reweighted least square (IRLS) by Efron et al. (2004), Beck and Teboulle (2009), Zhang and Zeng (2005) and Friedman et al. (2010).

**Bayesian Lasso** One drawback of the ordinary lasso is that it cannot properly capture the variation of inferential quantity. To address this issue, Tibshirani (1996) suggested extending the lasso estimate under the Bayesian framework. This can be achieved by assigning an independent and identically distributed Laplacian prior from Equation(1.4) and combining it with the likelihood form in Equation (1.5) to obtain the posterior mode Equation(1.6).

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmax}} P(\beta|y, \sigma^2, \tau), \quad \tau = \frac{\lambda}{2\sigma^2} \quad (1.4)$$

$$f(\beta_j) = \left(\frac{\tau}{2}\right) \exp(-\tau|\beta_j|), \quad (1.5)$$

$$p(y|\beta, \sigma^2) = N(y|X\beta, \sigma^2 I_n). \quad (1.6)$$

**Introduction of Bayesian Lasso Problem (Why Bayesian Lasso?)** Nevertheless, there is no tractable integration form for the Bayesian Lasso posterior until Park and Casella (2008) further explore the Lasso model under the setting of Bayesian framework. The choice of a conditional Laplace prior distribution over the regression coefficient  $\beta$  conditioning by standard error  $\sigma^2$  is added to the Lasso penalty formulation in the frequentist framework. This would ensure the unimodality of full posterior distribution. Based on closed form of tractable posterior distribution, a three-step Gibbs sampler is proposed to draw samples from Bayes Lasso posterior distribution, which can be utilized for further inference of parameter of interest. There are several benefits for using the Bayesian Lasso model. Firstly, it has easier implementation than the traditional Lasso, although it demands more computation of conditional distribution form is demanded. Secondly, we can generate

Bayesian credible intervals simultaneously for the parameters in a model, allowing for the modeling of uncertainty and guiding variable selection. Thirdly, [Park and Casella \(2008\)](#) also state that the Bayesian Lasso model could be a potential solution for addressing the issue of attaining optimal tuning parameter  $\lambda$ . We can achieve this by using the marginal maximum likelihood method along with a suitable hyperprior, such as a gamma prior on the square of the tuning parameter ( $\lambda^2$ ). It could support a more stable automatic tuning process of choosing the most appropriate tuning parameter  $\lambda$  for the Lasso problem. This is in contrast to the inefficient  $K$ -fold cross-validation approach under the ordinary Lasso model, which is time-consuming and computationally demanding. Lastly, the three-step Bayesian Lasso Gibbs Sampler proposed by [Park and Casella \(2008\)](#) would yield an exact posterior distribution that can be sampled given exact form of conditional distribution of each model parameter conditioning on rest of the other parameters. Theoretically, [Khare and Hobert \(2013\)](#) demonstrated a Bayesian Lasso Gibbs sampler version of central limit theorem (CLT), indicating Bayesian Lasso Gibbs sampler satisfies geometric ergodicity for any values of sample size  $n \geq 3$  and an arbitrary number of regression coefficient  $p$ , data matrix  $X$ , tuning parameter  $\lambda$ . This means, the Bayesian lasso Gibbs sampler is able to achieve asymptotically uncertainty of posterior estimation. To address this issue, [Rajaratnam and Sparks \(2015a\)](#) invented a reduced step Gibbs sampler instead, successfully accelerating the sampling procedure.

**Approximate Bayesian Inference Intuition** The Review of Bayesian inference intuition can be referred from Section 2.1, but challenges of Bayesian inference motivates the approximate Bayesian inference method thereafter. To be specific, [Bishop \(2006\)](#) states three main challenges of obtaining posterior distribution. Firstly, the dimension of target parameter might be high, which results in heavy computational cost for estimating posterior distribution. Secondly, it takes a long time to converge to the target posterior distribution. Thirdly, the computation of the posterior mean parameter  $\int_{\theta} \theta p(\theta|\mathcal{D})$  has high probability without having a simple calculation, resulting in the fact that there might not exist a closed form analytical solution for integration. Much efforts have been paid over the years, there are two main types of sampling approach that are effective currently, which are stochastic sampling algorithms and deterministic approximation algorithms.

**Introduction of ABI method** There are two genres of ABI methods, which include stochastic approaches such as MCMC, where an exact result can be obtained if infinite computational resource is assigned. Another category lies in deterministic approaches, which provide a faster substitution compared to stochastic approximation approaches. As stated above, an approximate inference methods such as MCMC are used for posterior distribution estimation. The need for approximate Bayesian inference methods arises from the difficulty and impracticality of estimating the exact posterior distribution parameters.

**Challenges of Bayesian Lasso model** Three-step Gibbs sampler belongs to the class of MCMC algorithms, which produces samples demonstrates gradually decreasing correlated samples as the burn-in period is increased. Burn-in period refers to the initial period during which the sample is discarded or ignored to eliminate any potential biases or inconsistencies in the measurement process. The purpose of the burn-in period is to ensure that the measurements are stable and consistent before beginning the actual analysis. The three-step Gibbs sampler is time-consuming and computational challenging, given the fact that it normally requires a long time to converge inside the interval of an acceptable tolerance.

**Approximation Algorithm: Deterministic type & VI.** As mentioned before, due to the limitations of stochastic algorithms, alternative methods such as deterministic Variational Inference (VI) have become popular due to their fast speed and simple computation. Numerous algorithms have been designed and utilized widely such as VB, Expectation Propagation algorithms etc. A common variation is the Coordinate Ascent Variational Inference approach produced by [Blei et al. \(2003\)](#), which assumes that the approximation originates from an analytically tractable class of distribution  $Q$ . Afterwards it attempts to search for the distribution from this family that is closest to the target posterior distribution with some discrepancy metric, such as the Kullback-Leibler divergence. An optimization based system in Equation (1.7) is established by iteratively updating variational parameter with an appropriate optimization algorithm. A traditional optimization algorithm can be Coordinate Ascent that could obtain approximated posterior distribution in the family of  $Q$ , while the most common choice is the Normal Distribution due to its simple form and adaptability to other distribution. To further illustrate the intuition, Figure 1.1 pro-

vides further explanation of aforementioned intuition, goal and procedure of Variational Inference.

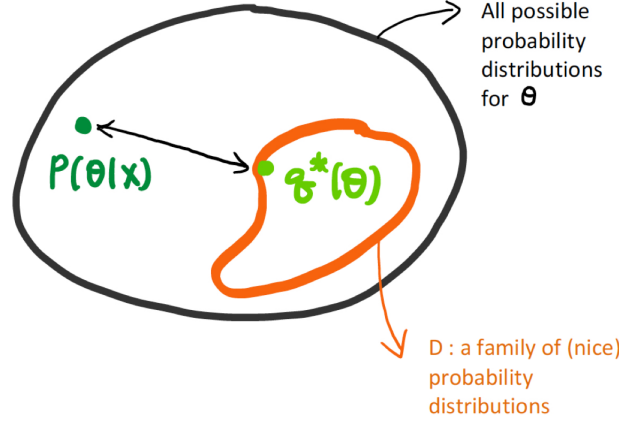


Figure 1.1: Variational Inference intuition, where  $X$  is data  $\mathcal{D}$ ,  $\mathcal{D}$  is equivalent to  $Q$  defined above

$$q^*(\theta) = \operatorname{argmin}_{q \in Q} \text{KL}(q(\theta) || p(\theta | \mathcal{D})) := \int q(\theta) \log\left(\frac{q(\theta)}{p(\theta | \mathcal{D})}\right) d\theta \quad (1.7)$$

$$\text{KL}(q || p(\cdot | \mathcal{D})) = - \int q(\theta) \log\left(\frac{p(\theta) p(\mathcal{D} | \theta)}{q(\theta)}\right) d\theta + \log p(\mathcal{D}). \quad (1.8)$$

In addition, the exact form of KL divergence can be found in Equation (1.8). In practice, minimizing the KL divergence from Equation 1.7 is difficult due to its complexity, and therefore it is often converted into an equivalent formulation in Equation 1.9 that maximizes the lower bound of  $\log(p(y))$ . This lower bound is known as the Evidence Lower Bound (ELBO), and can be defined by Equation 1.9).

$$q^*(\theta) = \operatorname{argmax}_{q \in Q} \text{ELBO}(q(\theta)), \quad (1.9)$$

$$\begin{aligned} \text{ELBO}(q(\theta)) &= \int q(\theta) \log\left(\frac{p(\theta) p(\mathcal{D} | \theta)}{q(\theta)}\right) d\theta = \mathbb{E}_{q(\theta)} \log\left(\frac{p(\theta) p(\mathcal{D} | \theta)}{q(\theta)}\right). \\ &= \mathbb{E}_{q(\theta)} [\log p(\theta, \mathcal{D})] - \mathbb{E}_{q(\theta)} [\log q(\theta)] \\ &= \mathbb{E}_{q(\theta)} [\log p(\theta, \mathcal{D})] - \text{KL}(q(\theta) || p(\theta)) \end{aligned} \quad (1.10)$$

## Mean Field Variational Bayes, History and Introduction

The most traditional Variational Inference algorithm is known as MFVB motivated by mean-field theory in statistical physics, as proposed by [Parisi and Shankar \(1988\)](#). The algorithm assumes that the approximated distribution is a product of independent parameter distributions from set  $Q$ , as described in [Equation 1.11](#), assuming there are  $k$  sub-parameters of the target parameter  $\theta$ .

$$q(\theta) = \prod_{i=1}^k q_i(\theta) \quad (1.11)$$

The MFVBs have been adapted and developed over the last two decades, especially in mixture modelling and probabilistic graphical modelling. It sometimes provides efficient variable selection as well, We will introduce more about the algebra of MFVB in Chapter

**2. Advantages of Variational Inference** One is that Variational Inference algorithm shows a descent computation cost and scalability, and thus Variational inference can be adaptive when the amount of data is huge. For instance, if there exists a billion image that requires to be fitted into probabilistic machine learning model, then exact precision method such as MCMC will be computataional demanding, while Variational Inference would sacrifice tiny accuracy with hundreds of time faster speed as return. Secondly, the descent time-efficiency of Variational Inference becomes another significant factor why it is popular, given the fact it only involves updating variational parameters iteratively until convergence, as opposed to MCMC that produce correlated samples that limit the ideal behavior of MCMC algorithm.

**Drawbacks of VI** Meanwhile, disadvantages of Variational Bayes include inexact approximation result under some scenarios, although it could capture marginal density. For example, it is suggested by [Blei et al. \(2017\)](#), that Variational Inference algorithm might underestimate the covariance between parameter of interest, if inter-parameter correlation is strong. It tends to ignore the correlation between parameters, results in unideal behavior as a resultt. [Figure 1.2](#) further demonstrates this phenomenon, the ture overall posterior of  $x_2$  and  $x_1$  have a exploded correlation with a eclipse-shaped density, while a circled-shape mean-field approximation is established instead due to its product density family limitation. We will expand properties and derivation of Variational Inference more in subsection [2.5](#).

Overall, Variational Inference has proven its effectiveness in distinct application fields

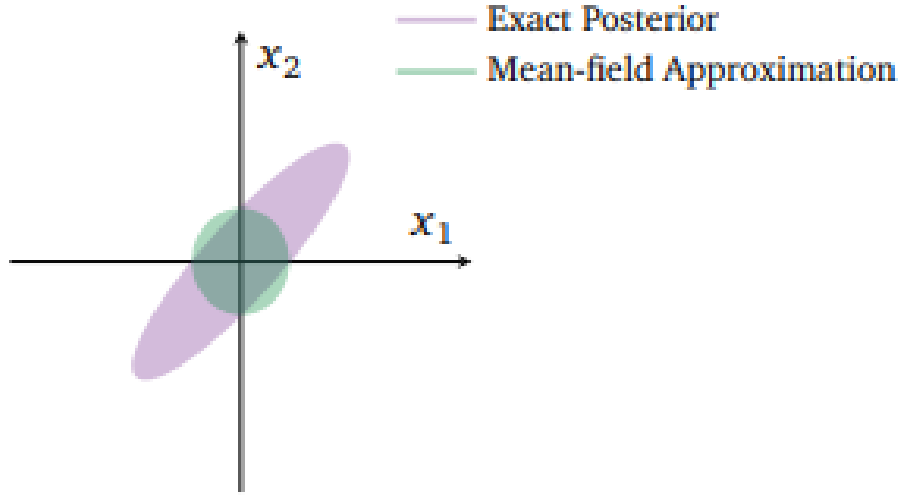


Figure 1.2: Visualization of Mean-Field Variational Approximation compared with exact posterior when correlation is large

such as speech recognition and document retrieval in natural language processing, computer vision etc. Despite small disadvantages of Variational Inference, the potential of variational approximation haven't been fully discovered yet by researchers, its ability to provide reliable posterior estimate is invaluable in the future in the era of big data and deep learning nowadays.

