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

Yuhao Li

Supervisor: A/Prof. John Ormerod and

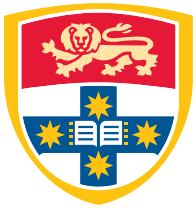
Dr. Mohammad Javad Davoudabadi

A thesis submitted in partial fulfillment of

the requirements for the degree of

Bachelor of Science(Honour)(Data Science)

School of 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 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 has sum of squared residual only for ordinary regression, generating and eliminating sparse coefficient estimates 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 λ, 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 being able to generate samples from the posterior distribution if the Markov Chain is run for enough iterations, it is slow and has a high computational cost.

Meanwhile, variational approximations, as a deterministic class of approximation algorithms for intractable posterior distributions, have been applied prevalently for fast Approximate Bayesian Inference (ABI) in 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 distributional forms, and subsequently identifying the density within that set that best approximates the target. The level of approximation is determined via the Kullback-Leibler (KL) divergence.

Nevertheless, variational approximation unlike the MCMC methods is not guaranteed to be exact. It can only find a density close to the target which means the approximation

accuracy might be a concern when it fails to provide reliable estimates of posterior variances. In order to address the slow speed of obtaining the posterior distribution for the Bayesian Lasso model, new alternative ABI methods need to be explored, especially for deterministic algorithms such as for variational Bayes and their variants.

This thesis presents two novel distributions, the univariate and bivariate Lasso distributions. These distributions can be used to assist in approximating marginal posterior distributions. In addition, this thesis introduces two highly efficient and precise variational approximation-based algorithms that we apply to solve the Bayesian Lasso regression problem.

The first method involves matching with a marginal univariate lasso distribution by updating a global parameter for each variable per iteration. In addition, 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 arises when an initial diagonal covariance matrix is assigned.

To verify the efficiency and accuracy of our algorithm, we conducted numerous experiments using benchmark datasets and evaluated it using several metrics such as l_1 accuracy and running speed.

Acknowledgements

Firstly, I would like to express my deep thanks to my supervisor: A/Prof: John Ormerod and Co-supervisor Dr. Mohammad Javad Davoudabadi, for their constant guidance and patience throughout this year. I would not make it so far without them. Secondly, I would like to thank my friends, my family, my mother, father. The entire honor year has been full of challenges not only in researching a novel area but also in the heavy coursework workload. I could not have overcome them without your support behind the scenes. Finally, I would like to thank the University of Sydney, for providing me a chance to study and research in this school. It is an unforgettable experience.

Contents

List of Figures

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 the class of shrinkage methods. As one of the traditional shrinkage methods, Lasso regression has been proven very useful in Statistical Community over the years.

The Lasso serves two purposes. One is the estimation of regression parameters and the second is the effective shrinkage of the coefficients to achieve variable selection. This is a fundamental difference between Lasso with some other shrinkage methods. The Lasso regression is helpful particularly for high-dimensional because of sparse estimates of coefficients.

Explain the purpose of Lasso: The linear regression model is

$$y = \mu 1_n + X\beta + \epsilon, \tag{1.1}$$

where β is a $p \times 1$ vector of the regression coefficient, y is a $n \times 1$ vector of the response variable, X is a $n \times p$ matrix of the covariate, μ is a $n \times 1$ vector of the population mean, and ϵ is the model uncertainty which is distributed normally with mean 0 and variance σ^2.

The least squares estimator suggests the sum of the square of the difference between the estimated response variable and the response variable should be used as a loss function as described in Equation 1.2

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} (\widetilde{y} - X\beta)^T (\widetilde{y} - X\beta). \tag{1.2}$$

The lasso regression estimates linear regression coefficients through L_1- constrained least squares. The regression coefficients are estimated based on the lasso method as follows:

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

where $\lambda \geq 0$, and $\widetilde{y} = y - \overline{y}\mathbf{1}_n$. The tuning parameter λ controls the strength of the penalty. Larger values of λ lead to a more sparse solution for the regression coefficient due to a square constraint set resulting from L_1 penalty function, resulting in the implicit variable selection of parameter. This can generate a higher prediction accuracy as well as a more interpretable model since we are able to drop the estimated regression coefficient that are estimated as 0 and state they have a weak effect for prediction according to Tibshirani (1996).

However, due to the non-existence of derivatives of absolute value of regression coefficient β, alternative improved algorithms have been purposed and deployed such as least angle regression (LARS), iterative soft-thresholding, subgradient methods, 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 quantities. To address this issue, Tibshirani (1996) suggested extending lasso estimation 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 1.6 to obtain the posterior mode Equation 1.5, i.e.,

$$\widehat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmax}}\ p(\beta|y, \sigma^2, \tau), \quad \text{where} \quad \tau = \frac{\lambda}{2\sigma^2} \tag{1.4}$$

with

$$p(y|\beta, \sigma^2) = N(X\beta, \sigma^2 I_n), \quad \text{and} \tag{1.5}$$

$$p(\beta_j) = \left(\frac{\tau}{2}\right) \exp(-\tau|\beta_j|). \tag{1.6}$$

Unfortunately, no tractable solution existed for fitting the Bayesian Lasso until Park and Casella (2008) who explored the lasso model within the Bayesian setting. The choice of a conditional Laplace prior distribution over the regression coefficient β conditioning by standard error σ^2 is added to the Lasso penalty formulation in the frequentist framework. They show that this ensures the unimodality of the posterior distribution. A three-step Gibbs sampler is proposed to draw samples from the Bayes Lasso posterior distribution, which can be utilized for further inference of parameters of interest.

There are several benefits to using the Bayesian Lasso formulation. Firstly, it is easier to implement than the traditional Lasso, although it is more computationally demanding. 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 selecting the tuning parameter λ. 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 (on λ or λ^2). This approach could support a stable automatic tuning process of choosing an appropriate tuning parameter λ for the Lasso model. This is in contrast to the inefficient K-fold cross-validation approach typically used to tune the frequentist 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 the exact forms of the full conditional distributions of each model parameter.

Theoretically, Khare and Hobert (2013) demonstrate a Bayesian Lasso Gibbs sampler version of central limit theorem (CLT) which indicates that Bayesian Lasso Gibbs sampler satisfies geometric ergodicity for any values of sample size $n \geq 3$ for an arbitrary number of regression coefficient p, data matrix X, tuning parameter λ.

Approxiamte Bayesian Inference: A review of Bayesian inference occurs in Section 2.1, but challenges of Bayesian inference motivate the approximate Bayesian inference method thereafter. To be specific, Bishop (2006) states three main challenges of obtaining a posterior distribution. Firstly, the dimension of the target parameter might be high, which results in heavy computational costs 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 $\int_\theta \theta\, p(\theta|\mathcal{D})d\theta$ does not have a closed-form analytical solution for integration. Much effort has been made over the years, there are two main types of sampling approaches that are effective currently, which are stochastic sampling algorithms and deterministic approximation algorithms.

There are two genres of ABI methods, which include stochastic approaches such as Markov Chain Monte Carlo (MCMC), where an exact result can be obtained if infinite

computational resources are assigned. The other category lies in deterministic approaches, which provide a faster substitution compared to stochastic approximation approaches. As stated above, approximate inference methods such as MCMC are used for posterior distribution estimation. The need for approximate Bayesian inference methods arises from the difficulties and impracticality of estimating the posterior distribution of the model parameters.

Challenges of Bayesian Lasso model: The three-step Gibbs sampler belongs to the class of MCMC algorithms and is a technique for sampling from a probability distribution by constructing a Markov chain that has the posterior distribution as its equilibrium distribution, where each variable is sampled in turn given the current values of the other variables. The burn-in period in MCMC methods is the initial set of iterations that are discarded before collecting samples with the purpose of ensuring that the measurements are stable and consistent before beginning the actual analysis. The rationale behind this is to reduce the dependence on the initial values and mitigate the impact of starting the chain from a poor position in the state space. The length of the burn-in period can vary based on the specific characteristics of the chain, and determining the appropriate length often involves some level of subjective judgment. The three-step Gibbs sampler, however, has several challenges. One of the fundamental challenges is its slow convergence. This issue is more highlighted in high-dimensional problems, and it increases the execution time of the algorithm to reach the stationary distribution.

Approximation Algorithms: 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 variational Bayes (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. Afterward, 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 (KL) divergence. An optimization-based system in Equation 1.7 is established by iter-

atively updating variational parameters with an appropriate optimization algorithm. An easy-to-implement optimization algorithm in this context is Coordinate Ascent which 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 distributions. To further illustrate the intuition, Figure 1.1 provides further explanation of the aforementioned intuition, goal, and procedure of Variational Inference.

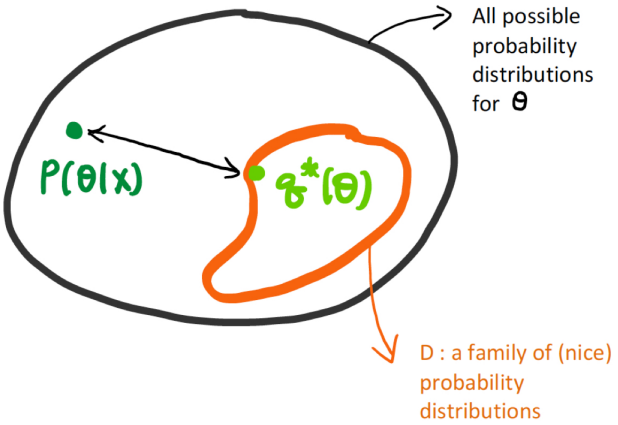


Figure 1.1: Variational Inference intuition, where X is data $\mathcal{D}$, $\mathcal{D}$ is equivalent to Q defined above

$$q^*(\theta) = \underset{q_\theta \in Q}{\operatorname{argmin}} \quad \mathrm{KL}(q(\theta)||p(\theta|\mathcal{D})) \coloneqq \int q(\theta) \log\left(\frac{q(\theta)}{p(\theta|\mathcal{D})}\right) d\theta, \tag{1.7}$$

where

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

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.

$$q^*(\theta) = \underset{q_\theta \in Q}{\operatorname{argmax}} \ \mathrm{ELBO}(q(\theta)). \tag{1.9}$$

Note that

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

Mean Field Variational Bayes: The most common Variational Inference algorithm is known as Mean Field Variational Bayes (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 parameter θ, i.e.,

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

MFVB has been adapted and developed over the last three decades, where it has been used in mixture modeling, probabilistic graphical modeling, and variable selection. We will introduce more about the algebra of MFVB in Chapter 2.

