

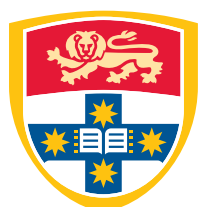
# A new Approximate Bayesian Inference algorithm for Bayesian Lasso: A local approximation adjusting approach

Yuhao Li

Supervisor: A/Prof. John Ormerod

A thesis submitted in partial fulfillment of  
the requirements for the degree of  
Bachelor of Science(Honour)(Data Science)

Mathematics and Statistics



THE UNIVERSITY OF  
**SYDNEY**

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 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 discovering the Lasso problem from Bayesian perspective, the Bayesian Lasso problem use a double-exponential prior for modelling variation of inferential quantity and obtaining interval estimate of coefficients. In addition, an automatic tuning process of tuning parameter  $\lambda$  that control the strength of penalization can also be performed in the Bayesian framework instead of time-consuming  $n$ -fold cross validation technique under the ordinary Lasso.

On the other hand, obtaining the posterior of the Bayesian Lasso model involves using Monte Carlo Markov Chain method such as Gibbs sampler for exact posterior distribution. Even though MCMC is famous for its oracle property such as generating arbitrary exact target samples if burn-in period is enough, MCMC is slow with 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, while it is also a faster alternative to Monte Carlo methods such as Markov Chain and Monte Carlo(MCMC).

The main idea behind Variational Approximation is: given an assumed distribution set, it will search for an optimal posterior distribution by continuing minimizing the gap between true posterior and estimated posterior such as using Kullback–Leibler divergence(KL-divergence) as a distance metric.

Nevertheless, elegant property in MCMC such as obtaining exact posterior if infinite burn-in time period is assigned, doesn't occur in Variational Inference, which means the

approximation accuracy will be a pivotal concern as unsatisfied distribution such as under-estimating the variance when the correlation of variables becomes large.

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

In this thesis, we will firstly introduce univariate and the multivariate lasso distribution, which are newly discovered distribution that could be matched for aiding marginal distribution approximation, followed by the introduction of two fast and more accurate Variational Approximation algorithms and their application in the Bayesian Lasso regression problem. By assuming the global parameter assuming Gaussian Approximation, the information of local parameter distribution would be accommodated by the lasso distribution so that global approximated distribution would be obtained by product of local distribution and conditional Gaussian distribution.

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

To verify the efficiency and accuracy of our algorithm, numerous experiments would be conducted under real-world datasets such as Hitters Dataset using several evaluation metric such as  $l_1$  accuracy and matrix norm. Our result suggest their high Variational Approximation accuracy with a descent time efficiency, compared with the traditional Monte Carlo methods and Mean-Field Variational Bayes (MFVB).

## **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 Definition and 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 Bayesian Expectation Maximization . . . . .	17
2.5 Variational Inference . . . . .	17
2.5.1 Mean Field Variational Bayes(MFVB) . . . . .	17
2.5.2 MFVB for Bayesian Lasso . . . . .	17
<b>3 Methodology</b>	<b>18</b>
3.1 Lasso distribution . . . . .	18
3.1.1 Univariate Lasso Distribution . . . . .	18
3.1.2 Multivariate Lasso Distribution . . . . .	18
3.2 Local-Global Algorithm . . . . .	18

<b>4</b>	<b>Experiment Result and Analysis</b>	<b>19</b>
4.1	Experimental Setting . . . . .	19
4.1.1	Parameter selection . . . . .	19
4.1.2	Evaluation metric . . . . .	19
4.1.3	Experimental datasets . . . . .	19
4.2	Experimental Result . . . . .	19
<b>5</b>	<b>Discussion and Conclusion</b>	<b>20</b>
	<b>Bibliography</b>	<b>21</b>

# List of Figures

1.1	Variational Inference intuition, where $X$ is data $\mathcal{D}$ , $D$ 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



# List of Tables

# 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 for his success in Statistical Community over the years.

The Lasso serves as two purposes, one is the estimation of regression parameter, the other is to 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 data because of its sparsity nature.