**Motivation** This study is motivated by the desire to improve the approximation accuracy of Variational Bayes and to take advantage of the oracle property of Bayesian Lasso regression coefficient estimation for variable selection and standard error estimation, there is an increasing demand for fast approximate inference. We would like to design new VI based algorithms for the Bayesian Lasso regression problem, for the purpose of obtaining Bayesian lasso posterior distribution in a much faster and more accurate manner. We fed the initial values of the mean  $\mu$  and covariance  $\Sigma$  and wrote out the marginal likelihood form of the regression coefficient  $\beta_j$  for the  $j_{th}$  variable. Using this information, we invented a new distribution called the univariate Lasso distribution to match the marginal likelihood, represented by  $\log(p(\mathcal{D})|\beta_j)$ , we have invented a new distribution called univariate lasso distribution for matching the marginal likelihood. Mixing each marginal likelihood with a gaussian approximation. We have demonstrated that our algorithm's approximation accu-



racy surpasses that of every existing algorithm, including MFVBs. Even though the speed of our algorithm is slightly slower than MFVB, the approximation accuracy illustrates a small gap between exact estimation from MCMC, with hundred times faster time complexity. Nevertheless, there is a drawback of this method, given the fact that the global covariance matrix would remain diagonal if the initial global covariance matrix is diagonal. To remedy this issue, we have also purposed another algorithm based on marginal likelihood estimation by a bivariate Lasso distribution. Instead of updating corresponding mean, covariance matrix for each variable in each iteration, the marginal likelihood of each pair of variables would be matched, so that further generalize our algorithm. Our conclusions are the univariate Lasso algorithm is faster with a lower accuracy while bivariate Lasso algorithm is slower with a higher accuracy, since it updates each pair of variables at a time resulting in  $\binom{p}{2}$  of unique pairs. We will show the full intuition and idea later in Chapter 3. By utilizing and fitting a both univariate and bivariate Lasso distribution to each of the marginal distribution, an improved estimate for global gaussian approximation can be obtained as defined in Equation (1.12). Finally, we will show our experiment result in Chapter 4 using various accuracy metrics.

Our contribution have been listed in the following subsection 1.2.

$$q^*(\theta) \approx N(\mu^*, \Sigma^*) \tag{1.12}$$

## 1.2 Contribution

Our main contribution could be concluded as the following parts:

- Introduction of univariate and multivariate Lasso distribution
- Derivation of properties for Univariate Lasso distribution, such as the expectation, variance, cumulative density function form etc.
- Derivation of properties for multivariate Lasso distribution such as the expectation, variance, cumulative density Function form etc.
- Implementation of univariate Lasso distribution and multivariate Lasso distribution property in R.

- Design of two new VI approaches based on local approximation by univariate lasso distribution and multivariate lasso distribution respectively.
- Conduct of experiment to testify two algorithms under dataset by several evaluation metrics for approximation accuracy such as Hitters dataset etc.

## 1.3 Thesis Organization

The thesis is organized as follows. Chapter 1 briefly illustrates the motivation and background of the Lasso problem, Bayesian Lasso Problem, and Approximate Bayesian Inference, with a specific focus on deterministic Variational Approximation. Chapter 2 briefly reviews and explains the details of the methods in previous work such as the Lasso problem, Approximate Bayesian Inference algorithm, MCMC, Bayesian Expectation Maximization algorithm and their variants and MFVB. We present our main methodology of variational algorithm in Chapter 3, followed by a comprehensive experiment for testing the effectiveness of algorithm in Chapter 4.

# Chapter 2

## Literature Review

### 2.1 Bayesian Inference Paradigm

**Why Bayesian** Bayesian inference approaches offer numerous advantages in the statistical community and application areas, especially in situations where data is scarce. In cases where effective data is extremely rare and insufficient, an appropriate prior choice can provide significant benefits, particularly in medical problems. Unlike frequentist inference approaches that treat parameter estimates as fixed values, Bayesian inference approaches regard parameter estimates as random variables that have probability distributions. This unique feature allows for interval estimates and error variances to be generated, providing a more comprehensive understanding of uncertainty and increasing confidence in interpreting parameter estimates. By incorporating prior knowledge and probability distributions, Bayesian Inference approaches offer a more flexible and intuitive framework for statistical analysis.

**Bayesian Inference Intuition** Bayesian inference approach stems from the Bayes rule, which is defined as Equation (2.1) based on theory developed by [Beech et al. \(1959\)](#). Suppose  $\theta$  is our model parameter of interest,  $\mathcal{D}$  is data, then  $p(\theta)$  is known as prior distribution, which offers pre-existing knowledge or information about  $\theta$ . Posterior distribution  $p(\theta|\mathcal{D})$  refers to the likelihood conditioning on the data  $\mathcal{D}$ . Incorporating information from current data and prior knowledge, posterior distribution can be then inferred and simplified to Equation (2.2), as the value of  $p(\mathcal{D})$  is constant and insignificant for determining the overall posterior distribution.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \quad (2.1)$$

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta), \quad (2.2)$$

## 2.2 Least Absolute Shrinkage and Selection Operator (LASSO) penalized regression

### 2.2.1 Lasso penalty formulation

The constraint form of lasso can be shown by Equation 2.3, where  $t \geq 0$  is denoted as a tuning term, regression coefficient is  $\beta$ ,  $\|\beta\|_1$  is the  $l_1$  norm of beta,  $\|y - X\beta\|_2$  is the  $l_2$  norm: of residual value, data matrix is  $X$ , response variable is  $y$ . The estimation for lasso estimate  $\hat{\beta}_{lasso}$  is defined by Equation 2.3.

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2, s.t. \|\beta\|_1 \leq t, t \geq 0. \quad (2.3)$$

In order to transform constraint form of lasso to penalty form, Lagrange multiplier method, as a pivotal technique from transforming a constraint optimization system into an unconstrained penalty formulation of system has been used. The Lagrangian function for constrained Lasso Regression is constructed by Equation 2.4

$$\mathcal{L}(\beta, \lambda) = \|y - X\beta\|_2 + \lambda \|\beta\|_1 - \lambda t, \lambda \geq 0 \quad (2.4)$$

Since the objective function contains a quadratic term  $\|y - X\beta\|_2$  with a linear term  $\lambda \|\beta\|_1 - \lambda t$ , leading to a convex optimization problem. Due to strong duality theorem in convex optimization system, therefore the penalty formulation of lasso regression can be deduced as Equation 1.3, is equivalent to constraint form Equation 2.3 after ignoring the unaffected constant  $-\lambda t$ .

Graphical demonstration of the lasso for Equation 2.3 and Equation 1.3 can also be found on the left hand side of the Figure 2.1, where the squared constraint set is drawn, in addition to the contour line of regression coefficient. The penalty term  $\lambda$  controls the strength of the penalization in Lasso regression. When the value of  $\lambda$  is set higher, a more sparse solution is facilitated. This forces the estimated coefficients to lie closer to the axis of each parameter, as shown in 2.1. As a result, Lasso regression coefficients are more likely to intersect with the corners of the squared constraint set, leading to the occurrence of sparse estimated regression coefficients. By encouraging sparsity, Lasso regression provides a useful tool for variable selection and reducing model complexity, leading to more interpretable and



Figure 2.1: Graphical comparison between lasso regression and ridge regression

generalizable models. On the other hand, ridge regression uses a  $l_2$  penalty to estimate the regression coefficients, but it tends to gain a non-sparse solution due to circled constraint set for  $\beta$ .

## 2.3 Bayesian Lasso

### 2.3.1 Bayesian Lasso model

[Park and Casella \(2008\)](#) proposed an alternative formula for the conditional Laplace prior, which takes the form of [Equation 1.5](#) and is expanded in [Equation 2.5](#). This approach offers a Bayesian interpretation of the Lasso penalty and provides a framework for incorporating prior information into the variable selection process. By using the conditional Laplace prior, the Bayesian Lasso regression model can be tuned to strike a balance between sparsity and estimation accuracy, resulting in a more robust and interpretable model.

$$\pi(\beta) = \prod_{j=1}^p \frac{\lambda}{2} e^{-\lambda|\beta_j|} \quad (2.5)$$

For a given variance, the mode of posterior form in Equation 2.6 is consistent with the estimate of lasso equation in Equation 1.3, but it will hinder the bayesian interpretation, inference and variable selection since the bayesian predictive distribution make future inference via a posterior mean instead of posterior mode. In addition, if a variance is unknown, the posterior will be a multimodal distribution, the derivation has been provided by the appendix from Park and Casella (2008).

$$\pi(\beta, \sigma^2 | \tilde{y}) \propto \pi(\sigma^2) (\sigma^2)^{-(n-1)/2} \exp\left(\frac{1}{2\sigma^2} (\tilde{y} - X\beta)^T (\tilde{y} - X\beta) - \lambda \sum_{j=1}^p |\beta_j|\right) \quad (2.6)$$

$$\pi(\beta | \sigma^2) = \prod_{j=1}^p \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \quad (2.7)$$

To remedy this issue, a conditional Laplacian prior from 2.7 with respect to Equation(2.5) has been designed, ensuring the unimodality of the posterior for  $\beta$ , and the current prior with respect to  $\beta$ ,  $\sigma^2$  can be written as

$$\pi(\beta, \sigma^2) \propto \pi(\sigma^2) \prod_{j=1}^p \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}, \quad (2.8)$$

which can result in the unimodal joint posterior distribution  $\pi(\beta, \sigma^2 | \tilde{y})$  of  $\beta$  and  $\sigma^2 > 0$  under the new prior 2.8, given an improper prior selection for  $\pi(\sigma^2) = \frac{1}{\sigma^2}$  and  $\lambda \geq 0$ . Additionally, an additional latent variable  $\tau$  is introduced as a scale mixture of Gaussians for reformulation of conditional prior 2.7 as 2.9, which can be regarded as corresponding weight assigned to each regression coefficient. If  $\tau_j$  goes to 0 then the corresponding regression coefficient will be shrunk towards zero accordingly.

$$\frac{\alpha}{2} e^{-\alpha|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{\alpha^2}{2} e^{-\alpha^2 s/2} ds, \alpha > 0 \quad (2.9)$$

Finally, the hierarchical bayesian Lasso model functional form can be written as Equation (2.10).

$$\begin{aligned} y | \mu, X, \beta, \sigma^2 &\sim N_n(\mu + X\beta, \sigma^2 I) \\ \beta | \tau_1^2, \dots, \tau_p^2 &\sim N_p(0, \sigma^2 D_\tau) \\ D_\tau &= \text{diag}(\tau_1^2, \dots, \tau_p^2) \\ \tau_1^2, \dots, \tau_p^2 &\sim \prod_{j=1}^p \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \tau_1^2, \dots, \tau_p^2 > 0 \\ \sigma^2 &\sim \pi(\sigma^2) = 1/\sigma^2, \sigma^2 > 0 \end{aligned} \quad (2.10)$$

## 2.3.2 Bayesian Lasso Gibbs Sampler

### Gibbs Sampler

[Geman and Geman \(1984\)](#) introduced a special case of the Metropolis-Hastings algorithm called the Gibbs sampler. As a Markov Chain Monte Carlo sampling algorithm, it can be used for efficient sampling of any probability density function, given the posterior form from the corresponding conditional distribution. In each iteration, each parameter of interest is sampled once by conditional distribution for the current iteration. After running the chain for long enough iterations, it returns posterior distribution samples with descent approximation accuracy after discarding samples in burn-in period. Notice that, the functional form of conditional distribution given any other parameter of interest has to be acquired. In this study, we need to have the marginal distributions of  $(\beta, \sigma^2, \tau)$ :

$$\begin{aligned} p(\beta|\mathcal{D}, \sigma^2, \tau^2) \\ p(\sigma^2|\mathcal{D}, \beta, \tau^2) \\ p(\tau|\mathcal{D}, \beta, \sigma^2) \end{aligned}$$

After getting functional form and ignoring the normalizing constant, we can infer their category of probability distribution for each expression, and we can sample from the corresponding distribution.

**Initial Setting:** In the initial setting we consider that:  $\lambda$  is fixed, and  $\sigma^2$  has a Gamma distribution with parameter  $a$  and  $b$ :  $\pi(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-b/\sigma^2}$ ,  $\sigma^2 > 0, a > 0, b > 0$

Our first step is to write the joint distribution:  $p(\beta, \tau^2, \sigma^2, \mathcal{D})$

**Joint distributional form:** Given [Equation 2.10](#), we can write the joint distribution as

$$\begin{aligned} p(\tilde{y}, \beta, \tau^2, \sigma^2) &= p(\tilde{y}|\beta, \sigma^2, \tau)p(\sigma^2) \prod_{j=1}^p p(\beta|\sigma^2, \tau_j)p(\tau_j^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{-(\tilde{y}-X\beta)^T(\tilde{y}-X\beta)}{2\sigma^2}} \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} e^{-b/\sigma^2} \prod_{j=1}^p \frac{1}{(2\sigma^2\tau_j^2)^{1/2}} e^{-\frac{-1}{2\sigma^2\tau_j^2}\beta_j^2} \frac{\lambda^2}{2} e^{-\lambda^2\tau_j^2/2} \end{aligned} \quad (2.11)$$

**Full conditional distribution of  $\beta$ :**

$$p(\beta|\tilde{y}, \tau, \sigma^2) \propto p(\tilde{y}, \beta, \tau, \sigma^2) \quad (2.12)$$

Recognizing the term without  $\beta$  as constant, the conditional distribution of  $\beta$  can be simplified to

$$\begin{aligned} p(\beta|\tilde{y}, \sigma^2, \tau^2) &\propto \exp\left(\frac{\beta^T X^T X \beta - 2\tilde{y}^T X \beta + \lambda^2 \beta^T A^{-1} \beta}{-2\sigma^2}\right), A = \text{diag}(\tau) \\ &= \exp\left(-\frac{1}{2}\beta^T \left(\frac{X^T X + \lambda^2 A^{-1}}{-2\sigma^2}\right)\beta + \frac{\tilde{y}^T X \beta}{\sigma^2}\right) \\ &\sim \text{MVN}(\mu^*, \Sigma^*) \end{aligned} \quad (2.13)$$

$$\text{where } \mu^* = (X^T X + \lambda^2 A^{-1})^{-1} X^T \tilde{y}, \Sigma^* = (X^T X + \lambda^2 A^{-1})^{-1} \sigma^2$$

So we can sample  $\beta$  from a multivariate normal distribution with its corresponding mean and variance.