Advantages of MFVB: As mentioned above, one advantage of the MFVB is that it has a lower computation cost and so scales well with the dimensionality of the data, making it a desirable choice for high-dimensional datasets. For instance, if there exist a billion images that require to be fitted into a probabilistic machine learning model, then exact methods such as MCMC will be computationally demanding, while Variational Inference would have a higher chance to sacrifice accuracy with hundreds of times faster speed. Secondly, the time-efficiency of MFVB becomes another significant factor why it is popular, given the fact it only involves updating variational parameters iteratively until convergence, as opposed to MCMC which produces correlated samples that limit the ideal behavior of the MCMC algorithm.

Drawbacks of MFVB: Disadvantages of MFVB include inexact approximation results under some scenarios, although it could capture some marginal density information. For example, it is suggested by Blei et al. (2017), that the Variational Inference algorithm might underestimate the covariance between the parameters of interest if the inter-parameter correlation is strong. It tends to ignore the correlation between parameters, resulting in unideal

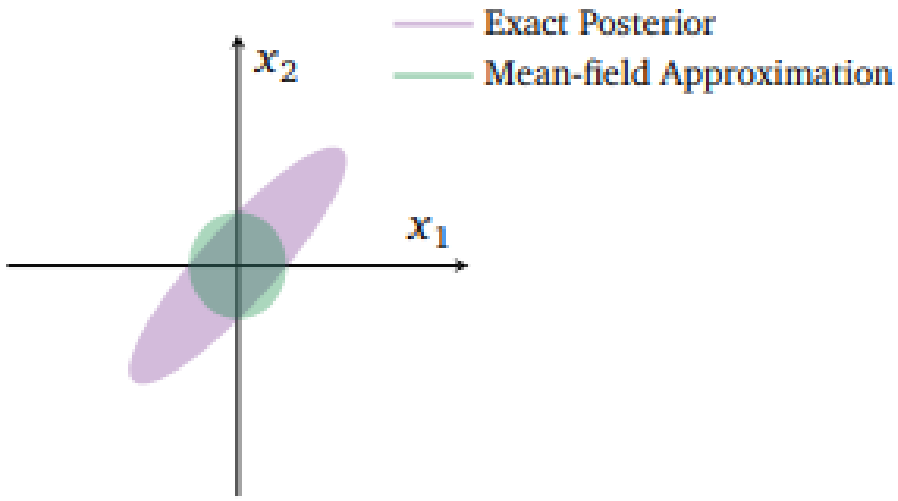


Figure 1.2: Visualization of Mean-Field Variational Approximation compared with exact posterior when the correlation is large

behavior as a result. Figure 1.2 demonstrates this phenomenon, the true overall posterior of x_2 and x_1 are correlated with an eclipse-shaped density, while a circled-shape mean-field approximation is established instead due to its product density family specification. 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 the small disadvantages of Variational Inference, the potential of variational approximation has not been fully explored by researchers, its ability to provide a 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 properties 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.

The aim of our approach is to use the posterior mean μ and covariance Σ for the re-

gression coefficient β_j for the j_{th} variable and use these to form an improved approximation for the marginal distributions. These approximated marginal distributions have the form of a novel distribution, a univariate Lasso distribution. We then use the mean and covariance of the Lasso distribtion to improve the VB Gaussian approximation of the posterior distribution for β. We have demonstrated that our algorithm's approximation accuracy surpasses that of several existing algorithms, 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 a hundred times faster time complexity.

Nevertheless, the drawback to this method is that if the initial global covariance matrix is diagonal it will remain diagonal after it is updated.

To remedy this issue, we have also purposed another algorithm based on marginal likelihood estimation by a bivariate Lasso distribution. Instead of updating the corresponding mean, and 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 the 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 both univariate and bivariate Lasso distribution to each of the marginal distributions, an improved estimate for global Gaussian approximation can be obtained as defined in Equation 1.12.

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

Finally, we will show our experiment result in Chapter 4 using various accuracy metrics.

1.2 Contributions

Our main contributions could be summarised as:

- Introduction of univariate and bivariate Lasso distributions.
- Derivation of properties for univariate Lasso distribution, such as the expectation, variance, cumulative density function, etc.

- Derivation of properties for bivariate Lasso distribution such as the expectation, variance, cumulative density function, etc.
- Implementation of univariate and bivariate Lasso distributions in R.
- Design of two new VI approaches based on local approximation via the univariate Lasso distribution and bivariate Lasso distribution respectively.
- Conduct of experiment to test the two algorithms under several benchmark datasets using several evaluation metrics for approximation accuracy.

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 approximations. 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 the variational algorithm in Chapter 3, followed by a comprehensive experiment for testing the effectiveness of the algorithm in Chapter 4. Chapter 5 is allocated for discussion and conclusion.

Chapter 2

Literature Review

2.1 Bayesian Inference Paradigm

Why the Bayesian paradigm?: 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. 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: Suppose θ is our model parameter of interest, $\mathcal{D}$ is a set of data, then $p(\theta)$ is known as a prior distribution, which offers pre-existing knowledge or information about θ. Then Bayesian inference approach stems from the Bayes rule, which is defined as Equation 2.1.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}. \tag{2.1}$$

The 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.

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta). \tag{2.2}$$

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, the regression coefficient is β, $||\beta||_1$ is the l_1 norm of beta, $||y - X\beta||_2$ is the l_2 norm: of residual value, the data matrix is X, the response variable is y. The estimation for lasso estimate $\hat{\beta}_{lasso}$ is defined by Equation 2.3.

$$\widehat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2^2, \quad s.t. \quad ||\beta||_1 \leq t, \quad \text{and} \quad t \geq 0. \tag{2.3}$$

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

$$\mathcal{L}(\beta, \lambda) = ||y - X\beta||_2^2 + \lambda||\beta||_1 - \lambda t, \quad \lambda \geq 0. \tag{2.4}$$

Since the objective function contains a quadratic term $||y - X\beta||_2^2$ with a linear term $\lambda||\beta||_1 - \lambda t)$, leading to a convex optimization problem. Due to the 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 Figure 2.1, where the squared constraint set is drawn, in addition to the contour line of the regression coefficient. The penalty term λ controls the strength of the penalization in Lasso regression. When the value of λ 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

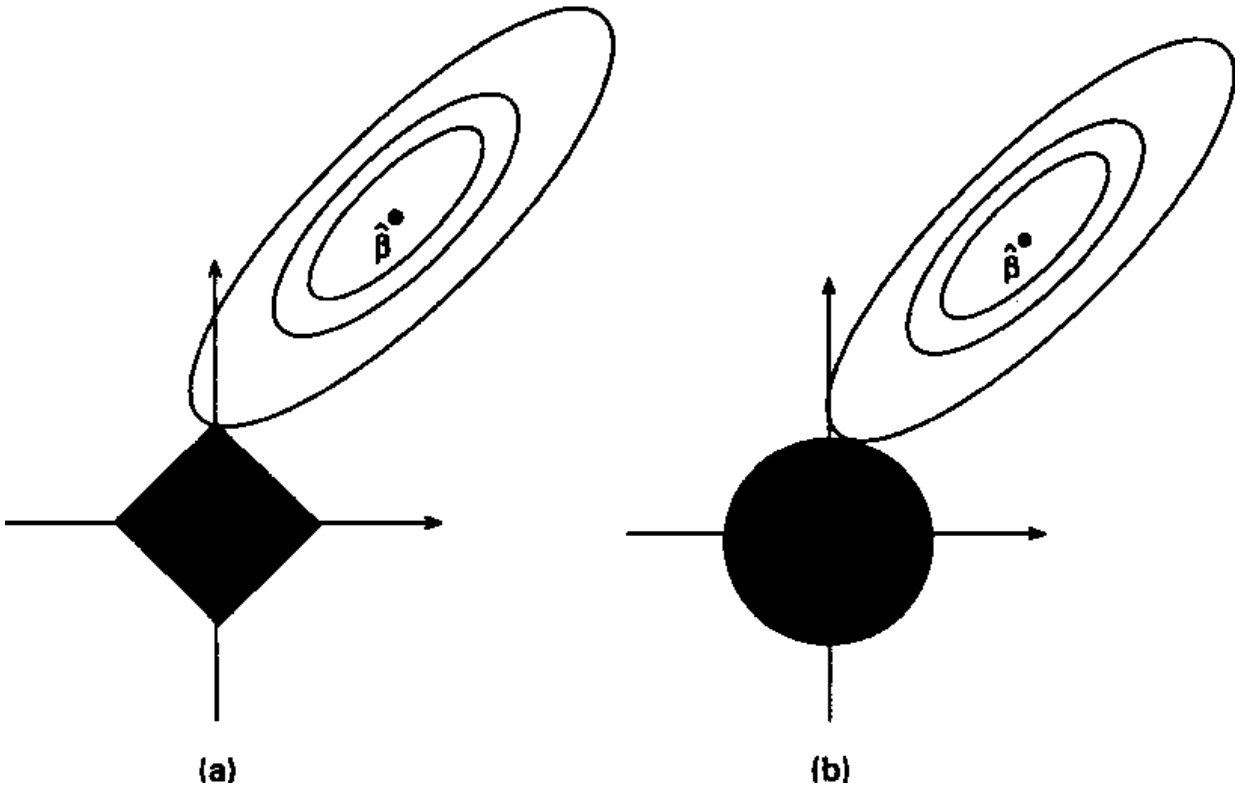


Figure 2.1: Graphical comparison between Lasso regression and ridge regression

and 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 the circular constraint set for β.

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 1.6 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.

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

For a given variance, the mode of posterior form in Equation 2.6 is consistent with the estimate of the Lasso equation in Equation 1.3, but it will hinder the Bayesian interpretation, inference, and variable selection since the Bayesian predictive distribution makes 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).

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

To remedy this issue, a conditional Laplacian prior from Equation 2.7 with respect to Equation 2.5 has been designed,

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

This prior ensures the unimodality of the posterior for β, and the current prior with respect to β, σ^2. The joint prior over β and σ^2 can be written as

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

which can result in the unimodal joint posterior distribution $\pi(\beta, \sigma^2|\tilde{y})$ of β and $\sigma^2 > 0$ under the new prior Equation 2.8, given an improper prior selection for $\pi(\sigma^2) = 1/\sigma^2$ and $\lambda \geq 0$.

To facilitate inference, an additional latent variable τ^2 is introduced. Then the Laplace distribution can be represented as a scale mixture of Gaussian densities for reformulation of conditional prior replacing Equation 2.7 with Equation 2.9, which can be regarded as corresponding weight assigned to each regression coefficient. If τ_j is near 0 then the corresponding regression coefficient will be shrunk towards zero accordingly.

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

Finally, the hierarchical Bayesian Lasso model functional form can be written as Equa-

tion 2.10.