**Explain purpose of Lasso** The definition of linear regression model of interest can be referred based on the following definition defined by [Tibshirani \(1996\)](#): the  $n \times 1$  vector of regression coefficient  $\beta$ ,  $y$  is the response variable with a dimension of  $n$ ,  $X$  is data matrix after standardization with a dimension of  $n \times p$ ,  $\mu$  is the population mean with a dimension of  $n \times 1$ ,  $\epsilon$  is independent and identically distributed normal noise with expectation of 0 and variance of  $\sigma^2$ . Then the linear model can be explained by the Equation [1.1](#)

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

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 Lasso estimate of regression coefficient is based on Equation [\(1.3\)](#), where the main distinction of Lasso is adding a penalty term of absolute value of regression coefficients  $\beta$  in

addition to the sum of squared value of residuals from ordinary regression objective function. The value of tuning parameter  $\lambda$  is served as a measure of the extent of penalization.

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (\tilde{y} - X\beta)^T (\tilde{y} - X\beta) + \lambda \|\beta\|_1, \lambda \geq 0. \tilde{y} = y - \bar{y}\mathbf{1}_n \quad (1.3)$$

Larger penalization leads to more sparse solution of regression coefficient due to a square constraint set results from  $L_1$  penalty function, so that achieve variable selection of parameter, generating a higher prediction accuracy as well as 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 of the detriments of the ordinary lasso is that the variation of inferential quantity can't be captured properly. To resolve this issue, [Tibshirani \(1996\)](#) also suggests that the lasso estimate can also be extended under the Bayesian framework, which can be described as posterior mode Equation(1.4) if independent and identically distributed Laplacian prior from Equation (1.5) is assigned, together with the likelihood form on Equation(1.6).

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

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

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

**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, where 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 to ensure unimodality of full posterior distribution. Based on closed form of tractable posterior distribution, a three-step Gibbs sampler is proposed to draw approximate 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 more computation of conditional distribution form is demanded. Secondly Bayesian credible interval of parameters can be generated simultaneously for modelling uncertainty and therefore can also guide variable selection based on the interpretation that lasso estimated is regarded as the mode of posterior distribution of  $\beta$ . Thirdly, [Park and Casella \(2008\)](#) also state that the Bayesian Lasso model could also be a potential solution for addressing the issue of attaining optimal tuning parameter  $\lambda$  by marginal maximum likelihood method together with a suitable hyperprior such as gamma prior on the square of tuning parameter  $\lambda$ :  $\lambda^2$ . It could support a more stable automatic tuning process of choosing the most appropriate tuning parameter  $\lambda$  for the Lasso problem, as opposed to inefficient  $K$ -fold cross validation approach under the ordinary Lasso model that is time consuming and computational 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 \(2013\)](#) has demonstrated a Bayesian Lasso Gibbs sampler version of central limit theorem (CLT), indicating Bayesian Lasso Gibbs Sampler satisfy geometrically ergodicity if any values of sample size  $n \geq 3$  and arbitrary number of regression coefficient  $p$ , data matrix  $X$ , tuning parameter  $\lambda$  are assigned. 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**

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, the exact posterior distribution form might be too complicated to be tractable,. Thirdly, the computation of the posterior mean parameter  $\int_{\theta} \theta p(\theta|\mathcal{D})$  has high probability without having an simple calculation,

resulting in the fact that there might not exist an 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 Approximate Bayesian Inference method, which includes stochastic approach such as MCMC, where an exact result can be obtained if infinite computational resource is assigned. Another category lies in deterministic approach that as an faster substitution compared with stochastic approximation approaches. As stated above, an approximate inference method such as MCMC is used for posterior distribution estimation. The necessity of approximate bayesian inference method is due to its' expensive and infeasible exact posterior distribution parameter estimate.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \quad (1.7)$$

**Challenges of Bayesian Lasso model** Returning back to the Bayesian Lasso model, the three-step Gibbs sampler belongs to the class of MCMC algorithm, which produce samples demonstrates gradually decreasing correlated samples as the burn-in period is increased. Burn-in period refers to the time the sample before that period have to be abandoned. On the other hand, the three-step Gibbs sampler from Markov Chain Monte Carlo(MCMC) under the class of stochastic approximation algorithm 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, especially when  $n$  is small and  $p$  is large derivation according to the upper bound of convergence rate by [Rajaratnam and Sparks \(2015b\)](#).