**Full conditional distribution of  $\sigma^2$ :**

$$\begin{aligned} p(\sigma^2|\tilde{y}, \beta, \tau^2) &\propto p(\tilde{y}, \beta, \tau^2, \sigma^2) \\ &= (\sigma^2)^{-\frac{n}{2}-\frac{p}{2}-a-1} \exp\left(-\frac{1}{2\sigma^2}(\tilde{y} - X\beta)^T(\tilde{y} - X\beta) + \frac{1}{2\sigma^2}\beta^T D_\tau \beta + \frac{b}{\sigma^2}\right). \\ &\sim \text{Inverse-Gamma}(\alpha^*, \beta^*) \end{aligned} \quad (2.14)$$

$$\text{where } \alpha^* = \frac{n}{2} + \frac{p}{2} + a, \beta^* = (\tilde{y} - X\beta)^T(\tilde{y} - X\beta)/2 + \beta^T D_\tau \beta/2 + b$$

**Full conditional distribution of  $\tau_j^2$ :**

$$\begin{aligned} p(\tau_j^2|\tilde{y}, \beta, \sigma^2) &\propto p(\tilde{y}, \beta, \tau^2, \sigma^2) = \frac{1}{\sqrt{\frac{2\pi\sigma^2\tau_j}{\lambda^2}}} \exp\left(-\frac{\beta_j^2 \lambda^2}{2\sigma^2 \tau_j}\right) \exp\left(-\frac{1}{2}\tau_j\right) \\ &\sim GIG(a^*, b^*, p^*) \end{aligned}$$

$$\text{where GIG is generalized inverse gaussian distribution with parameters } a^* = 1, b^* = \frac{\beta_j^2 \lambda^2}{\sigma^2}, p = \frac{1}{2} \quad (2.15)$$

**Summary** The gibbs sampler can be established by the following algorithm.

**Automatic selection of the penalty parameter  $\lambda$ :** Common choice of penalty parameter  $\lambda$  in the non-bayesian paradigm involves cross-validation approach, which is time-consuming and computational challenging especially for large datasets. [Park and](#)



---

**Algorithm 1** Gibbs Sampler for the Bayesian Lasso

---

```
1: Given  $\lambda^2 > 0, \tau^{(1)} = \mathbf{1}_n, \sigma^{2(1)} = 1, t = 1$  ▷ Initial Setting
2: while  $t \leq 10^5$  do
3:   Sampling  $\beta^{(t+1)} \sim \text{MVN}((X^T X + \lambda^2 A^{-1})^{-1} X^T y, (X^T X + \lambda^2 A^{-1})^{-1} \sigma^2)$  ▷ Generate sample  $\beta$ 
4:   Sampling  $\sigma^{2(t+1)} \sim IG(\frac{n}{2} + \frac{p}{2} + a, \frac{\|y - X\beta\|_2^2}{2} + \frac{\lambda^2 \sum_j \beta_j^2}{2\tau_j} + b)$  ▷ Generate sample  $\sigma^2$ 
5:   for  $j=1, \dots, p$  do
6:     Sampling  $\tau_j^{2(t+1)} \sim GIG(1, \frac{\beta_j^2 \lambda^2}{\sigma^2}, 1/2)$  ▷ Generate sample  $\tau_j$ 
7:    $t \leftarrow t + 1$ 
8: return  $\beta, \sigma^2, \tau^2$ 
```

---

Casella (2008) set a hyperprior to  $\lambda^2$  as 2.16, instead of  $\lambda$  to facilitate conjugacy. According to Park and Casella (2008), there are some additional notification of choosing prior, which involves: firstly, to avoid mixing issue, the prior distribution for  $\lambda^2$  should reach zero asymptotically with a descent speed as  $\lambda^2$  goes to infinity, Secondly, the density at maximum likelihood estimate should be assigned with enough probability density with a overall flat distribution.

$$\pi(\lambda^2) = \frac{\delta^\gamma}{\Gamma(\gamma)} (\lambda^2)^{\gamma-1} e^{-\delta \lambda^2}, \text{ for } \delta > 0, \gamma > 0, \lambda^2 > 0 \quad (2.16)$$

The penalty parameter is the extent of penalization of non-zero coefficient, which is also a compromise between model simplicity and fitting capability to data in the frequentist lasso setting. According to the posterior form of  $\tau_j$ ,  $\lambda$  controls the shape of generalized inverse-Gaussian posterior distribution of  $\tau_j$  as shown before.

To obtain the posterior form of  $\lambda$ , we need to incorporate a proper hyperprior distribution to the joint distribution  $p(y, \beta, \sigma^2, \tau)$ . First, assuming the prior of  $\lambda^2$  is with shape and rate parameters  $\theta$  and  $\gamma$ , respectively.

$$\begin{aligned} p(\lambda^2 | \tilde{y}, \beta, \sigma^2, \tau^2) &\propto p(\tilde{y}, \beta, \tau^2, \sigma^2, \lambda) \\ &= \left( \prod_{j=1}^p \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2 / 2} \right) (\lambda^2)^{\gamma-1} e^{-\delta \lambda^2} \\ &= (\lambda^2)^{p+\gamma-1} e^{-\lambda^2 (\frac{1}{2} \sum_{j=1}^p \tau_j^2 + \delta)} \end{aligned} \quad (2.17)$$

Thus, the posterior distribution of  $\lambda^2$  is still following a Gamma distribution, with a shape

parameter  $p + \gamma - 1$  and rate parameter  $\sum_{j=1}^p \tau_j^2 + \delta$ . The  $\lambda^2$  can be sampled by 2.17, based on using an augmented Gibbs sampler.

## 2.4 Expectation Maximization

Even though the posterior distribution can be sampled by Gibbs sampler in the last subsection, the sparsity nature of the Bayesian lasso is not captured by posterior mean given by Gibbs Sampler. The posterior mode calculated by Bayesian Expectation Maximization, however, could capture the posterior mode and preserve the sparsity feature of the basic lasso.

### 2.4.1 Classical Expectation Maximization

The expectation maximization (EM) algorithm was proposed by Dempster et al. (1977). It is an iterative approach for seeking the maximum likelihood estimate of parameters for probabilistic models that have missing data or latent variables. The application of EM algorithm includes the inference of the parameters of the Gaussian Mixture model etc. The EM algorithm involves two main steps, which are E-steps and M-steps. Suppose  $Z$  is the set of latent variable,  $X$  is the set of entire set of observed variables,  $\theta$  is parameter.  $t$  refers to the step during iteration,  $\log(P(X, Z|\theta))$  refers to the complete log-likelihood of data, and  $\log(P(X|\theta))$  refers to the incomplete log-likelihood of data without considering the hidden variables.

#### E-steps

By calculating the posterior distribution of the hidden variable given by the observed data and current parameter estimates, the purpose of this step is to compute the expectation of the latent variables by observed data, which is equivalent to calculate the expected value of the complete log-likelihood given the current parameter estimation and observed data. Mathematically, the E-step involves calculating the expectation of the complete data log-likelihood with respect to the conditional distribution of the hidden data given the observed

data and current parameter estimates:

$$Q(\theta, \theta^{(t-1)}) = E_{Z|X, \theta^{(t-1)}}[\log(P(X, Z|\theta))]. \quad (2.18)$$

Overall, the purpose of this step is to use the observed data to estimate and update the values of the missing data

### M-steps

The purpose of this step is to update the parameters that could maximize the expected complete data log-likelihood generated by the E-step, according to the current estimates

$$\theta^{(t)} = \operatorname{argmax}_{\theta} Q(\theta, \theta^{(t-1)}). \quad (2.19)$$

The algorithm runs until the difference between  $\theta^{(t)}$  and  $\theta^{(t-1)}$  is within an acceptable tolerance. The advantages and disadvantages of EM algorithm are as following: EM algorithm guarantees the increase of likelihood for each iteration according to [Dempster et al. \(1977\)](#), which enable EM algorithm becoming a greedy algorithm. However, it might suffer from slow convergence speed, sensitivity to initial parameter value, and the convergence to a local optima if there are multiple local optimas in the optimisation error surface.

## 2.4.2 Bayesian Expectation Maximization

The Bayesian EM algorithm incorporates the idea of EM algorithm and Bayesian inference for estimation of the probabilistic model when the data has missing or hidden values. As opposed to the traditional EM approach, the Bayesian EM approach incorporates prior knowledge of the parameter, for the estimation of the posterior mode  $p(\theta|\mathcal{D})$ , considering the prior distribution as  $p(\theta)$ , the Bayes rule can be written in the log scale

$$\ln p(\theta|\mathcal{D}) = \ln(p(\mathcal{D}|\theta)) + \ln(p(\theta)) - \ln(p(\mathcal{D})) \quad (2.20)$$

We can then further expand [Equation 2.20](#) to [Equation 2.21](#)

$$\ln p(\theta|\mathcal{D}) = Q(\theta, \theta^{(old)}) + \text{KL}(q||p(Z|X)) + \ln(p(\theta)) - \ln(p(\mathcal{D})) \quad (2.21)$$

, where  $\text{KL}(P||Q)$  is defined by [Equation 1.8](#)

### 2.4.3 Bayesian EM for Bayesian Lasso model

In order to deploy Bayesian EM algorithm to the Bayesian Lasso model for attaining the posterior mode, our purpose is to iteratively calculate

$$\theta_1^{(t+1)} = \underset{\theta_1}{\operatorname{argmax}}[E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\log p(y, \theta_1, \theta_2)]]. \quad (2.22)$$

#### E-step

Using the same notation as before, firstly, the complete log-likelihood can be written as [Equation 2.23](#)

$$\log(p(\theta_1, \theta_2, \tilde{y})) \propto -\frac{n+p}{2}\log(\sigma^2) - \frac{\|\tilde{y} - X\beta\|_2^2}{2\sigma^2} - \frac{1}{2\sigma^2}\beta^T E_{\theta_2|\tilde{y},\theta_1^{(t)}}[D_\tau]\beta - \frac{b}{\sigma^2} \quad (2.23)$$

given  $\theta_1 = (\beta, \sigma^2)$  as set of observed variables,  $\theta_2 = \tau^2 = (\tau_1^2, \dots, \tau_j^2)$  as set of latent variables,  $y$  is response variable.

$$E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\log p(y, \theta_1, \theta_2)] = -\frac{n}{2}\log(\sigma^{2(t)}) - \frac{\|y - X\beta^{(t)}\|_2^2}{2\sigma^{2(t)}} - E_{\theta_2|\tilde{y},\theta_1^{(t)}}\left[\sum_{j=1}^p \frac{\lambda^2 \beta_j^2}{2\sigma^2 \tau_j^2}\right] - (a+1)\log(\sigma^{2(t)}) - \frac{b}{\sigma^{2(t)}}. \quad (2.24)$$

Next, we need to take expectation of hidden variables:  $E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\sum_{j=1}^p \frac{\lambda^2 \beta_j^2}{2\sigma^2 \tau_j^2}]$ . After extracting constant with respect to  $\theta_2$ , required formulation is  $E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\frac{1}{\tau_j^2}]$ . Given the fact that  $\tau_j^2|\sigma^2, \tilde{y}, \beta \sim GIG(1, \frac{\beta_j^2 \lambda^2}{\sigma^2}, \frac{1}{2})$  and the special property of Generalized Inverse Gaussian distribution that if  $X \sim GIG(a, b, p)$ , then  $\frac{1}{X} \sim GIG(b, a, -p)$ , the distribution of  $\frac{1}{\tau_j^2}$  can be rearranged to  $\frac{1}{\tau_j^2}|\sigma^2, \tilde{y}, \beta \sim GIG(\frac{\beta_j^2 \lambda^2}{\sigma^2}, 1, -\frac{1}{2})$ . However, taking expectation with respect to Generalized Gaussian Distribution is still complicated and require advanced mathematical operation and functional properties such as modified Bessel function. Thus, we can continue converting the distribution into a Inverse Gaussian distribution family, which render  $\frac{1}{\tau_j^2}|\sigma^2, \theta_1, \tilde{y} \sim \text{InverseGaussian}(b^{-\frac{1}{2}}, 1)$ . Rewriting [Equation 2.24](#) can lead to the final conditional expectation form can be written as:

$$Q(\theta, \theta^{(t)}) = -(\frac{n}{2} + \frac{p}{2} + a + 1)\log(\sigma^2) - \frac{b}{\sigma^2} - \frac{\|y - X\beta\|_2^2}{2\sigma^2} - \frac{\lambda^2}{2\sigma^2} \sum_{j=1}^p (\beta_j^2 E[\frac{1}{\tau_j^2}]), \quad E[\frac{1}{\tau_j^2}] = \frac{\sigma^{(t)}}{|\beta_j^{(t)}|\lambda} \quad (2.25)$$

During the iteration, the  $E[\frac{1}{\tau_j^2}]$  will be iteratively updated according to the updated  $\beta^{(t)}$  and  $\sigma^{2(t)}$ .