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

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 full conditional distributions. In each iteration, each parameter of interest is sampled once from its full conditional distribution. After running the chain for sufficiently long, the samples from the Gibbs sampler will approximate the posterior distribution after discarding samples in the burn-in period. Notice that, the functional form of the full conditional distributions given any other parameter of interest has to be easily sampled from. In this study, we need to have the marginal distributions of $(\beta, \sigma^2, \tau^2)$ given by:

$$
\begin{aligned}
&p(\beta|\mathcal{D}, \sigma^2, \tau^2), \\
&p(\sigma^2|\mathcal{D}, \beta, \tau^2), \quad \text{and} \\
&p(\tau^2|\mathcal{D}, \beta, \sigma^2).
\end{aligned}
$$

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

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

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

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

Full conditional distribution of β: Note

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

Recognizing the term without β as constant, the conditional distribution of β can be simplified to

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

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

Full conditional distribution of σ^2:

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

Full conditional distribution of τ_j^2:

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

where GIG is generalized inverse gaussian distribution with parameters

$$a^* = 1, \quad b^* = \frac{\beta_j^2 \lambda^2}{\sigma^2}, \quad \text{and} \quad p = \frac{1}{2}.$$

Summary: The Gibbs sampler can be established by the following algorithm 1.

Algorithm 1 Gibbs sampler for the Bayesian Lasso

1: Given $\lambda^2 > 0, \tau^{(1)} = \mathbf{1_n}, \sigma^{2(1)} = 1, t = 1$

2: **while** $t \leq 10^5$ **do**

3: Sampling $\beta^{(t+1)} \sim \text{MVN}((X^T X + \lambda^2 A^{-1})^{-1} X^T y, (X^T X + \lambda^2 A^{-1})^{-1} \sigma^2)$

4: Sampling $\sigma^{2(t+1)} \sim IG\left(\frac{n}{2} + \frac{p}{2} + a, \frac{||y - X\beta||_2^2}{2} + \frac{\lambda^2 \sum_j \beta_j^2}{2\tau_j} + b\right)$

5: **for** j=1,...,p **do**

6: Sampling $\tau_j^{2(t+1)} \sim \text{GIG}\left(1, \frac{\beta_j^2 \lambda^2}{\sigma^2}, 1/2\right)$

7: $t \leftarrow t + 1$

8: return β, σ^2, τ^2

Automatic selection of the penalty parameter λ: A common choice of penalty parameter λ in the non-Bayesian paradigm involves a cross-validation approach, which is time-consuming and computationally challenging, especially for large datasets. Park and Casella (2008) propose a hyperprior on λ^2 of the form Equation 2.16, instead of λ to facilitate conjugacy. According to Park and Casella (2008), there are some additional constraints on choosing the prior on λ^2. Firstly, to avoid mixing issues, the prior distribution for λ^2 should reach zero asymptotically with a decent speed as λ^2 goes to infinity. Secondly, the density at maximum likelihood estimate should be assigned with enough probability density with an overall flat distribution.

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

The penalty parameter is the extent of penalization of the 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 τ_j, λ controls the shape of the Generalized Inverse Gaussian posterior distribution of τ_j as shown before.

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

and rate parameters θ and γ, respectively.

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

Thus, the posterior distribution of λ^2 is still following a Gamma distribution, with a shape parameter $p+\gamma$ and rate parameter $\sum_{j=1}^{p}\tau_j^2+\delta$. The λ^2 can be sampled by Equation 2.17, based on using an augmented Gibbs sampler.

2.4 Expectation Maximization

Even though the posterior distribution can be sampled by the Gibbs sampler in the last subsection, the sparsity nature of the Bayesian Lasso is not captured by the 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 the EM algorithm includes the inference of the parameters of the Gaussian Mixture model etc. The EM algorithm involves two main steps, which are an E-step and an M-step. Suppose Z is the set of latent variables, X is the set of the entire set of observed variables, and θ is a target parameter. The value t refers to the iteration number, $\log p(X,Z|\theta)$ 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-step

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 calculating 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:

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

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

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

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

2.4.2 Bayesian Expectation Maximization

The Bayesian EM algorithm incorporates the idea of EM algorithm and Bayesian inference for the 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

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D}). \tag{2.20}$$

We can then further expand Equation 2.20 to Equation 2.21

$$\log p(\theta|\mathcal{D}) = Q(\theta, \theta^{(old)}) + \mathrm{KL}(q||p(Z|X)) + \log p(\theta) - \log p(\mathcal{D}), \tag{2.21}$$

where $\mathrm{KL}(P||Q)$ is defined by Equation 1.8.

2.4.3 Bayesian EM for Bayesian Lasso model

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

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

E-step

Using the same notation as before, firstly, the complete log-likelihood can be written as Equation 2.22

$$\log(p(\theta_1, \theta_2, \tilde{y})) \propto -\frac{n+p}{2}\log(\sigma^2) - \frac{||\tilde{y} - X\beta||_2^2}{2\sigma^2} - \frac{1}{2\sigma^2}\beta^T E_{\theta_2|\tilde{y},\theta_1^{(t)}}[D_\tau]\beta - \frac{b}{\sigma^2}. \tag{2.22}$$

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

$$\begin{aligned} E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\log\, p(y, \theta_1, \theta_2)] = &-\frac{n}{2}\log(\sigma^{2(t)}) - \frac{||y - X\beta^{(t)}||_2^2}{2\sigma^{2(t)}} - E_{\theta_2|\tilde{y},\theta_1^{(t)}}\left[\sum_{j=1}^{p}\frac{\lambda^2\beta_j^2}{2\sigma^2\tau_j^2}\right] \\ &- (a+1)\log(\sigma^{2(t)}) - \frac{b}{\sigma^{2(t)}}. \end{aligned} \tag{2.23}$$

Next, we need to take the 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 θ_2, required formulation is $E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\frac{1}{\tau_j}]$. Given the fact that $\tau_j^2|\sigma^2, \tilde{y}, \beta \sim \mathrm{GIG}(1, \frac{\beta_j^2\lambda^2}{\sigma^2}, \frac{1}{2})$ and the special property of Generalized Inverse Gaussian distribution that if $X \sim \mathrm{GIG}(a, b, p)$, then $\frac{1}{X} \sim \mathrm{GIG}(b, a, -p)$, the distribution of $\frac{1}{\tau_j^2}$ can be rearranged to $\frac{1}{\tau_j^2}|\sigma^2, \tilde{y}, \beta \sim \mathrm{GIG}(\frac{\beta_j^2\lambda^2}{\sigma^2}, 1, -\frac{1}{2})$. However, taking expectations with

respect to generalized Gaussian distribution is still complicated and requires advanced mathematical operation and functional properties such as modified Bessel function. Thus, we can continue converting the distribution into an Inverse Gaussian distribution family, which renders $\frac{1}{\tau_j^2}|\sigma^2, \theta_1, \tilde{y} \sim \text{InverseGaussian}(b^{-\frac{1}{2}}, 1)$. Rewriting Equation 2.23 can lead to the final conditional expectation form can be written as:

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

where $E\left[\frac{1}{\tau_j^2}\right] = \frac{\sigma^{(t)}}{|\beta_j^{(t)}|\lambda}$. During the iteration, the $E\left[\frac{1}{\tau_j^2}\right]$ 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 setting them to 0 respectively provides a closed-form solution for updating observed parameters repeatedly:

$$\frac{\partial Q}{\partial \beta} = -\frac{1}{2\sigma^2}(-X^T y + 2X^T X\beta) - \frac{\lambda^2}{2\sigma^2}X^T X\beta = 0. \quad (2.25)$$

Rearranging the Equation 2.25, the updated formula for $\beta^{(t)}$ can be written as

$$\beta^{(t)} = (X^T X + \lambda^2 A)^{-1} X^T y, \quad \text{where} \quad A = \text{diag}\left(\frac{\sigma^{(t-1)}}{|\beta^{(t-1)}|\lambda}\right). \quad (2.26)$$

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

$$\frac{\partial Q}{\partial \sigma^2} = -\frac{(n + p + 2a + 2)}{2\sigma^2} + \frac{4b + 2||y - X\beta||_2^2 2 + \lambda^2(\beta^T A\beta)}{4\sigma^4} = 0. \quad (2.27)$$

Rearranging the Equation 2.27, the updated formula for $\sigma^{2(t)}$ can be written as:

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

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

Algorithm 2 Bayesian Expectation Maximization algorithm for the Bayesian Lasso

Given initial value $\theta_1^{(0)} = (\beta^{(0)}.\sigma^{2(0)})$, $\theta_2^0 = \mathbf{1}_p$, $t = 1$

while $||\theta_1^{(t)} - \theta_1^{(t-1)}||_2^2 < \epsilon$ **do**

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

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

  $A = \text{diag}\left(\frac{\sigma^{(t)}}{|\beta_j^{(t)}|\lambda}\right)$ ▷ Estimate expectation of hidden variable $E[\frac{1}{\tau_j^2}]$

  $t \leftarrow t + 1$

return $\theta_1^{(t)}$

2.5 Variational Inference

2.5.1 Introduction

One of the core challenges of statisticians in a Bayesian setting is to approximate complex probability density functions in a fast and efficient manner. VI serves as an effective alternative to the 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 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 minimizes the KL divergence to the exact posterior distribution as shown in Equation 1.7 for each subelement of parameter θ. The complexity of finding optimal distribution relies heavily on the complexity of Q. Nevertheless, due to the difficulty of computing marginal logarithm evidence $p(\mathcal{D}) = \int_\theta P(\mathcal{D}, \theta) d\theta$, as well as the implicit dependency nature of $p(\mathcal{D})$ to KL divergence as explained in Equation 2.29, additional conversion is required for further processing this

optimization system, which transforms Equation 1.7 to Equation 1.9.

$$\begin{aligned} \mathrm{KL}(q(\theta)||p(\theta|\mathcal{D})) &= \mathbb{E}_{q(\theta)}[\log q(\theta)] - \mathbb{E}_{q(\theta)}[\log p(\theta|\mathcal{D})] \\ &= \mathbb{E}_{q(\theta)}[\log q(\theta)] - \mathbb{E}_{q(\theta)}[\log p(\theta, \mathcal{D})] + \log p(\mathcal{D}). \end{aligned} \tag{2.29}$$

KL divergence

KL-divergence defined by Equation 2.29 is a distance metric for measuring the discrepancy of two probability distributions. This metric has several theoretical properties including non-negativity, and asymmetric property, i.e., $\mathrm{KL}(q||p) \neq \mathrm{KL}(p||q)$.