**Approximation Algorithm: Deterministic type & VI.** Due to limitation of stochastic type of algorithm, alternative methods such as deterministic Variational Inference methods has becoming popular due to its fast speed and simple computation. Numerous algorithms have been designed and utilized widely such as Variational Bayes, Expectation Propagation algorithms etc. A common variation is Coordinate Ascent Variational Inference approach produced by [Blei et al. \(2003\)](#) , which assume the approximation originates from a analytically tractable class of distribution  $Q$  first and attempt to search

for the distribution from this family that is the closest to the target posterior distribution with some discrepancy metric such as the Kullback–Leibler divergence etc. An optimization based system in Equation (1.8) is established by iteratively updating variational parameter with an appropriate optimization algorithm such as Coordinate Ascent to obtain approximated posterior distribution that is 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 provides further explanation of aforementioned intuition, goal and procedure of Variational Inference.

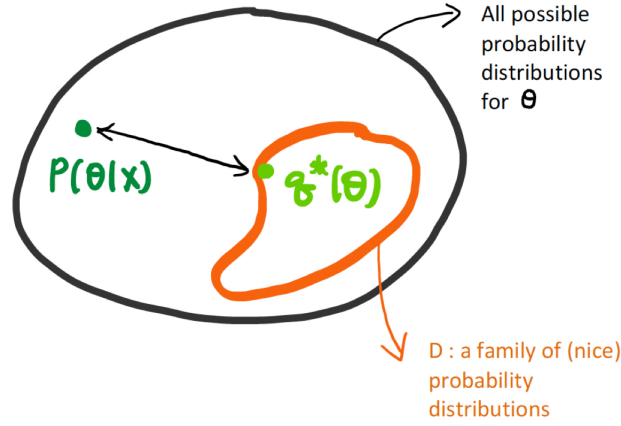


Figure 1.1: Variational Inference intuition, where  $X$  is data  $\mathcal{D}$ ,  $D$  is equivalent to  $Q$  defined above

$$q^*(\theta) = \underset{q \in Q}{\operatorname{argmin}} \operatorname{KL}(q(\theta) || p(\theta|\mathcal{D})) := \int q(\theta) \log\left(\frac{q(\theta)}{p(\theta|\mathcal{D})}\right) d\theta \quad (1.8)$$

$$\operatorname{KL}(q || p(\cdot|\mathcal{D})) = - \int q(\theta) \log\left(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)}\right) d\theta + \log p(\mathcal{D}). \quad (1.9)$$

In addition, the exact form of KL divergence can be found in Equation (1.9) In practice, the minimization of KL divergence from Equation (1.8) would be converted into an equivalent formulation in Equation (1.10) that maximize the lower bound of  $\log(p(y))$  due to complexity of minizing the original KL divergence, which can also be called Evidence Lower

Bound(ELBO) that can be defined by Equation (1.10).

$$q^*(\theta) = \operatorname{argmax}_{q_\theta \in Q} \text{ELBO}(q(\theta)), \quad (1.10)$$

$$\text{ELBO}(q(\theta)) = \int q(\theta) \log\left(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)}\right) = \mathbb{E}_{q(\theta)} \log\left(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)}\right). \quad (1.11)$$

### Mean Field Variational Bayes, History and introduction

The most traditional Variational Inference algorithm is known as Mean-Field Variational Bayes motivated by mean-field Theory in statistical physics yielded by [Jordan et al. \(1998\)](#) and [Attias \(1999\)](#), which assume the approximated distribution is from independent product of parameter distribution from set  $Q$  as in Equation (1.12) if assuming there are  $k$  sub-parameter of target parameter  $\theta$ .