## M-step

In order to maximize the expectation of complete log likelihood, taking derivative with respect to each target variable and set them to 0 respectively provides a closed-form solution for updating observed parameter repeatedly:

$$\frac{\partial Q}{\partial \beta} = -\frac{1}{2\sigma^2}(-X^T y + 2X^T X \beta) - \frac{\lambda^2}{2\sigma^2} X^T X \beta = 0. \quad (2.26)$$

Rearranging the Equation 2.26, the updated formula for  $\beta^{(t)}$  can be written as

$$\beta^{(t)} = (X^T X + \lambda^2 A)^{-1} X^T y, \text{ where } A = \text{diag}\left(\frac{\sigma^{(t-1)}}{|\beta^{(t-1)}|_\lambda}\right). \quad (2.27)$$

Similarly, set  $\frac{\partial Q}{\partial \sigma^2} = 0$ :

$$\frac{\partial Q}{\partial \sigma^2} = -\frac{(n + p + 2a + 2)}{2\sigma^2} + \frac{4b + 2\|y - X\beta\|_2^2 + \lambda^2(\beta^T A \beta)}{4\sigma^4} = 0 \quad (2.28)$$

Rearranging the Equation 2.28, the updated formula for  $\sigma^{2(t)}$  can be written as:

$$\sigma^{2(t)} = \frac{\|y - X\beta^{(t)}\|_2^2 + \lambda^2(\beta^{(t)T} A \beta^{(t)}) + 2b}{n + p + 2a + 2}. \quad (2.29)$$

---

### Algorithm 2 Bayesian Expectation Maximization algorithm for the Bayesian Lasso

---

- 1: Given initial value  $\theta_1^{(0)} = (\beta^{(0)}, \sigma^{2(0)})$ ,  $\theta_2^0 = \mathbf{1}_p$ ,  $t = 1$
  - 2: **while**  $\|\theta_1^{(t)} - \theta_1^{(t-1)}\|_2^2 < \epsilon$  **do**
  - 3:    $\beta^{(t)} = (X^T X + \lambda^2 A)^{-1} X^T y$ , where  $A = \text{diag}\left(\frac{\sigma^{(t-1)}}{|\beta^{(t-1)}|_\lambda}\right)$  ▷ Update  $\beta$
  - 4:    $\sigma^{2(t)} = \frac{\|y - X\beta^{(t)}\|_2^2 + \lambda^2(\beta^{(t)T} A \beta^{(t)}) + 2b}{n + p + 2a + 2}$  ▷ Update  $\sigma^2$
  - 5:    $A = \text{diag}\left(\frac{\sigma^{(t)}}{|\beta_j^{(t)}|_\lambda}\right)$  ▷ Estimate expectation of hidden variable  $E[\frac{1}{\tau_j^2}]$
  - 6:    $t \leftarrow t + 1$
  - 7: **return**  $\theta_1^{(t)}$
- 

After completing the iteration process of Bayesian Lasso, the posterior mode of Bayesian Lasso posterior distribution can be extracted from  $\beta$  generated by Bayesian Lasso algorithm, as a posterior model retaining variable selection nature.

## 2.5 Variational Inference

### 2.5.1 Introduction

One of the core challenges of statistician in a Bayesian setting is to approximate over-complex probability density function in a fast and efficient manner. VI serves as an effective alternative to MCMC algorithm especially for large datasets as mentioned in the previous chapter. By addressing an optimization-based system, it is possible to fit a proxy that accurately represents the posterior distribution. As the indispensable foundation of our proposed method, the purpose of this section is to provide detailed derivation and mathematical reasoning behind variational inference according to the detailed variational inference overview from [Blei et al. \(2017\)](#) and [Bishop \(2006\)](#).

### 2.5.2 KL divergence and Evidence Lower Bound(ELBO)

The purpose of the variational inference is to find a candidate approximation  $Q(\theta) \in Q$  after specifying a specific family of posterior distribution  $Q$  that minimize the  $KL$  divergence to the exact posterior distribution as shown in [Equation 1.7](#) for each subelement of parameter  $\theta$ . The complexity of finding optimal distribution relies heavily on the complexity of  $Q$ . Nevertheless, due to difficulty of computing marginal logarithm evidence  $p(\mathcal{D}) = \int_{\theta} P(\mathcal{D}, \theta) d\theta$ , as well as implicit dependency nature of  $p(\mathcal{D})$  to  $KL$  divergence as explained in [Equation 2.30](#), additional conversion is required for further processing this optimization system, which transform [Equation 1.7](#) to [Equation 1.9](#).

$$\begin{aligned} KL(q(\theta)||p(\theta|\mathcal{D})) &= \mathbb{E}_{q(\theta)}[\log(q(\theta))] - \mathbb{E}_{q(\theta)}[\log(p(\theta|\mathcal{D}))] \\ &= \mathbb{E}_{q(\theta)}[\log(q(\theta))] - \mathbb{E}_{q(\theta)}[\log(p(\theta, \mathcal{D}))] + \log[p(\mathcal{D})]. \end{aligned} \tag{2.30}$$

#### KL divergence

$KL$ -divergence defined by [Equation 2.30](#) is a distance metric for measuring the discrepancy of two probability distributions. This metric has several theoretical properties includes non-negativity, and asymmetric property of  $KL(q||p)$  and  $KL(p||q)$ .

## ELBO

The definition of the Evidence Lower bound is defined by [Equation 1.10](#), which is equivalent to the negative KL divergence despite of adding constant  $p(\mathcal{D})$  with respect to  $q(\theta)$ . Apart from the equivalence of optimization system, the explanation of why it is called "Evidence lower bound" can be shown in [Equation 2.31](#).

$$\log(p(\mathcal{D})) = \text{KL}(q(\theta)||p(\theta|\mathcal{D})) + \text{ELBO}(q(\theta)), \quad (2.31)$$

Given the fact that  $\text{KL}(.|.)$  is greater than 0, this explains log evidence is bounded below by *ELBO*, i.e:  $\log(p(\mathcal{D})) \geq \text{ELBO}(q(\theta))$ .

### 2.5.3 Mean-Field Variational Family

[Parisi and Shankar \(1988\)](#) proposed statistical mean-field theory. The mean-field variational family is the most common choice that is easier to optimize with a tractable solution form. The mean field variational family represents a group of probability distributions employed in variational inference. Its objective is to estimate intricate, infeasible distributions, such as the genuine posterior distribution within a Bayesian model, by using more tractable and simpler distributions. The generic member of mean field variational family is as described in [Equation 1.11](#), assuming mutual independence of each target parameter  $\theta_j$ , and the joint distribution is factorized as a product of individual distributions. Finally, no further assumption has been arranged to each individual distribution  $q(\theta_i)$ . The two dimensional visualization of mean-field variational family can be observed from [Figure 1.2](#), while the contour of mean-field variational family member forms a concentric contour line over the optimization surface.

### 2.5.4 Coordinate Ascent Variational Inference (CAVI)

Coordinate Ascent Variational Inference (CAVI) algorithm is the most frequently used optimization algorithm to find an optimum  $q^*(\theta)$  that maximizes the ELBO, which is preferred for its simplicity and computational efficiency. The main intuition behind coordinate ascent is to optimize the function with respect to one variable at a time while keeping the

other variables fixed. This is done iteratively until convergence is achieved. This is done iteratively until convergence is achieved. The convergence is achieved when either the difference between the current function value and the previous function value is less than a predefined threshold, or when the maximum number of iterations is reached. Similarly, the CAVI works by iteratively maximizing each component of  $q(\theta)$ :  $q_i(\theta_i)$ , while maintaining other factor of distribution unchanged,  $q_{-i}(\theta_{-i})$ , enforcing the final distribution can achieve a local optimum of the ELBO. Small changes regarding stopping criterion has also been proposed, where the stopping criterion has been transformed from difference of function value into difference of ELBO. Finally, we want to make note that MFVB belongs to a broader category of variational inference approaches, which can also serve as a purpose in the frequentist setting for maximum likelihood estimation. The CAVI algorithm is equivalent to MFVB algorithm in our setting.

## Derivation

Substituting the family of factorized target distribution  $\prod_i q_i$ , the ELBO can be rewritten as:

$$\begin{aligned}
\text{ELBO}(q(\theta)) &= \int q(\theta) \log\left(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)}\right) d\theta \\
&= \int \prod_i q_i(\theta_i) [\log[p(\mathcal{D}, \theta) - \sum_i \log(q_i(\theta_i))]] d\theta \\
&= \int q_j(\theta_j) \left[ \int \log p(\mathcal{D}, \theta) \prod_{i \neq j} q_i d\theta_j \right] d\theta_j - \int q_j(\theta_j) \log[q_j(\theta_j)] d\theta_j + \text{const} \\
&= \int q_j(\theta_j) \log[\tilde{p}(\mathcal{D}, \theta_j)] - \int q_j(\theta_j) \log[q_j(\theta_j)] d\theta_j
\end{aligned} \tag{2.32}$$

where  $\log[\tilde{p}(\mathcal{D}, \theta_j)] = \mathbb{E}_{i \neq j}[\log[p(\mathcal{D}, \theta)]] + \text{const}$ ,

$\mathbb{E}_{i \neq j}[\log[p(\mathcal{D}, \theta)]] = \int \log[p(\mathcal{D}, \theta)] \prod_{i \neq j} q_i(\theta_i) d\theta_i$  assuming optimizing the  $j$ th posterior parameter  $\theta_j$ . Identifying the 2.32 is negative KL-divergence between  $q_j(\theta_j)$  and  $\tilde{p}(\mathcal{D}, \theta_j)$ , the KL-divergence can be minimized proposed distribution is close enough to the target distribution. As a consequence, the optimal  $q^*(\theta_i)$  incidence of local maximum is achieved when  $q_j(\theta_j) = \tilde{p}(\mathcal{D}, \theta_j)$ . Thus, the general optimum distribution can be written in the following equation,

$$\log[q_j^*(\theta_j)] = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)] + \text{const} \tag{2.33}$$



Recovering from log scale and removing additional constant by normalization of  $q_j^*(\theta_j)$ , the optimum  $q_j^*(\theta_j)$  can be written as:

$$q_j^*(\theta_j) = \frac{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)]}{\int \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)] d\theta_j} \propto \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)] \quad (2.34)$$

The algorithm then iteratively replaces  $q_j(\theta_j)$  using the current estimation for all of other factors, and convergence criteria is break when differences of ELBOs changes less than a predefined tolerance. Theoretically, it is proved that the convergence of CAVI algorithm is guaranteed given the fact that the optimization problem is convex [Boyd and Vandenberghe \(2004\)](#). The following pseudo-algorithm demonstrates the entire procedure for CAVI.

---

**Algorithm 3** Coordinate Ascent Variational Inference (CAVI)

---

- 1: Input:  $p(\mathcal{D}, \theta)$ , data  $\mathcal{D}$ , Initialize Variational parameters for each  $q_j(\theta_j)$
  - 2: **while** ELBO has not converged **do**
  - 3:   **for**  $j=1, \dots, p$  **do**
  - 4:      $q_j(\theta_j) \propto \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)]$
  - 5:   Compute  $ELBO(q(\theta)) = \mathbb{E}[\log[p(\mathcal{D}, \theta)]] + \mathbb{E}[\log[q(\theta)]]$
  - 6: return  $q(\theta)$
- 

Similar to MFVB, CAVI has a similiar updating rule, while the only modification is that it converts the stopping criterion from using consecutive difference of ELBO into variation in the parameter  $\theta$  during two consecutive.

### 2.5.5 MFVB for Bayesian Lasso

We now propose MFVB algorithm for the Bayesian Lasso problem. Our goal is to attain Bayesian Lasso posterior approximation in a faster and efficient manner than that of gibbs sampler under the framework of VI. Note in this section, slight modification has been imposed to the definition of  $\tau_j$ .  $a_j \sim \text{Gamma}(1, 1/2)$  for  $1 \leq j \leq p$ , abandoning the dependence of  $\lambda$  to  $\tau$  to simplify the derivation and algorithm. The assumption of tractable distribution family  $Q$  can be written as the based on MFVB:

$$q(\theta) = q(\beta)q(\sigma^2) \prod_{i=1}^p q(a_j). \quad (2.35)$$

## Derivation

In addition, we would like to derive the optimal variational distribution for each parameter: ie.  $q_{\beta}^*(\beta)$ ,  $q_{\sigma^2}^*(\sigma^2)$  according to [Equation 2.34](#). We consider the parameters of  $a_j$  distribution are fixed, therefore, no hyper-priors are required.