ELBO

The definition of the Evidence Lower bound is defined by Equation 1.10, which is equivalent to the negative KL divergence despite adding constant $p(\mathcal{D})$ with respect to $q(\theta)$. Apart from the equivalence of the optimization system, the explanation of why it is called "Evidence lower bound" can be shown in Equation 2.30

$$\log p(\mathcal{D}) = \mathrm{KL}(q(\theta)||p(\theta|\mathcal{D})) + \mathrm{ELBO}(q(\theta)). \tag{2.30}$$

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

2.5.3 Mean-Field Variational Family

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 the mean-field variational family is as described in Equation 1.11, assuming mutual independence of each target parameter θ_j, and the joint distribution is factorized as a product of individual distributions. This approach is known as mean-field Variational Bayes (MFVB) in the Bayesian paradigm. Finally, no further assumption has been arranged for each individual

distribution $q(\theta_i)$. The two-dimensional visualization of mean-field approximation density is shown in Figure 1.2.

2.5.4 Coordinate Ascent Variational Inference (CAVI)

The 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. 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 $q(\theta)$: optimizing $q_i(\theta_i)$, while maintaining other factors of distribution unchanged, $q_{-i}(\theta_{-i})$, enforcing the final distribution can achieve a local optimum of the ELBO. Small changes regarding the stopping criterion have also been proposed, where the stopping criterion has been transformed from the difference of function value into the 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 the MFVB algorithm in our setting.

Derivation

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

$$\begin{aligned}
\text{ELBO}(q(\theta)) &= \int q(\theta) \log \left(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)} \right) d\theta \\
&= \int \prod_i q_i(\theta_i) \left[\log p(\mathcal{D}, \theta) - \sum_i \log q_i(\theta_i) \right] d\theta \\
&= \int q_j(\theta_j) \left[\int \log p(\mathcal{D}, \theta) \prod_{i \neq j} q_i d\theta_j \right] d\theta_j - \int q_j(\theta_j) \log q_j(\theta_j) d\theta_j + const \\
&= \int q_j(\theta_j) \log \tilde{p}(\mathcal{D}, \theta_j) - \int q_j(\theta_j) \log q_j(\theta_j) d\theta_j,
\end{aligned} \tag{2.31}$$

where $\log \tilde{p}(\mathcal{D}, \theta_j) = \mathbb{E}_{i \neq j} \log p(\mathcal{D}, \theta) + const$, and $\mathbb{E}_{i \neq j} \log p(\mathcal{D}, \theta) = \int \log p(\mathcal{D}, \theta) \prod_{i \neq j} q_j(\theta_j) d\theta_j$ assuming optimizing the jth posterior parameter θ_j. Identifying the Equation 2.31 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,

$$\log[q_j^*(\theta_j)] = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)] + const. \tag{2.32}$$

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

$$q_j^*(\theta_j) = \frac{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)]}{\int \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta) d\theta_j}} \propto \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)]. \tag{2.33}$$

The algorithm then iteratively replaces $q_j(\theta_j)$ using the current estimation for all of the other factors, and convergence criteria are broken when differences of ELBOs change less than a predefined tolerance. Theoretically, it is proved that the convergence of the 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. Similar to MFVB, CAVI has a similar updating rule, while the only modification is that it converts the stopping criterion from using the consecutive difference of ELBO values instead of the difference of two consecutive parameters θ values.

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,...,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 $q(\theta)$

2.5.5 MFVB for Bayesian Lasso

We now propose the MFVB algorithm for the Bayesian Lasso problem. Our goal is to attain Bayesian Lasso posterior approximation in a faster and more efficient manner than that of the Gibbs sampler under the framework of VI. Note in this section, slight modification has been imposed to the definition where the parameters of a_j distribution to be fixed, therefore, no hyperpriors are required. This also abandons the dependence of λ to τ to simplify the derivation and algorithm. The assumption of tractable distribution family Q can be written as based on MFVB:

$$q(\theta) = q(\beta)q(\sigma^2)\prod_{i=1}^{p} q(a_j). \tag{2.34}$$

Derivation

In addition, we would like to derive the optimal variational distribution for each parameter: i.e., $q^*_\beta(\beta)$, $q^*_{\sigma^2}(\sigma^2)$ according to Equation 2.33. The following derivation is based on the conditional posterior distributional form and expectation with respect to the corresponding standard distribution β, $q^*_\beta(\beta)$ can be derived by:

$$\begin{aligned} q^*_\beta(\beta) &\propto \exp\left[\mathbb{E}_{-\beta} \log p(\beta, y, \mathbf{a}, \sigma^2)\right] \\ &\sim MVN\left[(X^TX + \lambda^2 A)^{-1}X^T y, (X^TX + \lambda^2 A)^{-1}\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)^{-1}\right]. \end{aligned} \tag{2.35}$$

where $A = \text{diag}(\mathbf{a})$, $\mathbb{E}_{\sigma^2}(\frac{1}{\sigma^2})^{-1} = \frac{\tilde{b}}{\tilde{a}}$.

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

corresponding standard distribution.

$$
\begin{aligned}
q^*_{\sigma^2}(\sigma^2) &\propto \exp\left[\mathbb{E}_{-\sigma^2} \log p(\beta, y, \sigma^2, \mathbf{a})\right] \\
&\sim \text{InverseGamma}\left(\frac{n+p}{2}, \frac{1}{2}\mathbb{E}_q||y - X\beta||^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\mathbb{E}_q(\beta^T A\beta)\right).
\end{aligned} \tag{2.36}
$$

Therefore, the update procedure of MFVB for the Bayesian Lasso can be written as follows:

- Update Procedure of MFVB for the Bayesian Lasso
  - $Q = X^T X + \lambda^2 A$, where $A = \text{diag}(\mathbf{a})$.
  - The update for beta leads to

$$\widetilde{\mu} = Q^{-1}X^T y \quad \text{and} \quad \widetilde{\Sigma} = \mathbb{E}_q\left[\frac{1}{\sigma^2}\right]^{-1} Q^{-1}.$$

  - The update for σ^2 leads to

$$\widetilde{a} = \frac{n+p}{2}, \quad \text{and} \quad \widetilde{b} = \frac{E_q||y - X\beta||^2 + \lambda^2\mathbb{E}_q[\beta^T A\beta]}{2}.$$

The aforementioned update procedure for MFVB estimates the parameters of the posterior distributions of β and σ^2 with a descent speed, even though the posterior variance is usually underestimated due to the mean-field restriction for the approximated density.

Chapter 3

Method

3.1 Introduction

The main idea behind the method introduced in this study is to refine the MFVB posterior parameter estimate for the regression coefficients β (represented by $\tilde{\mu}$ and $\tilde{\Sigma}$). By assuming a mean field variational family $q(\theta) = \prod_i q(\theta_i)$ and using Gaussian approximation, denoted as $q^*(\theta) \sim N(\tilde{\mu}, \tilde{\Sigma})$, the aim is to approximate the global Bayesian Lasso posterior $p(\theta|\mathcal{D})$ while incorporating local parameter information. It is observed that the marginal likelihood approximately follows a lasso distribution when a Laplace prior is assigned. Consequently, the objective is to determine the mathematical expression of the local mean parameter μ_j^* and local variance Σ_{jj}^* from the Lasso distribution at a local level, as well as the global mean $\tilde{\mu}$ and global variance $\tilde{\Sigma}$ of $q^*(\theta)$ for Gaussian approximation. This can be achieved through iterative updates that correct the global parameter estimate based on the local parameter expression for each β_j.

3.2 Basic Setting for the Bayesian Lasso Problem

Based on MFVB approach, we can approximate the posterior distribution as follows:

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

We can divide up the set of parameters of interest $\theta = \beta_1, \ldots, \beta_p, \sigma^2$ into two parts θ_1: β_j current variable and θ_2 : β_{-j} , other variables. The marginal log-likelihood of θ_1 can be divided up into the ELBO part and the KL divergence part as follows:

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

Since the KL divergence is greater than 0, the marginal log-likelihood of θ_1 and $\mathcal{D}$ has a more tractable ELBO:

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

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

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

In Bayesian Lasso $\theta = (\beta, \sigma^2)$, however, in this study we do not discuss the update of σ^2. Thus, the conditional distribution $q(\beta_{-j}|\beta_j)$ for any jth variable can be derived by $q(\beta_{-j}|\beta_j) \propto q(\beta)$, resulting another multivariate normal distribution with dimension of $p-1$ as shown in Equation 3.4

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

Based on Equation 3.1, Equation 3.4, and Equation 3.1, the result from Equation 3.4, the estimated marginal log likelihood for each β_j after taking expectation with respect to $q(\theta)$ is:

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

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 σ^2, μ, Σ are posterior parameters for β. One of the key improvements of our method compared to MFVB is that it effectively captures the correlation between θ_2 and θ_1. Note that in Equation 3.5, the expectation with respect to $q(\theta_2|\theta_1)$ is performed. To illustrate this, let's consider the case of MFVB, where an expectation is taken with respect to $q(\theta)$ under the assumption of independence among the parameters θ. Furthermore, it is evident that our method reduces to the MFVB approach when $t = 0$, resulting in $s = \mu_{-j}$. In the following subsection, we introduce the Lasso distribution.

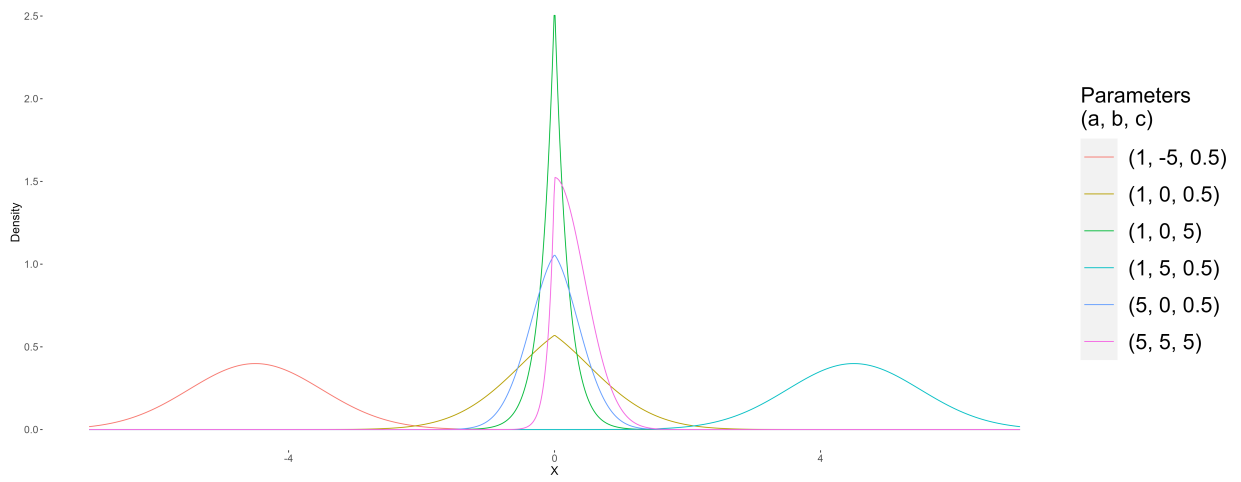


Figure 3.1: Visualization of the univariate Lasso distribution PDF for different parameter settings

3.3 Lasso distribution

Figure 3.1 demonstrates the shape, scale, and location of the univaraite Lasso distribution with different parameter settings which are shown in the legends. From the yellow and dark-blue line, we can observe that parameter A control the size of the curvature of the tuning point, larger a implies a smoother turning point. From the red line, yellow line, and sky-blue line, we can observe that changing in b changes the location of the curve. From the yellow line and green line, c controls the sharpness of the curve, a larger c implies distribution with smaller variance and a discontinuous derivative.