$$q(\theta) = \prod_{i=1}^k q_i(\theta) \quad (1.12)$$

Mean-Field Variational Bayes have been adapted and developed over the next two decades, especially in mixture modelling and probabilistic graphical model. It sometimes provide efficient variable selection as well, We will introduce more about the algebra of MFVB in Chapter 2. Variational Inference community benefits from MFVB due to its tractable family of

**Advantages of Variational Inference** Variational approach is more superior and appropriate at present for several reasons according to [Blei et al. \(2017\)](#). 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

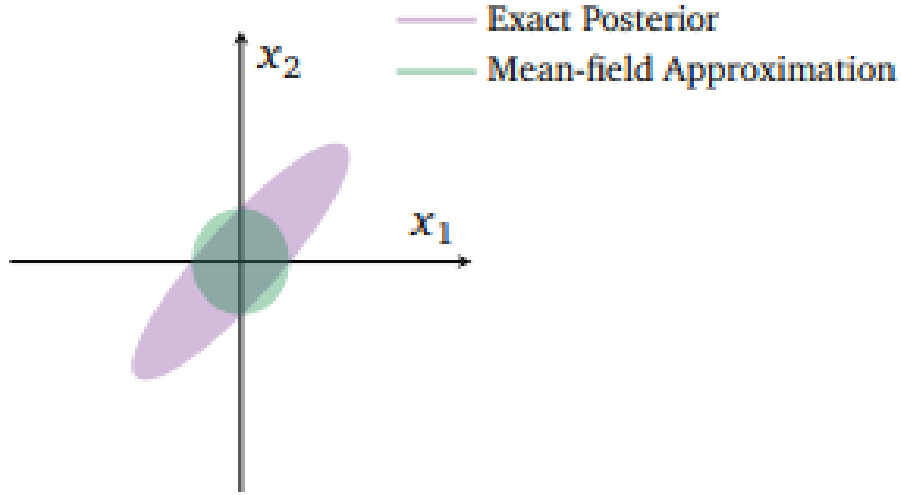


Figure 1.2: Visualization of Mean-Field Variational Approximation compared with exact posterior when correlation is large

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 result. Figure 1.2 further demonstrates this phenomenon, the true 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 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** Motivated by the intention of further enhancing the approximation accuracy of Variational Bayes, as well as the oracle property of Bayesian Lasso regression coefficient estimation for variable selection and standard error estimation leads to more



desired demand for fast approximate inference. We would like to design new Variational Inference 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 based on the Mean-Field Variational Bayes assumption.

Feeding initial value of mean  $\mu$  and covariance parameter  $\Sigma$  and writing out the marginal likelihood form of regression coefficient  $\beta_i$  for  $j$  th variable,  $\log(p(\mathcal{D})|\beta_j)$ , we've invented a new distribution called univariate lasso distribution for matching the marginal likelihood. Mixing each marginal likelihood with a Gaussian Approximation, we've shown its' approximation accuracy have surpassed every existing algorithm such as Mean-Field Variational Bayes. Even though the speed of our algorithm is slightly slower than MFVB, the approximation accuracy illustrates a tiny 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've also purposed another algorithm based on marginal likelihood estimation by a bivariate lasso distribution. Instead of updating corresponding mean, covariance matrix for each variable for each iteration, the marginal likelihood of each pair of variables would be matched, so that further generalize our algorithm. Our conclusion are the univariate lasso algorithm would be faster with a lower accuracy while bivariate lasso algorithm would be 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 multivariate Lasso distribution to each of the marginal distribution, an improved estimate for global Gaussian Approximation can be obtained as defined in Equation (1.13). Finally, we will show our experiment result in 4 using various accuracy metrics.