The following derivation is based on the conditional posterior distributional form and expectation with respect to corresponding standard distribution  $\beta$  ??,  $q_{\beta}^*(\beta)$  can be derived by:

$$\begin{aligned} q_{\beta}^*(\beta) &\propto \exp[\mathbb{E}_{-\beta}[\log(p(\beta, y, \mathbf{a}, \sigma^2))]] \\ &\sim MVN(((X^T X + \lambda^2 \mathbb{E}_q(A))^{-1} X^T y, (X^T X + \lambda^2 \mathbb{E}_q(A))^{-1} \mathbb{E}_q(\sigma^2))) \end{aligned} \quad (2.36)$$

where  $A = \text{diag}(\mathbf{a})$ ,  $\mathbb{E}_{\sigma^2}(\sigma^2) = \frac{\tilde{b}}{\tilde{a}}$ ,  $\mathbb{E}_a(A) = \text{diag}(\frac{\sigma}{|\beta|\lambda})$ .

$$\mathbb{E}_q[\sigma^2] = \frac{\tilde{b}}{\tilde{a}}$$

Also,  $q_{\sigma^2}^*(\sigma^2)$  can be derived by the following equation. The following derivation are based on the known conditional posterior distributional form and expectation with respect to corresponding standard distribution.

$$\begin{aligned} q_{\sigma^2}^*(\sigma^2) &\propto \exp[\mathbb{E}_{-\sigma^2}[\log(p(\sigma^2, y, \mathbf{a}, \beta))]] \\ &\sim \text{InverseGamma}(\frac{n+p}{2}, \frac{1}{2} \mathbb{E}_q[\|y - X\beta\|^2] + \frac{1}{2} \sum_{j=1}^p \mathbb{E}_q(\beta^T A \beta)) \end{aligned} \quad (2.37)$$

$$\mathbb{E}_q[\|y - X\beta\|^2]$$

$$\mathbb{E}_q[\beta^T A \beta] = \frac{\sqrt{\frac{\tilde{b}}{\tilde{a}(\tilde{\mu}^T \tilde{\mu} + \text{diag}(\tilde{\Sigma}))}}}{\lambda}$$

The aforementioned algorithm 4 estimates the parameters of the posterior distributions of  $\beta$  and  $\sigma^2$  with descent speed.

---

**Algorithm 4** Mean Field Variational Bayes for Bayesian Lasso

---

- 1: Input: response variable  $y$ , data matrix  $X$ , Tuning parameter  $\lambda$ ,  $\tilde{\mu}^{(0)} = \mathbf{1}_p$ ,  $\tilde{\Sigma}^{(0)} = \text{diag}(1)$ ,  $\tilde{a}^{(0)} = 0.0001$ ,  $\tilde{b}^{(0)} = 0.0001$ ,  $d = \mathbf{1}$
  - 2: **while**  $\|\theta^{(t)} - \theta^{(t-1)}\|_2^2 < \epsilon$  **do**
  - 3:      $Q = X^T X + \lambda^2 A$ , where  $A = \text{diag}(d)$
  - 4:      $\tilde{\mu} = Q^{-1} X^T y$  ▷ Update posterior parameter for  $\beta$
  - 5:      $\tilde{\Sigma} = Q^{-1} \mathbb{E}_q[\sigma^2]$  ▷ Update posterior parameter for  $\beta$
  - 6:      $\tilde{a} = \frac{n+p}{2}$  ▷ Update posterior parameter for  $\sigma^2$
  - 7:      $\tilde{b} = \frac{E_q[\|y - X\beta\|^2] + \lambda^2 \mathbb{E}_q[\beta^T A \beta]}{2}$  ▷ Update posterior parameter for  $\sigma^2$
  - 8: return  $\tilde{a}, \tilde{b}, \tilde{\mu}, \tilde{\Sigma}$
-

# Chapter 3

## Methodology

### 3.1 Introduction

The main intuition behind the method is to adjust MFVB posterior parameter estimate result for posterior of regression coefficients  $\beta$ :  $\tilde{\mu}$  and  $\tilde{\Sigma}$ . Continuing utilizing mean field variational family  $q(\theta) = \prod_i q(\theta_i)$  assumption and Gaussian Approximation:  $q^*(\theta) \sim N(\tilde{\mu}, \tilde{\Sigma})$  to approximate the global Bayesian Lasso Posterior  $p(\theta|\mathcal{D})$ , with the correction of local parameter information that the marginal likelihood is approximately coincides with a lasso distribution if a Laplace prior is assigned. As a consequence, the goal is to seek for the mathematical expression of each of the local mean parameter  $\mu_j^*$ , local variance  $\Sigma_{jj}^*$  from a lasso distribution locally, a global mean  $\tilde{\mu}$  and global variance  $\tilde{\Sigma}$  of  $q^*(\theta)$  for Gaussian Approximation can be updated by iteratively to correct global parameter estimate by local parameter expression, for each  $\beta_j$ .

### 3.2 Basic Setting for the Bayesian Lasso Problem

Firstly, the Bayesian Lasso Posterior approximation can be restricted by mean field variational family:

$$p(\beta, \sigma^2|\mathcal{D}) \approx q(\beta, \sigma^2) = q(\beta)q(\sigma^2) \quad (3.1)$$

Under the same setting in Variational Inference in Mean-Field-Variational-Bayes by [Ormerod and Wand \(2010\)](#): the parameter of interest  $\theta$  can be divided up into two parts  $\theta_1$ :  $\beta_j$  current variable and  $\theta_2$ :  $\beta_{-j}$ , other variables. The marginal log likelihood of  $\theta_1$  can be divided up into ELBO part and KL divergence part by the following derivation:

$$\log(\mathcal{D}, \theta_1) = \mathbb{E}_{q(\theta_2|\theta_1)}[\log(\frac{p(\mathcal{D}, \theta_1, \theta_2)}{q(\theta_2|\theta_1)})] + KL(q(\theta_2|\theta_1), p(\theta_2|\mathcal{D}, \theta_1)) \quad (3.2)$$

Since the KL divergence is greater than 0, the marginal log likelihood of  $\theta_1$  and  $\mathcal{D}$  has a more tractable Evidence Lower Bound:

$$\log(\mathcal{D}, \theta_1) \geq \mathbb{E}_{q(\theta_2|\theta_1)}[\log(\frac{p(\mathcal{D}, \theta_1, \theta_2)}{q(\theta_2|\theta_1)})] \quad (3.3)$$

When  $q(\theta_2|\theta_1) = p(\theta_2|\mathcal{D}, \theta_1)$  then

$$\log(\mathcal{D}, \theta_1) = \mathbb{E}_{p(\theta_2|\mathcal{D}, \theta_1)}[\log p(\mathcal{D}, \theta_2, \theta_1)]$$

In Bayesian Lasso:  $\theta = (\beta, \sigma^2)$ , however, the update for  $\sigma$  will not be discussed while we will only discuss the an approach for updating  $\beta$ , assuming  $q(\beta) \sim N(\mu, \Sigma)$ . Thus, the conditional distribution  $q(\beta_{-j}|\beta_j)$  for any  $j_{th}$  variable can be derived by the fact that  $q(\beta_{-j}|\beta_j) \propto q(\beta)$ , resulting another multivariate normal distribution with dimension of  $p-1$  as shown in [Equation 3.4](#)

$$q(\beta_{-j}|\beta_j) = N_{p-1}(\mu_{-j} + \Sigma_{-j,j}\Sigma_{j,j}^{-1}(\beta_j - \mu_j), \Sigma_{-j,j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}) \quad (3.4)$$

With the mean field restriction in [Equation 3.1](#), the result from [Equation 3.4](#), the fact that the products of log density  $Y|\beta, \sigma^2$ ,  $\beta|\sigma^2, \lambda$  and a Laplacian prior for  $\beta|\sigma^2, \lambda$ , the estimated marginal log likelihood for each  $\beta_j$  after taking expectation with respect to  $q(\theta)$ :

$$\begin{aligned} \log p(\mathcal{D}, \beta_j) &= \mathbb{E}_{\beta_{-j}, \sigma^2|\mathcal{D}, \beta_j} \log(p(\beta_j|\mathcal{D}, \beta_{-j}, \sigma^2)) \\ &\approx \mathbb{E}_{q(\beta_{-j}|\beta_j)q(\sigma^2)} \log(p(\beta_j|\mathcal{D}, \beta_{-j}, \sigma^2)) \\ &\propto \mathbb{E}_{q(\beta_{-j}|\beta_j)q(\sigma^2)} \left[ -\frac{\|X_j\|_2^2}{2\sigma^2}\beta_j^2 + \frac{X_j^T(y - X_{-j}\beta_{-j})}{\sigma^2}\beta_j - \frac{\lambda}{\sigma}|\beta_j| \right] \\ &= \frac{\tilde{a}}{\tilde{b}}(y - X_{-j}s)\beta_j - \frac{\tilde{a}}{2\tilde{b}}(X_j^T X_j + X_j^T X_{-j}t)\beta_j^2 - \frac{\lambda\Gamma(\tilde{a} + 1/2)}{\Gamma(\tilde{a})\sqrt{\tilde{b}}}|\beta_j|. \end{aligned} \quad (3.5)$$

where  $s = \mu_{-j} - \Sigma_{-j,j}\Sigma_{j,j}^{-1}\mu_j$  and  $t = \Sigma_{-j,j}\Sigma_{j,j}^{-1}$ ,  $\tilde{a}$  and  $\tilde{b}$  are posterior parameters for  $\sigma^2$ ,  $\mu, \Sigma$  are posterior parameters for  $\beta$ .

### 3.3 Lasso distribution

Before continuing presenting the methodology, the introduction of lasso distribution is essential for local approximation correction. [Equation 3.5](#) can be matched to an univariate lasso distribution as a local approximation of Bayesian Lasso posterior. In addition, the joint likelihood of a pair of variables can also be matched by a bivariate lasso distribution.

As shown in [Figure 3.1](#)

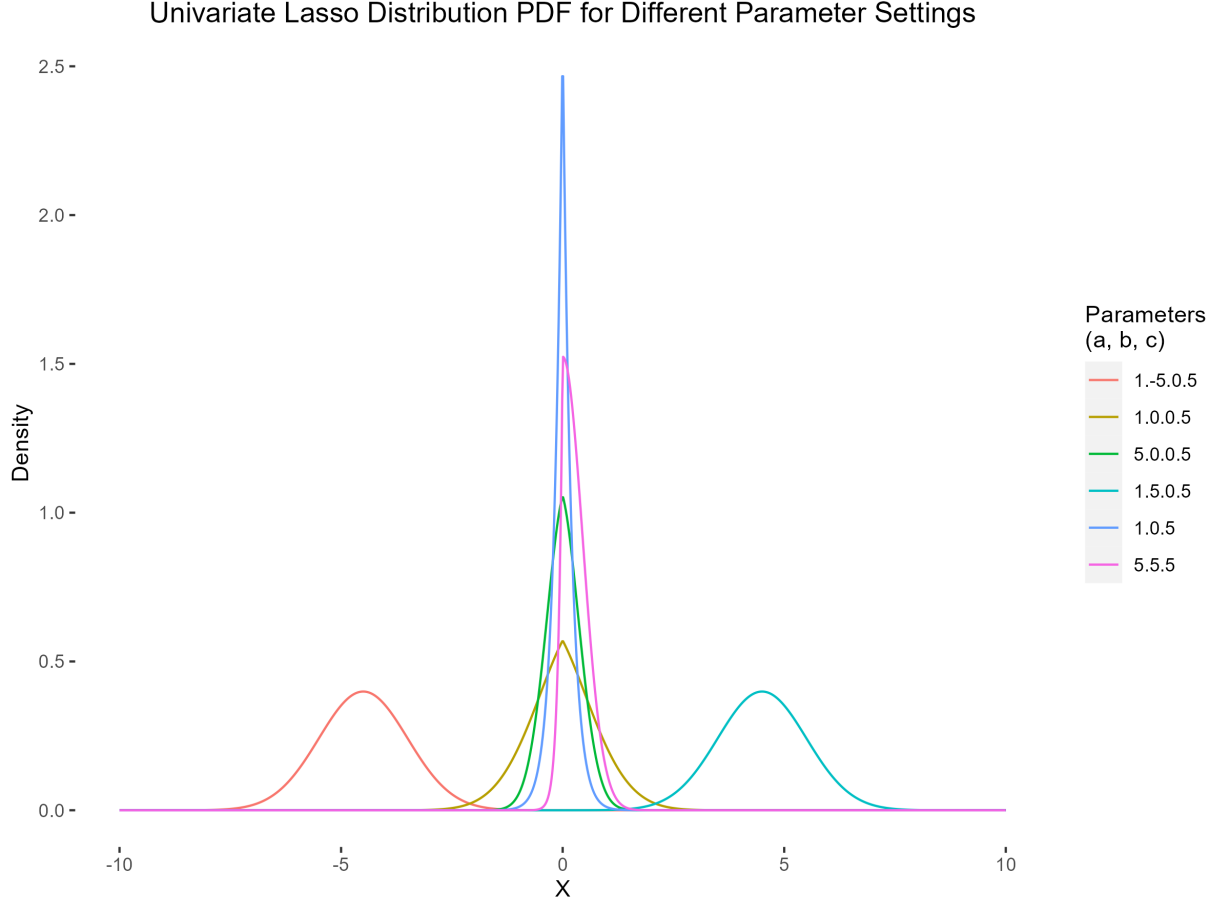


Figure 3.1: Visualization of Lasso Distribution PDF for different parameter setting

### 3.3.1 Univariate Lasso Distribution

If  $x \sim \text{Lasso}(a, b, c)$ , then the probability density function can be written as:

$$p(x, a, b, c) = Z^{-1} \exp\left(-\frac{1}{2}ax^2 + bx - c|x|\right) \quad (3.6)$$

where  $a > 0, b \in \mathbb{R}, c > 0$ ,  $Z$  is normalizing constant. Certain property of a probability distribution can also be formed to demonstrate the effectiveness, in our algorithm normalizing constant  $Z$ , expectation  $\mathbb{E}(x)$ , second moment  $\mathbb{E}(x^2)$  and variance  $\mathbb{V}(x)$  are necessary.