3.3.1 Univariate Lasso Distribution

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

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

where $a \geq 0, b \in \mathbb{R}, c \geq 0$, there are also certain restrictions to certain parameter settings:

- a and c cannot be 0 simultaneously.
- When $a = 0$, Lasso distribution will become an asymmetric Laplace distribution.
- When $c = 0$, Lasso distribution will become a normal distribution.

The probability density function of univariate Lasso distribution can be divided up into four components:

- A normalization constant Z, to enable the integration of the probability density function to be 1.
- A quadratic term ax^2 to control the curvature of the curve.
- A linear term bx to control the location of the curve.
- An absolute term $c|x|$ to control the sharpness of the discontinuous derivative.

Certain properties of a lasso distribution such as normalizing constant Z, expectation $\mathbb{E}(x)$, second moment $\mathbb{E}(x^2)$ and variance $\mathbb{V}(x)$ are necessary for our algorithm which is computed in the following.

Basic Property

Derivation of normalizing constant

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

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

Derivation of Moments

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

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

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

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

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. The function $\zeta_1(x)$ represents the inverse Mills ratio, which requires careful consideration. Hence,

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

Equation 3.5 can be matched to a univaraite 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.

In the following subsection we introduce bivariate Lasso distribution and its features.

3.3.2 Bivariate Lasso Distribution

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

$$p(x) = Z^{-1} \exp\left(-\frac{1}{2}x^T Ax + b^T x - c||x||_1\right), \tag{3.7}$$

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

Similarly, the probability density function of bivariate Lasso distribution can be divided up into four components:

- A normalization constant Z, to enable the integration of the probability density function to be 1.
- A quadratic term $x^T Ax$ to control the curvature of the curve.
- A linear term bx to control the location of the curve.
- $c||x||_1$ as l_1 norm of x to control the sharpness of the discontinuous derivative.

Similar to the univariate Lasso distribution, closed form normalizing constant is useful for constructing a valid probability density function, while expectation and covariance matrix are essential for the bivariate local global algorithm to calculate the local approximated mean and covariance.

Derivation of Normalizing Constant

The normalizing constant can be calculated via integrate the unnormalized probability density function.

$$
\begin{aligned}
Z(a,b,c) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}x^TAx + b^Tx - c1^T|x|_1\right] dx \\
&= \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}x^TAx + (b^T - c1^T)x\right] dx \\
&\quad + \int_{0}^{\infty}\int_{-\infty}^{0} \exp\left[-\tfrac{1}{2}x^TAx + (b^T - c[1,-1]^T)x\right] dx \\
&\quad + \int_{-\infty}^{0}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}x^TAx + (b^T - c[-1,1]^T)x\right] dx \\
&\quad + \int_{-\infty}^{0}\int_{-\infty}^{0} \exp\left[-\tfrac{1}{2}x^TAx + (b^T + c1^T)x\right] dx \\
&= \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}x^TAx + (b^T - c1^T)x\right] dx \\
&\quad + \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}x^TA^*x + (b_1 - c, -b_2 - c)^Tx\right] dx \\
&\quad + \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}x^TA^*x + (-b_1 - c, b_2 - c)^Tx\right] dx \\
&\quad + \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}x^TAx - (b^T + c1^T)x\right] dx \\
&= \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \tfrac{(A\mu_1)^T\Sigma_1(A\mu_1)]}{2}\right] dx \\
&\quad + \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) + \tfrac{(A^*\mu_2)^T\Sigma_2(A^*\mu_2)]}{2}\right] dx \\
&\quad + \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}(x-\mu_3)^T\Sigma_2^{-1}(x-\mu_3) + \tfrac{(A^*\mu_3)^T\Sigma_2(A^*\mu_3)]}{2}\right] dx \\
&\quad + \int_{0}^{\infty}\int_{0}^{\infty} \exp\left[-\tfrac{1}{2}(x-\mu_4)^T\Sigma_1^{-1}(x-\mu_4) + \tfrac{(A\mu_4)^T\Sigma_1(A\mu_4)]}{2}\right] dx \\
&= 2\pi|\Sigma_1|^{\frac{1}{2}}[\exp\left[\tfrac{(A\mu_1)^T\Sigma_1(A\mu_1)}{2}\right]\int_{0}^{\infty}\int_{0}^{\infty}\phi_2(x;\mu_1,\Sigma_1)dx \\
&\quad + \exp\left[\tfrac{(A\mu_4)^T\Sigma_1(A\mu_4)}{2}\right]\int_{0}^{\infty}\int_{0}^{\infty}\phi_2(x;\mu_4,\Sigma_1)dx] \\
&\quad + 2\pi|\Sigma_2|^{\frac{1}{2}}[\exp\left[\tfrac{(A^*\mu_2)^T\Sigma_2(A^*\mu_2)]}{2}\right]\int_{0}^{\infty}\int_{0}^{\infty}\phi_2(x;\mu_2,\Sigma_2)dx \\
&\quad + \exp\left[\tfrac{(A^*\mu_3)^T\Sigma_2(A^*\mu_3)}{2}\right]\int_{0}^{\infty}\int_{0}^{\infty}\phi_2(x;\mu_3,\Sigma_2)dx] \\
&= |\Sigma_1|\left(\tfrac{\int_0^\infty\int_0^\infty\phi_2(x;\mu_1,\Sigma_1)dx}{\phi_2(A\mu_1,\Sigma_1^{-1})} + \tfrac{\int_0^\infty\int_0^\infty\phi_2(x;\mu_4,\Sigma_1)dx}{\phi_2(A\mu_4,\Sigma_1^{-1})}\right) \\
&\quad + |\Sigma_2|\left(\tfrac{\int_0^\infty\int_0^\infty\phi_2(x;\mu_2,\Sigma_2)dx}{\phi_2(A^*\mu_2,\Sigma_2^{-1})} + \tfrac{\int_0^\infty\int_0^\infty\phi_2(x;\mu_3,\Sigma_2)dx}{\phi_2(A^*\mu_3,\Sigma_2^{-1})}\right)
\end{aligned}
$$

where $\mu_1 = A^{-1}(b - c1)^T$, $\mu_2 = A^{*-1}(b_1 - c, -b_2 - c)^T$, $\mu_3 = A^{*-1}(-b_1 - c, b_2 - c)^T$, $\mu_4 = A^{-1}(-b - c1^T)^T$ and $\Sigma_1 = A^{-1}$, $\Sigma_2 = A^{*-1}$, $A^* = A \odot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, here $\odot$ denotes the element-wise product.

Derivation of Expectation

The expectation of lasso distribution can be derived by the expectation property: $\mathbb{E}[x] = \int x f(x) dx$.

$$
\begin{aligned}
\mathbb{E}[x] &= Z^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \odot \exp\left[-\tfrac{1}{2} x^T A x + b^T x - c 1^T ||x||_1\right] dx \\
&= Z^{-1} \int_{0}^{\infty} \int_{0}^{\infty} x \odot \exp\left[-\tfrac{1}{2} x^T A x + (b^T - c 1^T) x\right] dx \\
&+ \int_{0}^{\infty} \int_{0}^{\infty} [1,-1]^T \odot x \odot \exp\left[-\tfrac{1}{2} x^T A^* \odot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} x + (b_1 - c, -b_2 - c)^T x\right] dx \\
&+ \int_{0}^{\infty} \int_{0}^{\infty} [-1,1]^T \odot x \odot \exp\left[-\tfrac{1}{2} x^T A^* \odot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} x + (-b_1 - c, b_2 - c)^T x\right] dx \\
&- \int_{0}^{\infty} \int_{0}^{\infty} x \odot \exp\left[-\tfrac{1}{2} x^T A x + (b^T + c 1^T) x\right] dx \\
&= Z^{-1} [|\Sigma_1| \left(\tfrac{\int_0^\infty \int_0^\infty x \odot \phi_2(x; \mu_1, \Sigma_1) dx}{\phi_2(A\mu_1, \Sigma_1^{-1})} - \tfrac{\int_0^\infty \int_0^\infty x \odot \phi_2(x; \mu_4, \Sigma_1) dx}{\phi_2(A\mu_4, \Sigma_1^{-1}))} \right) \\
&+ |\Sigma_2| \left([1,-1]^T \tfrac{\int_0^\infty \int_0^\infty x \odot \phi_2(x; \mu_2, \Sigma_2) dx}{\phi_2(A^* \mu_2, \Sigma_2^{-1})} + [-1,1]^T \tfrac{\int_0^\infty \int_0^\infty x \odot \phi_2(x; \mu_3, \Sigma_2) dx}{\phi_2(A^* \mu_3, \Sigma_2^{-1})}) \right)] \\
&= Z^{-1} [|\Sigma_1| (\tfrac{\mathbb{E}[\mathbf{A}] \int_0^\infty \int_0^\infty \phi_2(x; \mu_1, \Sigma_1) dx}{\phi_2(A\mu_1, \Sigma_1^{-1})} - \tfrac{\mathbb{E}[D] \int_0^\infty \int_0^\infty \phi_2(x; \mu_4, \Sigma_1) dx}{\phi_2(A\mu_4, \Sigma_1^{-1})}) \\
&+ |\Sigma_2| ([1,-1]^T \tfrac{\mathbb{E}[B] \int_0^\infty \int_0^\infty \phi_2(x; \mu_2, \Sigma_2) dx}{\phi_2(A^* \mu_2, \Sigma_2^{-1})} + [-1,1]^T \tfrac{\mathbb{E}[C] \int_0^\infty \int_0^\infty \phi_2(x; \mu_3, \Sigma_2) dx}{\phi_2(A^* \mu_3, \Sigma_2^{-1})})]
\end{aligned}
$$

where $\mu_1 = A^{-1}(b - c1)^T$, $\mu_2 = A^{*-1}(b_1 - c, -b_2 - c)^T$, $\mu_3 = A^{*-1}(-b_1 - c, b_2 - c)^T$, $\mu_4 = A^{-1}(-b - c1^T)^T$, $\Sigma_1 = A^{-1}$, $\Sigma_2 = A^{*-1}$, and $A^* = A \odot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, here $\odot$ denotes the element-wise product. $\mathbf{A} \sim MTN_+(\mu_1, \Sigma_1)$, $B \sim MTN_+(\mu_2, \Sigma_2)$, $C \sim MTN_+(\mu_3, \Sigma_2)$, $D \sim MTN_+(\mu_4, \Sigma_1)$ denote the multivariate positively truncated normal distribution. Note $\mathbf{A}$ represents a multivariate positively truncated normal random variable, while A represents the bivariate Lasso parameter.