Our contribution have been listed in the following subsection 1.2.

$$q^*(\theta) \approx N(\mu^*, \Sigma^*) \tag{1.13}$$

## 1.2 Contribution

Our main contribution could be concluded as the following part:

- 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 Variational Inference 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

This paper will be divided up into 5 chapters. Chapter 1 briefly illustrate the motivation and background of the Lasso problem, Bayesian Lasso Problem, the motivation for Approximate Bayesian Inference, with a specific focus on deterministic Variational Approximation. Chapter 2 will briefly review and explain the details of the methodology in previous work such as the lasso problem, Approximate Bayesian Inference algorithm, MCMC(Monte Carlo Method), Bayesian Expectation Maximization algorithm and their variants and Mean-Field Variational Bayes(MFVB). We will 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

## Definition and Literature Review

### 2.1 Bayesian Inference Paradigm

**Why Bayesian** Bayesian Inference approaches shares numerous advantages in statistic community and application areas, particularly for the circumstance when there is lack of data. An appropriate prior choice can be beneficial in aforementioned case, especially for medical problem where the amount of effective data is extremely rare and untenable. Additionally, unlike frequentist inference approaches which treat parameter estimate as a fixed value, Bayesian Inference approaches regard parameter estimate as a random variable that have probability distribution, which means interval estimate and error variance would be generated for capturing uncertainty, offering belief and confidence for interpreting parameter estimates.

**Bayesian Inference Intuition** Bayesian inference approach stems from the Bayes rule, which is defined as Equation (1.7) 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 (2.1) since  $p(\mathcal{D})$  is equal to constant and is also insignificant to know overall posterior distribution.

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

## 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 2.2, where  $t \geq 0$  is denoted as a tuning term  $t$ , regression coefficient is  $\beta$ ,  $||\beta||_1$  is the  $l_1$  norm of beta,  $||\beta||_2$  is the  $l_2$  norm of  $\beta$ , data matrix is  $X$ , response variable is  $y$ . The estimation for lasso estimate  $\hat{\beta}_{lasso}$  is defined by Equation 2.2.

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2, s.t. ||\beta||_1 \leq t, t \geq 0. \quad (2.2)$$

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 2.3

$$\mathcal{L}(\beta, \lambda) = ||y - X\beta||_2 + \lambda ||\beta||_1 - \lambda t, \lambda \geq 0 \quad (2.3)$$

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 2.2 after ignoring the unaffected constant  $-\lambda t$ .

Graphical demonstration of the lasso for Equation 2.2 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. Given that  $\lambda$  is set as penalty term that control the strength of penalization, larger penalization facilitate a more sparse solution, so that further enclosing the estimated coefficient to lies on the axis of each parameter as shown in Figure 2.1. Lasso regression coefficient would have higher chance to render the contour line of  $\beta$  intersect with the corner of the squared constraint set, causing the occurrence of sparse estimated regression coefficient. Compared with the ridge regression



Figure 2.1: Graphical comparison between lasso regression and ridge regression

where a sum of square of penalty term is yielded instead on the right side of Figure 2.1, ridge regression tends to gain a non-sparse solution due to circled constraint set for  $\beta$ .

**Problem** In addition, the optimal estimated  $\beta_{lasso}$  can be generated by taking the derivative with respect to  $\beta$  and solving the normal equation, denoted as Equation (??). In addition, lasso estimated can be efficiently computed via Least Angle Regression algorithm by

## 2.3 Bayesian Lasso

### 2.3.1 Bayesian Lasso model

Park and Casella (2008) has proposed an alternative conditional Laplace prior formula of the form Equation (1.5), expanded by Equation(2.4)

$$\pi(\beta) = \prod_{j=1}^p \frac{\lambda}{2} e^{-\lambda|\beta_j|} \quad (2.4)$$

For a given variance, the mode of posterior form in Equation(2.5) is consistent with the estimate of lasso equation in 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 appendeix from [Park and Casella \(2008\)](#).