## Basic Property

### Derivation of normalizing constant

The normalizing constant  $Z$  can be written as a function of  $a$ ,  $b$ ,  $c$ .

$$\begin{aligned}
Z(a, b, c) &= \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2}ax^2 + bx - c|x| \right] dx \\
&= \int_0^{\infty} \exp \left[ -\frac{1}{2}ax^2 + (b - c)x \right] dx + \int_{-\infty}^0 \exp \left[ -\frac{1}{2}ax^2 + (b + c)x \right] dx \\
&= \int_0^{\infty} \exp \left[ -\frac{1}{2}ax^2 + (b - c)x \right] dx + \int_0^{\infty} \exp \left[ -\frac{1}{2}ay^2 - (b + c)y \right] dy \\
&= \int_0^{\infty} \exp \left[ -\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{\mu_1^2}{2\sigma^2} \right] dx + \int_0^{\infty} \exp \left[ -\frac{(x-\mu_2)^2}{2\sigma^2} + \frac{\mu_2^2}{2\sigma^2} \right] dy \\
&= \sqrt{2\pi\sigma^2} \left[ \exp \left\{ \frac{\mu_1^2}{2\sigma^2} \right\} \int_0^{\infty} \phi(x; \mu_1, \sigma^2) dx + \exp \left\{ \frac{\mu_2^2}{2\sigma^2} \right\} \int_0^{\infty} \phi(y; \mu_2, \sigma^2) dy \right] \\
&= \sqrt{2\pi\sigma^2} \left[ \exp \left\{ \frac{\mu_1^2}{2\sigma^2} \right\} \{1 - \Phi(-\mu_1/\sigma)\} + \exp \left\{ \frac{\mu_2^2}{2\sigma^2} \right\} \{1 - \Phi(-\mu_2/\sigma)\} \right] \\
&= \sqrt{2\pi\sigma^2} \left[ \exp \left( \frac{\mu_1^2}{2\sigma^2} \right) \Phi \left( \frac{\mu_1}{\sigma} \right) + \exp \left( \frac{\mu_2^2}{2\sigma^2} \right) \Phi \left( \frac{\mu_2}{\sigma} \right) \right] \\
&= \sigma \left[ \frac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} + \frac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)} \right]
\end{aligned}$$

### Derivation of Moments

Note, the expectation is the first moment, and variance of lasso distribution can be computed by the property  $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ .

$$\begin{aligned}
E(x^r) &= Z^{-1} \int_{-\infty}^{\infty} x^r \exp \left[ -\frac{1}{2}ax^2 + bx - c|x| \right] dx \\
&= Z^{-1} \int_0^{\infty} x^r \exp \left[ -\frac{1}{2}ax^2 + (b - c)x \right] dx + \int_{-\infty}^0 x^r \exp \left[ -\frac{1}{2}ax^2 + (b + c)x \right] dx \\
&= Z^{-1} \int_0^{\infty} x^r \exp \left[ -\frac{1}{2}ax^2 + (b - c)x \right] dx + (-1)^r \int_0^{\infty} y^r \exp \left[ -\frac{1}{2}ay^2 - (b + c)y \right] dy \\
&= Z^{-1} \sqrt{2\pi\sigma^2} \exp \left( \frac{\mu_1^2}{2\sigma^2} \right) \int_0^{\infty} x^r \phi(x; \mu_1, \sigma^2) dx \\
&\quad + (-1)^r \sqrt{2\pi\sigma^2} \exp \left( \frac{\mu_2^2}{2\sigma^2} \right) \int_0^{\infty} y^r \phi(y; \mu_2, \sigma^2) dy \\
&= \frac{\sigma}{Z} \left[ \frac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} \frac{\int_0^{\infty} x^r \phi(x; \mu_1, \sigma^2) dx}{\Phi(\mu_1/\sigma)} + (-1)^r \frac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)} \frac{\int_0^{\infty} y^r \phi(y; \mu_2, \sigma^2) dy}{\Phi(\mu_2/\sigma)} \right] \\
&= \frac{\sigma}{Z} \left[ \frac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} \mathbb{E}(A^r) + (-1)^r \frac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)} \mathbb{E}(B^r) \right]
\end{aligned} \tag{3.7}$$

where  $A \sim TN_+(\mu_1, \sigma^2)$ ,  $B \sim TN_+(\mu_2, \sigma^2)$  and  $TN_+$  is denotes the positively truncated normal distribution;  $\mu_1 = (b - c)/a$ ,  $\mu_2 = -(c + b)/a$  and  $\sigma^2 = 1/a$ . Note that

$$\mathbb{E}(A) = \mu_1 + \frac{\sigma \phi(\mu_1/\sigma)}{\Phi(\mu_1/\sigma)} = \mu_1 + \sigma \zeta_1(\mu_1/\sigma)$$

and

$$\mathbb{V}(A) = \sigma^2 [1 + \zeta_2(\mu_1/\sigma)]$$

where  $\zeta_k(x) = d^k \log \Phi(x)/dx^k$ ,  $\zeta_1(t) = \phi(t)/\Phi(t)$ ,  $\zeta_2(t) = -t \zeta_1(t) - \zeta_1(t)^2$ . Here  $\zeta_1(x)$  is the inverse Mills ratio which too needs to be treated with care. Hence,

$$\mathbb{E}(A^2) = \mathbb{V}(A) + \mathbb{E}(A)^2 = \sigma^2 [1 + \zeta_2(\mu_1/\sigma)] + [\mu_1 + \sigma \zeta_1(\mu_1/\sigma)]^2$$

We now have sufficient information to calculate the moments of the Lasso distribution. We also have sufficient information to implement a VB approximation.

### 3.3.2 Bivariate Lasso Distribution

If  $\mathbf{x} \sim \text{Bilasso}(A, b, c)$  with then it has density given by

$$p(\mathbf{x}) = Z^{-1} \exp\left(-\frac{1}{2} \mathbf{x}^T A \mathbf{x} + b^T \mathbf{x} - c \|\mathbf{x}\|_1\right) \quad (3.8)$$

where  $A \in S_d^+$ : positive definite matrix with dimension  $d$ ,  $b \in \mathbb{R}^2$ ,  $c > 0$



## Derivation of Normalizing Constant

$$\begin{aligned}
Z(a, b, c) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} x^T A x + \mathbf{b}^T x - c \mathbf{1}^T |x|_1 \right] d\mathbf{x} \\
&= \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} x^T A x + (\mathbf{b}^T - c \mathbf{1}^T) x \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_{-\infty}^0 \exp \left[ -\frac{1}{2} x^T A x + (\mathbf{b}^T - c[1, -1]^T) x \right] d\mathbf{x} \\
&\quad + \int_{-\infty}^0 \int_0^{\infty} \exp \left[ -\frac{1}{2} x^T A x + (\mathbf{b}^T - c[-1, 1]^T) x \right] d\mathbf{x} \\
&\quad + \int_{-\infty}^0 \int_{-\infty}^0 \exp \left[ -\frac{1}{2} x^T A x + (\mathbf{b}^T + c \mathbf{1}^T) x \right] d\mathbf{x} \\
&= \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} x^T A x + (\mathbf{b}^T - c \mathbf{1}^T) x \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} x^T A^* x + (b_1 - c, -b_2 - c)^T x \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} x^T A^* x + (-b_1 - c, b_2 - c)^T x \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} x^T A x - (\mathbf{b}^T + c \mathbf{1}^T) x \right] d\mathbf{x} \\
&= \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \frac{(A\mu_1)^T \Sigma_1 (A\mu_1)}{2} \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \frac{(A^* \mu_2)^T \Sigma_2 (A^* \mu_2)}{2} \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} (x - \mu_3)^T \Sigma_2^{-1} (x - \mu_3) + \frac{(A^* \mu_3)^T \Sigma_2 (A^* \mu_3)}{2} \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} \exp \left[ -\frac{1}{2} (x - \mu_4)^T \Sigma_1^{-1} (x - \mu_4) + \frac{(A\mu_4)^T \Sigma_1 (A\mu_4)}{2} \right] d\mathbf{x} \\
&= 2\pi |\Sigma_1|^{\frac{1}{2}} \left[ \exp \left[ \frac{(A\mu_1)^T \Sigma_1 (A\mu_1)}{2} \right] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_1, \Sigma_1) d\mathbf{x} \right. \\
&\quad \left. + \exp \left[ \frac{(A\mu_4)^T \Sigma_1 (A\mu_4)}{2} \right] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_4, \Sigma_1) d\mathbf{x} \right] \\
&\quad + 2\pi |\Sigma_2|^{\frac{1}{2}} \left( \exp \left[ \frac{(A^* \mu_2)^T \Sigma_2 (A^* \mu_2)}{2} \right] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_2, \Sigma_2) d\mathbf{x} \right. \\
&\quad \left. + \exp \left[ \frac{(A^* \mu_3)^T \Sigma_2 (A^* \mu_3)}{2} \right] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_3, \Sigma_2) d\mathbf{x} \right) \\
&= |\Sigma_1| \left( \frac{\int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_1, \Sigma_1) d\mathbf{x}}{\phi_2(A\mu_1, \Sigma_1^{-1})} + \frac{\int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_4, \Sigma_1) d\mathbf{x}}{\phi_2(A\mu_4, \Sigma_1^{-1})} \right) \\
&\quad + |\Sigma_2| \left( \frac{\int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_2, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_2, \Sigma_2^{-1})} + \frac{\int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_3, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_3, \Sigma_2^{-1})} \right)
\end{aligned}$$

where  $\mu_1 = A^{-1}(b - c \mathbf{1})^T$ ,  $\mu_2 = A^{*-1}(b_1 - c, -b_2 - c)^T$ ,  $\mu_3 = A^{*-1}(-b_1 - c, b_2 - c)^T$ ,  $\mu_4 = A^{-1}(-b - c \mathbf{1}^T)^T$  and  $\Sigma_1 = A^{-1}$ ,  $\Sigma_2 = A^{*-1}$   $A^* = A \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ .

## Derivation of Expectation

Follow similar step as before

$$\begin{aligned}
E[X] &= Z^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \otimes \exp \left[ -\frac{1}{2} x^T A x + \mathbf{b}^T x - c \mathbf{1}^T \|x\|_1 \right] d\mathbf{x} \\
&= Z^{-1} \int_0^{\infty} \int_0^{\infty} x \otimes \exp \left[ -\frac{1}{2} x^T A x + (\mathbf{b}^T - c \mathbf{1}^T) x \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} [1, -1]^T \otimes x \otimes \exp \left[ -\frac{1}{2} x^T A^* \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} x + (b_1 - c, -b_2 - c)^T x \right] d\mathbf{x} \\
&\quad + \int_0^{\infty} \int_0^{\infty} [-1, 1]^T \otimes x \otimes \exp \left[ -\frac{1}{2} x^T A^* \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} x + (-b_1 - c, b_2 - c)^T x \right] d\mathbf{x} \\
&\quad - \int_0^{\infty} \int_0^{\infty} x \otimes \exp \left[ -\frac{1}{2} x^T A x + (\mathbf{b}^T + c \mathbf{1}^T) x \right] d\mathbf{x} \\
&= Z^{-1} \left[ |\Sigma_1| \left( \frac{\int_0^{\infty} \int_0^{\infty} x \otimes \phi_2(x; \mu_1, \Sigma) d\mathbf{x}}{\phi_2(A\mu_1, \Sigma_1^{-1})} - \frac{\int_0^{\infty} \int_0^{\infty} x \otimes \phi_2(x; \mu_4, \Sigma) d\mathbf{x}}{\phi_2(A\mu_4, \Sigma_1^{-1})} \right) \right. \\
&\quad \left. + |\Sigma_2| \left( [1, -1]^T \frac{\int_0^{\infty} \int_0^{\infty} x \otimes \phi_2(x; \mu_2, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_2, \Sigma_2^{-1})} + [-1, 1]^T \frac{\int_0^{\infty} \int_0^{\infty} x \otimes \phi_2(x; \mu_3, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_3, \Sigma_2^{-1})} \right) \right] \\
&= Z^{-1} \left[ |\Sigma_1| \left( \frac{E[\mathbf{A}] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_1, \Sigma_1) d\mathbf{x}}{\phi_2(A\mu_1, \Sigma_1^{-1})} - \frac{E[D] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_4, \Sigma_1) d\mathbf{x}}{\phi_2(A\mu_4, \Sigma_1^{-1})} \right) \right. \\
&\quad \left. + |\Sigma_2| \left( [1, -1]^T \frac{E[B] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_2, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_2, \Sigma_2^{-1})} + [-1, 1]^T \frac{E[C] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_3, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_3, \Sigma_2^{-1})} \right) \right]
\end{aligned}$$

where  $\mu_1 = A^{-1}(b - c\mathbf{1})^T$ ,  $\mu_2 = A^{*-1}(b_1 - c, -b_2 - c)^T$ ,  $\mu_3 = A^{*-1}(-b_1 - c, b_2 - c)^T$ ,  $\mu_4 = A^{-1}(-b - c\mathbf{1}^T)^T$  and  $\Sigma_1 = A^{-1}$ ,  $\Sigma_2 = A^{*-1}$   $A^* = A \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ .  $\mathbf{A} \sim MTN_+(\mu_1, \Sigma_1)$ ,  $B \sim MTN_+(\mu_2, \Sigma_2)$ ,  $C \sim MTN_+(\mu_3, \Sigma_2)$ ,  $D \sim MTN_+(\mu_4, \Sigma_1)$  is denotes the multivariate positively truncated normal distribution.