Derivation of Covariance Matrix

The second moment of lasso distribution can be derived by the expectation property: $\mathbb{E}[xx^T] = \int xx^T f(x) dx$.

$$\mathrm{Cov}(x) = \mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T$$

$$
\begin{aligned}
\mathbb{E}[xx^T] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xx^T \odot \exp\left(-\tfrac{1}{2}x^TAx + b^Tx - c1^T||x||_1\right)dx \\
&= Z^{-1}2\pi|\Sigma|^{\frac{1}{2}}[\exp\left(\tfrac{(A\mu_1)^T\Sigma(A\mu_1)]}{2}\right)\int_0^{\infty}\int_0^{\infty} xx^T \odot \phi_2(x;\mu_1,\Sigma_1)dx \\
&\quad + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \odot \exp\left(\tfrac{(A^*\mu_2)^T\Sigma(A^*\mu_2)]}{2}\right)\int_0^{\infty}\int_0^{\infty} xx^T \odot \phi_2(x;\mu_2,\Sigma_2)dx \\
&\quad + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \odot \exp\left(\tfrac{(A^*\mu_3)^T\Sigma(A^*\mu_3)]}{2}\right)\int_0^{\infty}\int_0^{\infty} xx^T \odot \phi_2(x;\mu_3,\Sigma_2)dx \\
&\quad - \exp\left(\tfrac{(A\mu_4)^T\Sigma(A\mu_4)]}{2}\right)\int_0^{\infty}\int_0^{\infty} xx^T \odot \phi_2(x;\mu_4,\Sigma_1)dx] \\
&= Z^{-1}|\Sigma|[\tfrac{\int_0^{\infty}\int_0^{\infty} xx^T\odot\phi_2(x;\mu_1,\Sigma)dx}{\phi_2(A\mu_1,\Sigma)} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \odot \tfrac{\int_0^{\infty}\int_0^{\infty} xx^T\odot\phi_2(x;\mu_2,\Sigma)dx}{\phi_2(A\mu_2,\Sigma)} \\
&\quad + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \odot \tfrac{\int_0^{\infty}\int_0^{\infty} xx^T\odot\phi_2(x;\mu_3,\Sigma_1)dx}{\phi_2(A\mu_3,\Sigma_1)} - \tfrac{\int_0^{\infty}\int_0^{\infty} xx^T\odot\phi_2(x;\mu_4,\Sigma_1)dx}{\phi_2(A\mu_4,\Sigma_1)}] \\
&= Z^{-1}[|\Sigma_1|\left(\tfrac{\mathbb{E}[\mathbf{A}\mathbf{A}^T]\int_0^{\infty}\int_0^{\infty}\phi_2(x;\mu_1,\Sigma_1)dx}{\phi_2(A\mu_1,\Sigma_1^{-1})} + \tfrac{\mathbb{E}[DD^T]\int_0^{\infty}\int_0^{\infty}\phi_2(x;\mu_4,\Sigma_1)dx}{\phi_2(A\mu_4,\Sigma_1^{-1})}\right) \\
&\quad + |\Sigma_2|\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \odot \tfrac{\mathbb{E}[BB^T]\int_0^{\infty}\int_0^{\infty}\phi_2(x;\mu_2,\Sigma_2)dx}{\phi_2(A^*\mu_2,\Sigma_2^{-1})} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \odot \tfrac{\mathbb{E}[CC^T]\int_0^{\infty}\int_0^{\infty}\phi_2(x;\mu_3,\Sigma_2)dx}{\phi_2(A^*\mu_3,\Sigma_2^{-1})}\right)]
\end{aligned}
$$

where $\mu_1 = A^{-1}(b-c1)^T$, $\mu_2 = A^{*-1}(b_1-c,-b_2-c)^T$, $\mu_3 = A^{*-1}(-b_1-c,b_2-c)^T$, $\mu_4 = A^{-1}(-b-c1^T)^T$ $\Sigma_1 = A^{-1}$, $\Sigma_2 = A^{*-1}$, and $A^* = A \odot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, here $\odot$ denotes the element-wise product. $\mathbf{A} \sim MTN_+(\mu_1,\Sigma_1)$, $B \sim MTN_+(\mu_2,\Sigma_2)$, $C \sim MTN_+(\mu_3,\Sigma_2)$, and $D \sim MTN_+(\mu_4,\Sigma_1)$ denote the multivariate positively truncated normal distributions. In addition, the second moment $\mathbb{E}[AA^T]$ and $\mathbb{E}[BB^T]$ can be derived similarly from the variance property in multivariate function:

$$\mathbb{E}[AA^T] = \mathrm{Cov}(A) - \mathbb{E}[A]\mathbb{E}[A]^T$$

Derivation of marginal distribution

The marginal distribution of the bivariate Lasso distribution is useful for evaluating the accuracy of l_1 norm approximation, allowing for comparison and visualization of the approximated density. The marginal distributions of x_1 and x_2 can be derived by $f(x_1) = \int_{-\infty}^{\infty} f(x_1,x_2)dx_2$ and $f(x_2) = \int_{-\infty}^{\infty} f(x_1,x_2)dx_1$, respectively. In the bivariate Lasso case,

we have

$$
\begin{aligned}
f(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2)dx_2 \\
&= Z^{-1} \int_{-\infty}^{\infty} \exp(-\tfrac{1}{2}x^T Ax + b^T x - c||x||_1))dx_2 \\
&= Z^{-1} \int_{-\infty}^{\infty} \exp(-0.5a_{11}x_1^2 + b_1 x_1 - c|x_1|)dx_2 \\
&\int_{-\infty}^{\infty} \exp(-\tfrac{1}{2}[(a_{12}+a_{21})x_1x_2 + a_{22}x_2^2] + b_2x_2 - c|x_2|]dx_2 \\
&= k \int_{-\infty}^{\infty} \exp[-(0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{22}x_2^2 + b_2x_2 - c|x_2|]dx_2 \\
&= k[\int_0^{\infty} \exp[-0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{22}x_2^2 + (b_2-c)x_2]dx_2 \\
&+ \int_{-\infty}^{0} \exp[-0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{22}x_2^2 + (b_2+c)x_2]dx_2 \\
&= k[\int_0^{\infty} \exp[-0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{22}x_2^2 + (b_2-c)x_2]dx_2 \\
&+ \int_0^{\infty} \exp[0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{22}x_2^2 - (b_2+c)x_2]dx_2] \\
&= k[\int_0^{\infty} \exp[-\tfrac{(x_2-\mu_1)^2}{2\sigma^2} + \tfrac{\mu_1^2}{2\sigma^2}]dx_2] + \int_0^{\infty} \exp[-\tfrac{(x_2-\mu_2)^2}{2\sigma^2} + \tfrac{\mu_2^2}{2\sigma^2}]dx_2] \\
&= k\sigma \left[\tfrac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} + \tfrac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)}\right]
\end{aligned}
$$

where $\mu_1 = (-\frac{a_{12}+a_{21}}{2a_{22}}x_1 + \frac{b_2-c}{a_{22}})$, $\mu_2 = (\frac{a_{12}+a_{21}}{2a_{22}}x_1 - \frac{b_2+c}{a_{22}})$, $\sigma^2 = 1/a_{22}$, $k = Z^{-1}\exp(-0.5a_{11}x_1^2 + b_1x_1 - c|x_1|)$. Also, the marginal distribution of x_2 is

$$
\begin{aligned}
f(x_2) &= \int_{-\infty}^{\infty} f(x_1, x_2)dx_1 \\
&= Z^{-1} \int_{-\infty}^{\infty} \exp(-\tfrac{1}{2}x^T Ax + b^T x - c||x||_1))dx_1 \\
&= Z^{-1} \int_{-\infty}^{\infty} \exp(-0.5a_{22}x_2^2 + b_2 x_2 - c|x_2|)dx_1 \\
&\int_{-\infty}^{\infty} \exp(-\tfrac{1}{2}[a_{12}a_{21}x_1x_2 + a_{11}x_1^2] + b_1x_1 - c|x_1|]dx_1 \\
&= k \int_{-\infty}^{\infty} \exp[-0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{11}x_1^2 + b_1x_1 - c|x_1|]dx_1 \\
&= k[\int_0^{\infty} \exp[-0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{11}x_1^2 + (b_1-c)x_1]dx_1 \\
&+ \int_{-\infty}^{0} \exp[-0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{11}x_1^2 + (b_1+c)x_1]dx_1] \\
&= k[\int_0^{\infty} \exp[-0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{11}x_1^2 + (b_1-c)x_1]dx_1 \\
&+ \int_0^{\infty} \exp[0.5(a_{12}+a_{21})x_1x_2 - 0.5a_{11}x_1^2 - (b_1+c)x_1]dx_1] \\
&= k[\int_0^{\infty} \exp[-\tfrac{(x_2-\mu_1)^2}{2\sigma^2} + \tfrac{\mu_1^2}{2\sigma^2}]dx_1] + \int_0^{\infty} \exp[-\tfrac{(x_2-\mu_2)^2}{2\sigma^2} + \tfrac{\mu_2^2}{2\sigma^2}]dx_1] \\
&= k\sigma \left[\tfrac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} + \tfrac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)}\right]
\end{aligned}
$$