$$\pi(\beta, \sigma^2 | \tilde{y}) \propto \pi(\sigma^2) (\sigma^2)^{-(n-1)/2} \exp\left(\frac{1}{2\sigma^2} (\tilde{y} - X\beta)^T (\tilde{y} - X\beta) - \lambda \sum_{j=1}^p |\beta_j|\right) \quad (2.5)$$

$$\pi(\beta | \sigma^2) = \prod_{j=1}^p \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \quad (2.6)$$

To remedy this issue, a conditional Laplacian prior from [2.6](#) with respect to Equation [\(2.4\)](#) has been designed, ensuring the unimodality of the posterior for  $\beta$ , and the current prior with respect to  $\beta$ ,  $\sigma^2$  can be written in

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

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 [2.7](#), given an improper prior selection for  $\pi(\sigma^2) = \frac{1}{\sigma^2}$  and  $\lambda \geq 0$ . Additionally, an additional latent variable  $\tau$  are introduced as a scale mixture of Gaussians for reformulation of conditional prior [2.6](#) as [2.8](#), 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.

$$\frac{\alpha}{2} e^{-\alpha|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{\alpha^2}{2} e^{-\alpha^2 s/2} ds, \alpha > 0 \quad (2.8)$$

Finally, the hierarchical bayesian lasso model functional form can be written as Equation [\(2.9\)](#).

$$\begin{aligned} y | \mu, X, \beta, \sigma^2 &\sim N_n(\mu + X\beta, \sigma^2 I) \\ \beta | \tau_1^2, \dots, \tau_p^2 &\sim N_p(0, \sigma^2 D_\tau) \\ D_\tau &= \text{diag}(\tau_1^2, \dots, \tau_p^2) \\ \tau_1^2, \dots, \tau_p^2 &\sim \prod_{j=1}^p \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \tau_1^2, \dots, \tau_p^2 > 0 \\ \sigma^2 &\sim \pi(\sigma^2) = 1/\sigma^2, \sigma^2 > 0 \end{aligned} \quad (2.9)$$

## 2.3.2 Bayesian Lasso Gibbs Sampler

### Gibbs Sampler

Gibbs sampler is a Markov Chain Monte Carlos sampling algorithm for efficient sampling of any probability density function, given corresponding conditional distribution form with an reasonable burn-in period. For each iteration, each parameter of interest will be sampled once by conditional distribution for the current iteration. After reasonable burn-in period, it will return an exact posterior distribution after discarding samples generated before burn-in period. After completion, the combination of each component of samples can formulate full samples of posterior distribution. As a result, the functional form of conditional distribution given any other parameter of interest has to be acquired. For the bayesian lasso model, this means the following exact functional form has to be derived

$$\begin{aligned} p(\beta|\mathcal{D}, \sigma^2, \tau) \\ p(\sigma|\mathcal{D}, \sigma^2, \tau) \\ p(\tau_j|\mathcal{D}, \sigma^2, \tau) \end{aligned} \tag{2.10}$$

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:** Before derivation, we would like to formulate our initial setting:  $\lambda$  is fixed,  $\pi(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-b^2/\sigma^2}$ ,  $\sigma^2 > 0, a > 0, b > 0$  follows an Gamma distribution with parameter  $a$  and  $b$ .

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

**Joint distributional form:** Given Equation 2.9, we can write joint distribution form as

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

**Formulation 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

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

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

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

**Formulation of  $\sigma^2$ :**

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

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

**Formulation of  $\tau_j$ :**

$$\begin{aligned} p(\tau_j|\tilde{y}, \beta, \sigma^2) &\propto p(\tilde{y}, \beta, \tau, \sigma^2) \\ &= \frac{1}{\sqrt{\frac{2\pi\sigma^2\tau_j}{\lambda^2}}} \exp\left(-\frac{\beta_j^2 \lambda^2}{2\sigma^2 \tau_j}\right) \exp\left(-\frac{1}{2}\tau_j\right) \sim GIG(a^*, b^*, p^*) \end{aligned} \quad (2.15)$$