## Derivation of Covariance Matrix

Follow similar steps as before

$$Cov(X) = E[XX^T] - E[X]E[X]^T$$

$$\begin{aligned}
E[XX^T] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xx^T \otimes \exp \left[ -\frac{1}{2}x^T Ax + \mathbf{b}^T x - c\mathbf{1}^T ||x||_1 \right] d\mathbf{x} \\
&= Z^{-1} 2\pi |\Sigma|^{\frac{1}{2}} \left[ \exp \left[ \frac{(A\mu_1)^T \Sigma (A\mu_1)}{2} \right] \int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_1, \Sigma_1) d\mathbf{x} \right. \\
&\quad + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \exp \left[ \frac{(A^* \mu_2)^T \Sigma (A^* \mu_2)}{2} \right] \int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_2, \Sigma_2) d\mathbf{x} \\
&\quad + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \exp \left[ \frac{(A^* \mu_3)^T \Sigma (A^* \mu_3)}{2} \right] \int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_3, \Sigma_2) d\mathbf{x} \\
&\quad \left. - \exp \left[ \frac{(A\mu_4)^T \Sigma (A\mu_4)}{2} \right] \int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_4, \Sigma_1) d\mathbf{x} \right] \\
&= Z^{-1} |\Sigma| \left[ \frac{\int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_1, \Sigma) d\mathbf{x}}{\phi_2(A\mu_1, \Sigma)} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \frac{\int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_2, \Sigma) d\mathbf{x}}{\phi_2(A\mu_2, \Sigma)} \right. \\
&\quad \left. + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \frac{\int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_3, \Sigma) d\mathbf{x}}{\phi_2(A\mu_3, \Sigma)} - \frac{\int_0^{\infty} \int_0^{\infty} xx^T \otimes \phi_2(x; \mu_4, \Sigma) d\mathbf{x}}{\phi_2(A\mu_4, \Sigma)} \right] \\
&= Z^{-1} [|\Sigma_1| \left( \frac{E[\mathbf{A}\mathbf{A}^T] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_1, \Sigma_1) d\mathbf{x}}{\phi_2(A\mu_1, \Sigma_1^{-1})} + \frac{E[DD^T] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_4, \Sigma_1) d\mathbf{x}}{\phi_2(A\mu_4, \Sigma_1^{-1})} \right) \\
&\quad + |\Sigma_2| \left( \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \frac{E[BB^T] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_2, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_2, \Sigma_2^{-1})} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \frac{E[CC^T] \int_0^{\infty} \int_0^{\infty} \phi_2(x; \mu_3, \Sigma_2) d\mathbf{x}}{\phi_2(A^* \mu_3, \Sigma_2^{-1})} \right)]
\end{aligned}$$

where  $\mu_1 = A^{-1}(b - c\mathbf{1})^T$ ,  $\mu_2 = A^{*-1}(b_1 - c, -b_2 - c)^T$ ,  $\mu_3 = A^{*-1}(-b_1 - c, b_2 - c)^T$ ,  $\mu_4 = A^{-1}(-b - c\mathbf{1}^T)^T$  and  $\Sigma_1 = A^{-1}$ ,  $\Sigma_2 = A^{*-1}$ ,  $A^* = A \otimes \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ .  $\mathbf{A} \sim MTN_+(\mu_1, \Sigma_1)$ ,  $B \sim MTN_+(\mu_2, \Sigma_2)$ ,  $C \sim MTN_+(\mu_3, \Sigma_2)$ ,  $D \sim MTN_+(\mu_4, \Sigma_1)$  is denotes the multivariate positively truncated normal distribution.

$$E[AA^T] = Cov(A) - E[A]E[A]^T$$

## Derivation of marginal distribution

$$\begin{aligned}
f(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \\
&= Z^{-1} \exp(-\frac{1}{2}x^T A x + b^T x - c||x||_1) dx_2 \\
&= Z^{-1} \exp(-0.5a_{11}x_1^2 + b_1x_1 - c|x_1|) \\
&\quad \int_{-\infty}^{\infty} \exp(-\frac{1}{2}[(a_{12} + a_{21})x_1x_2 + a_{22}x_2^2] + b_2x_2 - c|x_2|) dx_2 \\
&= k \int_{-\infty}^{\infty} \exp[-(0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{22}x_2^2 + b_2x_2 - c|x_2|)] dx_2 \\
&= k [\int_0^{\infty} \exp[-0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{22}x_2^2 + (b_2 - c)x_2] dx_2 \\
&\quad + \int_{-\infty}^0 \exp[-0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{22}x_2^2 + (b_2 + c)x_2] dx_2 \\
&= k [\int_0^{\infty} \exp[-0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{22}x_2^2 + (b_2 - c)x_2] dx_2 \\
&\quad + \int_0^{\infty} \exp[0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{22}x_2^2 - (b_2 + c)x_2] dx_2] \\
&= k [\int_0^{\infty} \exp[-\frac{(x_2 - \mu_1)^2}{2\sigma^2} + \frac{\mu_1^2}{2\sigma^2}] dx_2] + \int_0^{\infty} \exp[-\frac{(x_2 - \mu_2)^2}{2\sigma^2} + \frac{\mu_2^2}{2\sigma^2}] dx_2] \\
&= k\sigma [\frac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} + \frac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)}]
\end{aligned}$$

where  $\mu_1 = (-\frac{a_{12}+a_{21}}{2a_{22}}x_1 + \frac{b_2-c}{a_{22}})$ ,  $\mu_2 = (\frac{a_{12}+a_{21}}{2a_{22}}x_1 - \frac{b_2+c}{a_{22}})$ ,  $\sigma^2 = 1/a_{22}$ ,  $k = Z^{-1} \exp(-0.5a_{11}x_1^2 + b_1x_1 - c|x_1|)$

$$\begin{aligned}
f(x_2) &= \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 \\
&= Z^{-1} \exp(-\frac{1}{2}x^T A x + b^T x - c||x||_1) dx_1 \\
&= Z^{-1} \exp(-0.5a_{22}x_2^2 + b_2x_2 - c|x_2|) \\
&\quad \int_{-\infty}^{\infty} \exp(-\frac{1}{2}[a_{12}a_{21}x_1x_2 + a_{11}x_1^2] + b_1x_1 - c|x_1|) dx_1 \\
&= k \int_{-\infty}^{\infty} \exp[-0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{11}x_1^2 + b_1x_1 - c|x_1|] dx_1 \\
&= k [\int_0^{\infty} \exp[-0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{11}x_1^2 + (b_1 - c)x_1] dx_1 \\
&\quad + \int_{-\infty}^0 \exp[-0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{11}x_1^2 + (b_1 + c)x_1] dx_1] \\
&= k [\int_0^{\infty} \exp[-0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{11}x_1^2 + (b_1 - c)x_1] dx_1 \\
&\quad + \int_0^{\infty} \exp[0.5(a_{12} + a_{21})x_1x_2 - 0.5a_{11}x_1^2 - (b_1 + c)x_1] dx_1] \\
&= k [\int_0^{\infty} \exp[-\frac{(x_1 - \mu_1)^2}{2\sigma^2} + \frac{\mu_1^2}{2\sigma^2}] dx_1] + \int_0^{\infty} \exp[-\frac{(x_1 - \mu_2)^2}{2\sigma^2} + \frac{\mu_2^2}{2\sigma^2}] dx_1] \\
&= k\sigma [\frac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} + \frac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)}]
\end{aligned}$$

where  $\mu_1 = (-\frac{a_{12}+a_{21}}{2a_{11}}x_2 + \frac{b_1-c}{a_{11}})$ ,  $\mu_2 = (\frac{a_{12}+a_{21}}{2a_{11}}x_2 - \frac{b_1+c}{a_{11}})$ ,  $\sigma^2 = 1/a_{11}$ ,  $k = Z^{-1} \exp(-0.5a_{22}x_2^2 + b_2x_2 - c|x_2|)$

## 3.4 Local-Global Algorithm

### 3.4.1 Univariate local global algorithm

Continuing from Equation 3.5, it is evident to observe that  $p(\beta_j|\mathcal{D}) \propto p(\beta_j, \mathcal{D}) \sim \text{Lasso}(\frac{\tilde{a}}{b}(y - X_{-j}s), \frac{\tilde{a}}{2b}(X_j^T X_j + X_j^T X_{-j}t), \frac{\lambda\Gamma(\tilde{a}+1/2)}{\Gamma(\tilde{a})\sqrt{b}})$ .

The local approximation of mean  $\mu_j^*$  and variance  $\Sigma_{jj}^*$  can be obtained by the expression Equation 3.7 with given Lasso parameter  $a, b, c$ . A marginal normal approximation to the conditional distribution  $p(\theta_j|\mathcal{D})$ :  $q^*(\theta_j) \approx N(\mu_j^*, \Sigma_{jj}^*)$ . The optimal distribution can be updated via Equation 3.9

$$q^*(\theta) = q(\theta_{-j}|\theta_j)\phi(\theta_j; \mu_j^*, \Sigma_{jj}^*) \quad (3.9)$$

Since, both  $q(\theta_{-j}|\theta_j)$  and  $q(\theta_j)$  are Normal distribution, the joint distribution will also be a normal distribution as well. Additionally, the marginal mean and variance of the joint distribution for  $\theta_j$  will be  $\mu_j^*$  and  $\Sigma_{jj}^*$ . The derivation for  $\tilde{\mu}$  has been shown below using the property of the conditional expectation trick. It is known that  $q(\theta_2|\theta_1)$  and  $q(\theta_1)$  are Gaussian. Hence, their joint distribution will also be Gaussian. The marginal mean and variance of the joint distribution for  $\theta_1$  will be  $\mu_1^*$  and  $\Sigma_{11}^*$ . Suppose the mean and covariance of the joint distribution are  $\tilde{\mu}$  and  $\tilde{\Sigma}$  respectively. Then  $\tilde{\mu}_1 = \mu_1^*$ , and  $\tilde{\Sigma}_{11} = \Sigma_{11}^*$ . The rest of the derivation is to determine  $\tilde{\mu}_2$ ,  $\tilde{\Sigma}_{22}$  and  $\tilde{\Sigma}_{12}$ .

We have

$$\begin{aligned} \mathbb{E}[\theta_2] &= \mathbb{E}[\mathbb{E}(\theta_2 | \theta_1)] \\ &= \mathbb{E}[\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\theta_1 - \mu_1)] \\ &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mu_1^* - \mu_1). \end{aligned}$$

Similarly,

$$\begin{aligned} \text{Cov}(\theta_2) &= \mathbb{E}[\text{Cov}(\theta_2 | \theta_1)] + \text{Cov}[\mathbb{E}(\theta_2 | \theta_1)] \\ &= \mathbb{E}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}) + \text{Cov}[\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\theta_1 - \mu_1)] \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{21}\Sigma_{11}^{-1}\text{Cov}(\theta_1)\Sigma_{11}^{-1}\Sigma_{12} \\ &= \Sigma_{22} + \Sigma_{21}(\Sigma_{11}^{-1}\Sigma_{11}^*\Sigma_{11}^{-1} - \Sigma_{11}^{-1})\Sigma_{12}. \end{aligned}$$

Lastly,

$$\begin{aligned}
\text{Cov}(\theta_1, \theta_2) &= \mathbb{E}[(\theta_1 - \mathbb{E}(\theta_1))(\theta_2 - \mathbb{E}(\theta_2))^T] \\
&= \mathbb{E}[\{(\theta_1 - \mathbb{E}(\theta_1))(\theta_2 - \mathbb{E}(\theta_2))^T \mid \theta_1\}] \\
&= \mathbb{E}\left[(\theta_1 - \mu_1^*) (\Sigma_{21} \Sigma_{11}^{-1} (\theta_1 - \mu_1^*))^T\right] \\
&= \mathbb{E}\left[(\theta_1 - \mu_1^*) (\theta_1 - \mu_1^*)^T\right] \Sigma_{11}^{-1} \Sigma_{12} \\
&= \Sigma_{11}^* \Sigma_{11}^{-1} \Sigma_{12}.
\end{aligned}$$

Hence,  $q^*(\theta) = N(\tilde{\mu}, \tilde{\Sigma})$  where

$$\tilde{\mu} = \begin{bmatrix} \mu_1^* \\ \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mu_1^* - \mu_1) \end{bmatrix}$$

The derivation for  $\tilde{\Sigma}$  has been shown below using the property of the conditional covariance:

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma_{11}^* & \Sigma_{11}^* \Sigma_{11}^{-1} \Sigma_{12} \\ \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{11}^* & \Sigma_{22} + \Sigma_{21} \Sigma_{11}^{-1} (\Sigma_{11}^* - \Sigma_{11}) \Sigma_{11}^{-1} \Sigma_{12} \end{bmatrix}.$$

Hence,  $q^*(\theta) = N(\tilde{\mu}, \tilde{\Sigma})$  at the time when updating variable  $\beta_j$  is:

$$\tilde{\mu} = \begin{bmatrix} \mu_j^* \\ \mu_{-j} + \Sigma_{-j,j} \Sigma_{jj}^{-1} (\mu_j^* - \mu_j) \end{bmatrix} \tag{3.10}$$

and

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma_{jj}^* & \Sigma_{jj}^* \Sigma_{jj}^{-1} \Sigma_{j-j} \\ \Sigma_{-jj} \Sigma_{jj}^{-1} \Sigma_{jj}^* & \Sigma_{-j-j} + \Sigma_{-jj} \Sigma_{jj}^{-1} (\Sigma_{jj}^* - \Sigma_{jj}) \Sigma_{jj}^{-1} \Sigma_{j-j} \end{bmatrix}. \tag{3.11}$$

---

**Algorithm 5** Univariate-Local-Global-Algorithm

---

```
1: Input: data  $X$ , normalized response variable  $y$ , parameter from MFVB  $(\tilde{a}, \tilde{b}, \tilde{\mu}, \tilde{\Sigma})$ ,  
   Penalizing parameter:  $\lambda$   
2: while  $\tilde{\mu}$  is changing less than  $\epsilon$  do  
3:   for  $j = 1$  to  $p$  do  
4:      $a = \frac{\tilde{a}}{\tilde{b}}(X^T X)_{j,j} + (X^T X)_{j,-j} \tilde{\Sigma}_{-j,j} \tilde{\Sigma}_{j,j}^{-1}$  ▷ Obtain Lasso parameter  
5:      $b = \frac{\tilde{a}}{\tilde{b}} X_j (y - X_{-j}(\tilde{\mu}_{-j} - \tilde{\Sigma}_{-j,j} \tilde{\Sigma}_{j,j}^{-1} \tilde{\mu}_j))$  ▷ Obtain Lasso parameter  
6:      $c = \lambda(\exp(\Gamma(\tilde{a} + 0.5) - \Gamma(\tilde{a}) - 0.5 \log(\tilde{b})))$  ▷ Obtain Lasso parameter  
7:      $\tilde{\mu}_j = \text{elasso}(a, b, c) = \mu_j^*$  ▷ Replace by Local mean  
8:      $\mu_{-j}^* = \tilde{\mu}_{-j} + \tilde{\Sigma}_{-j,j} \tilde{\Sigma}_{j,j}^{-1} (\tilde{\mu}_j^* - \tilde{\mu}_j)$  ▷ Update Global Mean  
9:      $\tilde{\Sigma}_{j,j} = \text{vlasso}(a, b, c) = \Sigma_{jj}^*$  ▷ Replace by Updated Local Covariance  
10:     $\tilde{\Sigma}_{j,-j} = \Sigma_{jj}^* \tilde{\Sigma}_{jj}^{-1} \tilde{\Sigma}_{j,-j}$  ▷ Update Global Covariance  
11:     $\tilde{\Sigma}_{-j,j} = \Sigma_{j,-j}^T$  ▷ Update Global Covariance  
12:     $\tilde{\Sigma}_{-j,-j} = \tilde{\Sigma}_{-j,-j} + \tilde{\Sigma}_{-j,j} \tilde{\Sigma}_{j,j}^{-1} (\Sigma_{j,j}^* - \tilde{\Sigma}_{j,j}) \tilde{\Sigma}_{j,j}^{-1} \tilde{\Sigma}_{j,-j}$  ▷ Update Global Covariance  
13: return  $\tilde{\mu}, \tilde{\Sigma}$ 
```

---

The following bullet points summarized the procedure of our method.

- Our target:  $p(\theta|\mathcal{D}) = p(\beta, \sigma^2|\mathcal{D})$ , where a normal distribution is used to approximate, with global parameters  $\tilde{\mu}, \tilde{\Sigma}$ . The initial estimate of  $\tilde{\mu}$  and  $\Sigma$  has been obtained via MFVB. Define  $\theta = (\theta_j, \theta_{-j})$ , separated by a target variable and rest of the variables.  $\mu_j^*, \Sigma_{jj}^*$  as local parameter
- Compute the mean of lasso distribution and variance of lasso distribution with the current  $a, b, c$ :  $q^*(\theta_1) \approx N(\mu_j^*, \Sigma_{jj}^*)$
- We can use the update expression  $q^*(\theta) = q(\theta_{-j}|\theta_j)\phi(\theta_j; \mu_j^*, \Sigma_{jj}^*)$  to adjust the global mean and global covariance of  $q(\theta)$
- Iterates through each variable  $\theta_j$  and Update  $\tilde{\mu}$  and  $\tilde{\Sigma}$  each time by the derived expression.

### 3.4.2 Bivariate local global algorithm

---

**Algorithm 6** Bivariate-Local-Global-Algorithm

---

- 1: Input:  $p(\mathcal{D}, \theta)$ , data  $\mathcal{D}$ , Initialize Variational parameters for each  $q_j(\theta_j)$
  - 2: **while** ELBO has not converged **do**
  - 3:     **for**  $j=1, \dots, p$  **do**
  - 4:          $q_j(\theta_j) \propto \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)]$
  - 5:     Compute  $ELBO(q(\theta)) = \mathbb{E}[\log[p(\mathcal{D}, \theta)]] + \mathbb{E}[\log[q(\theta)]]$
  - 6: **return**  $q(\theta)$
-



# Chapter 4

## Experiment Result and Analysis

### 4.1 Experimental Setting

#### 4.1.1 Parameter selection

#### 4.1.2 Evaluation metric

- $l_1$  norm accuracy

$$\text{Acc}(f, g) = 1 - l_1(f, g) = \int |f(x) - g(x)| dx \quad (4.1)$$

- Common choice for comparing Probability Density Functions

(MORE) Matrix norm Posterior cov and estimated Cov

Objective: Use math to find posterior mode of lasso distribution: Given a,b,c Task: Check if posterior estimate reach Metric: Use Posterior TP/FP Rate(Soft thresholding operator) Check if a local parameter mode is close to 0, compared with true parameter and use this for Variable Selection

Expectation: posterior mode sparse, posterior mean not sparse

#### 4.1.3 Experimental datasets

- **Hitters:**
  - Type: Baseball statistics dataset
  - Predictors ( $p$ ): 20, Samples ( $n$ ): 263
  - Description: Contains baseball player statistics, including performance measures and salary information.

- **Kakadu:**

- Type: Environmental dataset
- Predictors ( $p$ ): 22, Samples ( $n$ ): 1828
- Description: Relates environmental factors to the abundance of amphibians in Kakadu National Park, Australia.

- **Bodyfat:**

- Type: Human body measurements dataset
- Predictors ( $p$ ): 15, Samples ( $n$ ): 250
- Description: Contains measurements of various body parts for a sample of individuals, such as weight, height, and circumferences.

- **Prostate:**

- Type: Medical dataset
- Predictors ( $p$ ): 8, Samples ( $n$ ): 97
- Description: Prostate cancer data with clinical measurements and Logarithm of Prostate Specific Antigen (lpsa).

- **Credit:**

- Type: Credit scoring dataset
- Predictors( $p$ ): 11, Samples ( $n$ ): 400
- Description: Contains information about loan applicants, such as credit history, employment, and demographics, to assess creditworthiness.

- **Eyedata:**

- Type: Medical dataset
- Predictors ( $p$ ): **200**, Samples ( $n$ ): **120**
- Description: Contains measurements related to glaucoma patients, such as intraocular pressure and visual acuity, to assess the effectiveness of treatment options.

## 4.2 Experimental Result

### 4.2.1 Approximation Accuracy Result

Algorithm	MCMC	VB	LG_Local	LG_Global
Min.	100	86.8	97.3	89.3
1st Qu.	100	92.1	99.2	97.0
Median	100	95.7	99.6	97.4
Mean	100	94.2	99.3	97.1
3rd Qu.	100	97.4	99.7	99.0
Max.	100	98.7	99.8	99.7
Run Time(s)	453.75	0.17	0.17	0.17

Table 4.1: Experiment Result on Hitters dataset

### 4.2.2 Approximation Density Visualization

Algorithm	MCMC	VB	LG_Local	LG_Global
Min.	100	93.2	95.9	92.2
1st Qu.	100	98.9	99.8	99.2
Median	100	99.2	99.8	99.4
Mean	100	98.6	99.4	98.8
3rd Qu.	100	99.3	99.8	99.8
Max.	100	99.7	99.8	99.8
Run Time(s)	6696.56	0.14	0.19	0.19

Table 4.2: Experiment Result on Kakadu dataset

Algorithm	MCMC	VB	LG_Local	LG_Global
Min.	100	91.3	96.3	90.7
1st Qu.	100	97.0	99.6	97.6
Median	100	98.0	99.7	98.4
Mean	100	97.0	99.2	97.2
3rd Qu.	100	98.4	99.7	98.6
Max.	100	99.3	99.7	99.7
Run Time(s)	398.59	0.14	0.17	0.17

Table 4.3: Experiment Result on bodyfat dataset

Algorithm	MCMC	VB	LG_Local	LG_Global
Min.	100	96.9	99.5	97.4
1st Qu.	100	97.2	99.5	97.9
Median	100	97.6	99.6	98.9
Mean	100	97.5	99.6	98.7
3rd Qu.	100	97.7	99.6	99.5
Max.	100	98.4	99.6	99.6
Run Time(s)	336.31	0.11	0.12	0.12

Table 4.4: Experiment Result on Prostate dataset

Algorithm	MCMC	VB	LG_Local	LG_Global
Min.	100	91.8	99.5	99.3
1st Qu.	100	98.3	99.7	99.5
Median	100	99.4	99.8	99.5
Mean	100	97.9	99.7	99.6
3rd Qu.	100	99.5	99.8	99.7
Max.	100	99.7	99.8	99.8
Run Time(s)	359.92	0.1	0.11	0.11

Table 4.5: Experiment Result on Credit dataset

Algorithm	MCMC	VB	LG_Local	LG_Global
Min.	100	78.4	97.3	86.1
1st Qu.	100	86.9	98.6	90.4
Median	100	90.3	98.7	91.3
Mean	100	88.9	98.7	91.8
3rd Qu.	100	91.2	98.8	92.3
Max.	100	93.1	99.1	99.2
Run Time(s)	18144.7	1.21	1.72	1.72

Table 4.6: Experiment Result on Eyedata dataset

# Chapter 5

## Discussion and Conclusion

Bodyfat, Prostate, Credit: no correlation; VB does do well because of that

Kakadu  $n = 2000$ ;

Hard for VB: Hitters: high correlation(between predictors) and Eyedata:  $p \ll n$ .

### 5.1 Limitation

### 5.2 Future Work

### 5.3 Conclusion

# Bibliography

- Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2, 183–202. doi:[10.1137/080716542](https://doi.org/10.1137/080716542).
- Beech, D.G., Kendall, M.G., Stuart, A., 1959. The advanced theory of statistics. volume 1, distribution theory. *Applied Statistics* 8, 61. URL: <https://doi.org/10.2307/2985818>, doi:[10.2307/2985818](https://doi.org/10.2307/2985818).
- Bishop, C.M., 2006. Pattern recognition and machine learning. Springer.
- Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112, 859–877. doi:[10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773).
- Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3, 993–1022.
- Boyd, S., Vandenberghe, L., 2004. Convex optimization. Cambridge university press.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39, 1–38. URL: <http://www.jstor.org/stable/2984875>.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. *The Annals of Statistics* 32. doi:[10.1214/009053604000000067](https://doi.org/10.1214/009053604000000067).
- Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33. doi:[10.18637/jss.v033.i01](https://doi.org/10.18637/jss.v033.i01).
- Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6, 721–741. doi:[10.1109/TPAMI.1984.4767596](https://doi.org/10.1109/TPAMI.1984.4767596).



- Khare, K., Hobert, J.P., 2013. Geometric ergodicity of the bayesian lasso. *Electronic Journal of Statistics* 7. doi:[10.1214/13-ejs841](https://doi.org/10.1214/13-ejs841).
- Ormerod, J.T., Wand, M.P., 2010. Explaining variational approximations. *The American Statistician* 64, 140 – 153.
- Parisi, G., Shankar, R., 1988. *Statistical field theory* .
- Park, T., Casella, G., 2008. The bayesian lasso. *Journal of the American Statistical Association* 103, 681–686. doi:[10.1198/016214508000000337](https://doi.org/10.1198/016214508000000337).
- Rajaratnam, B., Sparks, D., 2015a. Fast bayesian lasso for high-dimensional regression .
- Rajaratnam, B., Sparks, D., 2015b. Mcmc-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. URL: <https://arxiv.org/abs/1508.00947>.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 267–288. doi:[10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x).
- Zhang, N., Zeng, S., 2005. A gradient descending solution to the lasso criteria. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. doi:[10.1109/ijcnn.2005.1556393](https://doi.org/10.1109/ijcnn.2005.1556393).