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

3.4 Local-Global Algorithm

3.4.1 Univariate local global algorithm

The global approximation is a marginal normal approximation to the conditional distribution $p(\beta_j|\mathcal{D})$: $q^*(\beta_j) \approx N(\mu_j^*, \Sigma_{jj}^*)$. The local approximation of mean μ_j^* and variance Σ_{jj}^* can be obtained by the expression of Lasso expectation formula with given Lasso parameters a, b, and c. Based on Equation 3.5, we can show that the distribution of β_j given $\mathcal{D}$ is Lasso distribution, $p(\beta_j|\mathcal{D}) \propto p(\beta_j, \mathcal{D}) \sim Lasso\left(\frac{\tilde{a}}{\tilde{b}}(y - x_{-j}s), \frac{\tilde{a}}{2\tilde{b}}(X_j^T X_j + X_j^T X_{-j}t), \frac{\lambda\Gamma(\tilde{a}+1/2)}{\Gamma(\tilde{a})\sqrt{\tilde{b}}}\right)$, where $s = \mu_{-j} - \Sigma_{-j,j}\Sigma_{j,j}^{-1}\mu_j$, $t = \Sigma_{-j,j}\Sigma_{j,j}^{-1}$ are predefined value from original data, $\tilde{a}$ and $\tilde{b}$ are from MFVB output. Matching moments of $p(\beta_j|\mathcal{D})$ with a normal distribution and reforming the joint distribution we obtain Equation 3.8:

$$q^*(\beta) = q(\beta_{-j}|\beta_j)\phi(\beta_j; \mu_j^*, \Sigma_{jj}^*). \tag{3.8}$$

Since both $q(\beta_{-j}|\beta_j)$ and $q(\beta_j) = \phi(\beta_j; \mu_j^*, \Sigma_{jj}^*)$ are normal distributions, the joint distribution is also normal distribution. Suppose the mean and covariance of the joint distribution are $\widetilde{\mu}$ and $\widetilde{\Sigma}$ respectively, then $\widetilde{\mu}_1 = \mu_1^*$, and $\widetilde{\Sigma}_{11} = \Sigma_{11}^*$. The derivation of $\tilde{\mu}$ is shown below based on the law of total expectation. Denote $\theta_1 = \beta_j$, $\theta_2 = \beta_{-j}$ as in Chapter 1. It is known that $q(\theta_2|\theta_1)$ and $q(\theta_1)$ are Gaussian since $q(\theta)$ is Gaussian. Hence, their joint distribution is also Gaussian. The mean and variance of the joint distribution of θ_1 are μ_1^* and Σ_{11}^*. The rest of the derivation is to determine $\widetilde{\mu}_2$, $\widetilde{\Sigma}_{22}$ and $\widetilde{\Sigma}_{12}$. Denote the actual $q(\theta) \sim N(\mu, \Sigma)$, which approximate $p(\theta|\mathcal{D})$

We have

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

Similarly,

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

Lastly,

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

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

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

Based on the law of total variance $\tilde{\Sigma}$ is:

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

Hence, $q^*(\beta) = N(\widetilde{\mu}, \widetilde{\Sigma})$ at the time when updating variable β_j is:

$$\widetilde{\mu} = \begin{bmatrix} \mu_j^* \\ \mu_{-j} + \Sigma_{-j,j}\Sigma_{jj}^{-1}\left(\mu_j^* - \mu_j\right) \end{bmatrix} \tag{3.9}$$

and

$$\widetilde{\Sigma} = \begin{bmatrix} \Sigma_{jj}^{*} & \Sigma_{jj}^{*}\Sigma_{jj}^{-1}\Sigma_{j-j} \\ \Sigma_{-jj}\Sigma_{jj}^{-1}\Sigma_{jj}^{*} & \Sigma_{-j-j} + \Sigma_{-jj}\Sigma_{jj}^{-1}(\Sigma_{jj}^{*} - \Sigma_{jj})\Sigma_{jj}^{-1}\Sigma_{j-j} \end{bmatrix}. \tag{3.10}$$

In practise, since we do not know the actual μ and Σ, they will be replaced by initial $\widetilde{\mu}$ and initial $\widetilde{\Sigma}$ respectively, as denoted in the below algorithm.

Algorithm 4 Univariate Local-Global Algorithm

Input: X, y, parameters $(\tilde{a}, \tilde{b}, \tilde{\mu}, \tilde{\Sigma})$ from MFVB , λ

while $\tilde{\mu}$ is changing less than ϵ **do**

  for $j = 1$ to p **do**

    $a = \frac{\tilde{a}}{\tilde{b}}(X^T X)_{j,j} + (X^T X)_{j,-j}\tilde{\Sigma}_{-j,j}\tilde{\Sigma}_{j,j}^{-1}$ ▷ Obtain Lasso parameter

    $b = \frac{\tilde{a}}{\tilde{b}} X_j(y - X_{-j}(\tilde{\mu}_{-j} - \tilde{\Sigma}_{-j,j}\tilde{\Sigma}_{j,j}^{-1}\tilde{\mu}_j))$ ▷ Obtain Lasso parameter

    $c = \lambda(\exp(\Gamma(\tilde{a} + 0.5) - \Gamma(\tilde{a}) - 0.5\log(\tilde{b})))$ ▷ Obtain Lasso parameter

    $\tilde{\mu}_j = elasso(a, b, c) = \mu_j^*$ ▷ Replace by Local mean

    $\tilde{\mu}_{-j} = \tilde{\mu}_{-j} + \tilde{\Sigma}_{-j,j}\tilde{\Sigma}_{jj}^{-1}(\tilde{\mu}_j^* - \tilde{\mu}_j)$ ▷ Update Global Mean

    $\tilde{\Sigma}_{j,j} = vlasso(a, b, c) = \Sigma_{jj}^*$ ▷ Replace by Updated Local Covariance

    $\tilde{\Sigma}_{j,-j} = \Sigma_{jj}^*\tilde{\Sigma}_{jj}^{-1}\tilde{\Sigma}_{j,-j}$ ▷ Update Global Covariance

    $\tilde{\Sigma}_{-j,j} = \tilde{\Sigma}_{j,-j}^T$ ▷ Update Global Covariance

    $\tilde{\Sigma}_{-j,-j} = \tilde{\Sigma}_{-j,-j} + \tilde{\Sigma}_{-j,j}\tilde{\Sigma}_{j,j}^{-1}(\Sigma_{j,j}^* - \tilde{\Sigma}_{j,j})\tilde{\Sigma}_{j,j}^{-1}\tilde{\Sigma}_{j,-j}$ ▷ Update Global Covariance

return $\tilde{\mu}, \tilde{\Sigma}$

The following bullet points summarize 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 $\tilde{\mu}$ and $\tilde{\Sigma}$. The initial estimate of $\tilde{\mu}$ and $\tilde{\Sigma}$ is obtained via MFVB. Define $\beta = (\beta_j, \beta_{-j})$, separated by a target variable and the rest of the variables. μ_j^* and Σ_{jj}^* are local parameters.

- Compute the mean and variance of the Lasso distribution: elasso and vlasso functions in the aforementioned algorithm, with the current values a, b, and$c$: $q^*(\theta_1) \approx N(\mu_j^*, \Sigma_{jj}^*)$.

- Use the update expression $q^*(\beta) = q(\beta_{-j}|\beta_j)\phi(\beta_j; \mu_j^*, \Sigma_{jj}^*)$ to adjust the global mean and global covariance of $q(\beta)$.

- Iterate through each variable β_j and update $\tilde{\mu}$ and $\tilde{\Sigma}$ each time by the derived expression.

3.4.2 Bivariate local global algorithm

Apart from the univariate algorithm that matches each variable of interest into a univariate Lasso distribution, each pair of regression coefficients can be matched simultaneously by a bivariate Lasso distribution as mentioned in subsection 3.3.2. The number of index pairs can be calculated while the combination arithmetic. The total number of index pairs is $\binom{p}{2}$. For example, if $p = 4$ then index pair $\mathcal{I}$: can be $(1,2)$, $(1,3)$, $(1,4)$, $(2,3)$, $(2,4)$, $(3,4)$. The main intuition is similar to Equation 3.2, but the log likelihood of each index pair $\mathcal{I}$ now can be written as:

$$\begin{aligned} \log p(D, \beta_{\mathcal{I}}) &= \mathbb{E}_{\beta_{-\mathcal{I}}, \sigma^2 | \mathcal{D}, \beta_{\mathcal{I}}} \log p(\beta_{\mathcal{I}} \mid \mathcal{D}, \beta_{-\mathcal{I}}, \sigma^2) \\ &\approx \mathbb{E}_{q(\beta_{-\mathcal{I}}|\beta_{\mathcal{I}})q(\sigma^2)} \log p(\beta_{\mathcal{I}} \mid \mathcal{D}, \beta_{-\mathcal{I}}, \sigma^2). \end{aligned} \tag{3.11}$$

The log-likelihood of index pair can be written as the following:

$$\begin{aligned} \log p(\beta_{\mathcal{I}} \mid \mathcal{D}, \beta_{-I}, \sigma^2) &= -\tfrac{1}{2\sigma^2} \beta_{\mathcal{I}}^T X_{\mathcal{I}}^T X_{\mathcal{I}} \beta_{\mathcal{I}} - \frac{\lambda}{\sigma} \|\beta_{\mathcal{I}}\|_1 \\ &\quad + \frac{1}{2\sigma^2} (y - X_{-\mathcal{I}} \beta_{-\mathcal{I}})^T X_{\mathcal{I}} \beta_{\mathcal{I}} + \frac{1}{2\sigma^2} \beta_{\mathcal{I}} X_{\mathcal{I}}^T (y - X_{-\mathcal{I}} \beta_{-\mathcal{I}}). \end{aligned} \tag{3.12}$$

For the purpose of enabling $\beta_{-\mathcal{I}}$ to be symmetric for preventing generating a non-symmetric A, the log-likelihood of index pair $\mathcal{I}$ can be remodified as the following, denoting $S = \mu_{-\mathcal{I}} - T\mu_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}, T = \Sigma_{-\mathcal{I},\mathcal{I}} \Sigma_{\mathcal{I},\mathcal{I}}^{-1} \in \mathbb{R}^{(p-|\mathcal{I}|)\times|\mathcal{I}|}$ similar to the univariate case.

$$\begin{aligned} \log p(\mathcal{D}, \beta_{\mathcal{I}}) &= \tfrac{\widetilde{a}}{\widetilde{b}} \left(y - X_{-\mathcal{I}} S\right)^T X_{\mathcal{I}} \beta_{\mathcal{I}} - \tfrac{\lambda \Gamma(\widetilde{a}+1/2)}{\Gamma(\widetilde{a})\sqrt{\widetilde{b}}} \|\beta_{\mathcal{I}}\|_1 \\ &\quad - \tfrac{\widetilde{a}}{2\widetilde{b}} \beta_{\mathcal{I}}^T \left(X_{\mathcal{I}}^T X_{\mathcal{I}} + \tfrac{1}{2} X_{\mathcal{I}}^T X_{-\mathcal{I}} T + \tfrac{1}{2} T X_{-\mathcal{I}}^T X_{\mathcal{I}}\right) \beta_{\mathcal{I}}. \end{aligned} \tag{3.13}$$