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

**Summary** In summary, each conditional distribution can be written in the following



equation

$$\begin{aligned}
p(\beta|y, \sigma^2, a) &: \text{MVN}((X^T X + \lambda^2 A^{-1})^{-1} X^T y, (X^T X + \lambda^2 A^{-1})^{-1} \sigma^2) \\
p(\sigma^2|y, \beta, \tau_j) &: \text{Inverse-Gamma}(\frac{n}{2} + \frac{p}{2} + a, \frac{\|y - X\beta\|_2^2}{2} + \frac{\lambda^2 \sum_j \beta_j^2}{2\tau_j} + b) \\
p(\tau_j|y, \beta, \sigma^2) &: GIG(1, \frac{\beta_j^2 \lambda^2}{\sigma^2}, 1/2)
\end{aligned} \tag{2.16}$$

The gibbs sampler can be established by the following algorithm.

---

**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  $\tau_j^{(t+1)} \sim GIG(1, \frac{\beta_j^2 \lambda^2}{\sigma^2}, 1/2)$  ▷ Generate sample  $\tau_j$
  - 7:      $t \leftarrow t + 1$
- 

**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 Casella \(2008\)](#) arrange a hyperprior to  $\lambda^2$  from Gamma distribution as 2.17. 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.

$$\pi(\sigma^2) = \frac{\delta^\gamma}{\Gamma(\gamma)} (\lambda^2)^{\gamma-1} e^{-\delta \lambda^2}, \text{ for } \delta > 0, \gamma > 0, \lambda^2 > 0 \tag{2.17}$$

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 in the 2.16.

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 gamma distribution has shape  $\theta$  and rate parameter  $\gamma$ .

(2.18)

$$\begin{aligned} p(\lambda|\tilde{y}, \beta, \sigma^2, \tau) &\propto p(\tilde{y}, \beta, \tau, \sigma^2, \lambda) \\ &= \left(\prod_{j=1}^p \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2 / 2}\right) (\lambda^2)^{\gamma-1} e^{-\delta \lambda^2} \\ &= (\lambda^2)^{p+\gamma-1} e^{-\lambda^2 (\frac{1}{2} \sum_{j=1}^p \tau_j^2 + \delta)} \end{aligned} \quad (2.19)$$

Thus, the posterior distribution of  $\lambda$  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 2.19, based on using an Augmented Gibbs sampler.

## 2.4 Expectation Maximization

### 2.4.1 Bayesian Expectation Maximization

## 2.5 Variational Inference

### 2.5.1 Mean Field Variational Bayes(MFVB)

Suppose there are  $n$  number of parameters, then MFVB assumes target distribution  $q(\theta)$  is the product of single factorization of each parameter distribution  $q_i(\theta_i)$ , due to simplicity of product density form.

$$q(\theta) = \prod_{i=1}^n q_i(\theta_i) \quad (2.20)$$

To measure the similarity between true distribution and target distribution, KL divergence metric is selected to produce

$$KL(q(x)||p(x|\mathcal{D})) \quad (2.21)$$

### 2.5.2 MFVB for Bayesian Lasso

# Chapter 3

## Methodology

### 3.1 Lasso distribution

#### 3.1.1 Univariate Lasso Distribution

Basic Property

Derivation

#### 3.1.2 Multivariate Lasso Distribution

Basic Property

Derivation

### 3.2 Local-Global Algorithm

# Chapter 4

## Experiment Result and Analysis

### 4.1 Experimental Setting

#### 4.1.1 Parameter selection

#### 4.1.2 Evaluation metric

L1 accuracy

(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

toy dataset 3-4 datasets

### 4.2 Experimental Result

# Chapter 5

## Discussion and Conclusion

# Bibliography

- Attias, H., 1999. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. p. 21–30.
- 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.
- 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).
- Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1998. An introduction to variational methods for graphical models. *Learning in Graphical Models* , 105–161doi:[10.1007/978-94-011-5014-9\\_5](https://doi.org/10.1007/978-94-011-5014-9_5).
- 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).

- 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).