Then the marginal log-likelihood can be matched to a bivariate Lasso distribution:

$$\beta_{\mathcal{I}} \mid \mathcal{D} \overset{\text{approx.}}{\sim} \text{BiLasso}\left(\tfrac{\widetilde{a}}{\widetilde{b}} \left(X_{\mathcal{I}}^T X_{\mathcal{I}} + \tfrac{1}{2} X_{\mathcal{I}}^T X_{-j} T + \tfrac{1}{2} T^T X_{-j}^T X_{\mathcal{I}}\right), \tfrac{\widetilde{a}}{\widetilde{b}} X_{\mathcal{I}}^T \left(y - X_{-\mathcal{I}} S\right), \frac{\lambda \Gamma(\widetilde{a}+1/2)}{\Gamma(\widetilde{a})\sqrt{\widetilde{b}}} \right). \tag{3.14}$$

In order to update and refine the global mean and variance parameters, it is important to consider the conditional distribution of $q(\beta_{-\mathcal{I}}|\beta_{\mathcal{I}})$. This distribution can be expressed in the following manner:

$$\begin{aligned} q(\beta_{-\mathcal{I}} \mid \beta_{\mathcal{I}}) &= N_{p-|\mathcal{I}|} \left(\mu_{-\mathcal{I}} + \Sigma_{-\mathcal{I},\mathcal{I}} \Sigma_{\mathcal{I},\mathcal{I}}^{-1} (\beta_{\mathcal{I}} - \mu_{\mathcal{I}}), \Sigma_{-\mathcal{I},-\mathcal{I}} - \Sigma_{-\mathcal{I},\mathcal{I}} \Sigma_{\mathcal{I},\mathcal{I}}^{-1} \Sigma_{\mathcal{I},-\mathcal{I}}\right) \\ &= N_{p-|\mathcal{I}|} \left(S + T\beta_{\mathcal{I}}, \Sigma_{-\mathcal{I},-\mathcal{I}} - \Sigma_{-\mathcal{I},\mathcal{I}} \Sigma_{\mathcal{I},\mathcal{I}}^{-1} \Sigma_{\mathcal{I},-\mathcal{I}}\right). \end{aligned} \tag{3.15}$$

By utilizing Equation 3.15 and extending the propagation equation derived from Equation 3.8, the update formula for $\tilde{\mu}$ and $\tilde{\Sigma}$ can be modified using the same equation as in the univariate case, as presented in Equation 3.9 and Equation 3.10. The only alteration required is the substitution of the subscript j with the index pair $\mathcal{I}$.

Algorithm 5 Bivariate Local-Global Algorithm

Input: X, y, parameters $(\tilde{a}, \tilde{b}, \tilde{\mu}, \tilde{\Sigma})$ from MFVB, λ
while $\tilde{\mu}$ is changing less than ϵ **do**
  for $\mathcal{I} = (1, 1)$ to $(p-1, p)$ **do**
    $a = \frac{\tilde{a}}{\tilde{b}} \left(X_{\mathcal{I}}^T X_{\mathcal{I}} + \frac{1}{2} X_{\mathcal{I}}^T X_{-j} T + \frac{1}{2} T^T X_{-j}^T X_{\mathcal{I}} \right)$ ▷ Obtain Lasso parameter
    $b = \frac{\tilde{a}}{\tilde{b}} X_{\mathcal{I}}^T \left(y - X_{-\mathcal{I}} s \right)$ ▷ Obtain Lasso parameter
    $c = \lambda(\exp(\Gamma(\tilde{a} + 0.5) - \Gamma(\tilde{a}) - 0.5 \log(\tilde{b})))$ ▷ Obtain Lasso parameter
    $\tilde{\mu}_{\mathcal{I}} = Bielasso(a, b, c) = \mu_{\mathcal{I}}^*$ ▷ Replace by Local mean
    $\tilde{\mu}_{-\mathcal{I}} = \tilde{\mu}_{-\mathcal{I}} + \tilde{\Sigma}_{-\mathcal{I},\mathcal{I}} \tilde{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1} (\tilde{\mu}_{\mathcal{I}}^* - \tilde{\mu}_{\mathcal{I}})$ ▷ Update Global Mean
    $\tilde{\Sigma}_{\mathcal{I},\mathcal{I}} = Bivlasso(a, b, c) = \Sigma_{\mathcal{I}\mathcal{I}}^*$ ▷ Replace by Updated Local Covariance
    $\tilde{\Sigma}_{\mathcal{I},-\mathcal{I}} = \Sigma_{\mathcal{I}\mathcal{I}}^* \tilde{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1} \tilde{\Sigma}_{\mathcal{I},-\mathcal{I}}$ ▷ Update Global Covariance
    $\tilde{\Sigma}_{-\mathcal{I},\mathcal{I}} = \tilde{\Sigma}_{\mathcal{I},-\mathcal{I}}^T$ ▷ Update Global Covariance
    $\tilde{\Sigma}_{-\mathcal{I},-\mathcal{I}} = \tilde{\Sigma}_{-\mathcal{I},-\mathcal{I}} + \tilde{\Sigma}_{-\mathcal{I},\mathcal{I}} \tilde{\Sigma}_{\mathcal{I},\mathcal{I}}^{-1} (\Sigma_{\mathcal{I},\mathcal{I}}^* - \tilde{\Sigma}_{\mathcal{I},\mathcal{I}}) \tilde{\Sigma}_{\mathcal{I},\mathcal{I}}^{-1} \tilde{\Sigma}_{\mathcal{I},-\mathcal{I}}$ ▷ Update Global Covariance
return $\tilde{\mu}, \tilde{\Sigma}$

The following bullet points summarize the procedure of Bivariate Local-Global Algorithm

- Our target: $p(\theta|\mathcal{D}) = p(\beta, \sigma^2|\mathcal{D})$, where a normal distribution is used to approximate with global parameters $\tilde{\mu}$, and $\tilde{\Sigma}$. The initial estimate of $\tilde{\mu}$ and $\tilde{\Sigma}$ has been obtained via MFVB. Define $\theta = (\beta_j, \beta_{-j})$, separated by a target variable and the rest of the variables. μ_j^*, and Σ_{jj}^* as local parameter.
- Compute the mean and variance of bivariate Lasso distribution: bielasso and bivlasso functions in the aforementioned algorithm, with the current a, b, and c: $q^*(\beta_j) \approx N(\mu_j^*, \Sigma_{jj}^*)$.
- Use the update expression $q^*(\theta) = q(\beta_{-j}|\beta_j)\phi(\beta_j; \mu_j^*, \Sigma_{jj}^*)$ to adjust the global mean and global covariance of $q(\beta)$.

- Iterates through each variable β_j and update $\tilde{\mu}$ and $\tilde{\Sigma}$ each time by the derived expression.

Nevertheless, the Bivariate-Local-Global algorithm has an issue currently. One of the main issues can be the result of generating nan value for $\tilde{\mu}$ and components in $\tilde{\Sigma}$. The origin of this problem has not been found until now. Due to the limited amount of time, we would like to resolve this issue in the future. As a consequence, we would not experiment with the effectiveness and the approximation of it in the next Chapter.

Chapter 4

Experiment Result and Analysis

4.1 Experimental Setting

4.1.1 Evaluation metric

In this study, we employ a range of evaluation metrics to assess the effectiveness of our algorithms.

- l_1 norm accuracy

$$l_1(f, g) = \int |f(x) - g(x)|dx. \tag{4.1}$$

$$\text{Acc}(f, g) = 1 - \frac{1}{2} l_1(f, g). \tag{4.2}$$

- Run time: This metric represents the total time (in seconds) taken to generate the posterior density.

The l_1 norm is commonly employed to compare probability densities, particularly to assess the accuracy of the distribution's central tendency. The emphasis on calculating l_1 norm accuracy stems from the recognition that the accuracy of the tail distribution approximation is relatively less significant compared to that of the central distribution. The primary objective of our approximation for the posterior distribution centers around achieving precision in the central region. Additionally, it is important to evaluate the execution time required for generating posterior parameters. This examination allows for a comprehensive comparison of time complexity among the Local-Global algorithm, MFVB, and MCMC methods. The execution time is measured in seconds.

4.1.2 Experimental Datasets

The following bullet points demonstrate the dataset description, which includes the introduction to the purpose of the dataset, and a number of predictors and number of samples, respectively.

- **Hitters**:
  - Type: Baseball statistics dataset
  - Predictors (p): 20, Samples (n): 263
  - Description: Contains baseball player statistics, including performance measures and salary information.
  - Further description: High correlation between predictors
- **Kakadu**:
  - Type: Environmental dataset
  - Predictors (p): 22, Samples (n): 1828
  - Description: Relates environmental factors to the abundance of amphibians in Kakadu National Park, Australia.
  - Further description: Approximately normal posterior distribution due to large number of samples.
- **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.
  - Further description: Approximately normal distribution, strong correlation between predictors.
- **Prostate**:

  - Type: Medical dataset
  - Predictors (p): 8, Samples (n): 97
  - Description: Prostate cancer data with clinical measurements and Logarithm of Prostate Specific Antigen (lpsa).
  - Further description: Approximately normal posterior distribution, no strong correlation between predictors.

- **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.
  - Further description: Approximately normal posterior distribution, no strong correlation between predictors.

- **Eye data**:
  - 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.
  - Further description: The number of predictors is higher than the sample size.

In particular, the Hitters dataset and Eye data dataset deserve additional attention due to their distinct characteristics. Hitters exhibits a substantial correlation among predictors, while Eyedata possesses a larger number of predictors compared to the number of samples, making both datasets more challenging to approximate accurately. However, if the Local-Global algorithm can achieve higher accuracy and better approximation density than MFVB on these challenging datasets, it suggests that the algorithm's success can extend to other datasets that are comparatively easier to approximate. The experimental results,

which will be presented in a later subsection, will primarily focus on Hitters and Eye data while also comparing the findings to those obtained from other datasets.

4.2 Experimental Result

4.2.1 Approximation Density Visualization

The following six plots illustrate the approximated density plots for each of the datasets. Since each dataset contains a large number of predictors, we will present two density plots for each dataset. The first plot represents the best density approximation achieved by the Local-Global algorithm with MCMC, while the second plot represents the worst density approximation, exhibiting the largest deviation from MCMC. To enhance readability, the best plot will be positioned on the left, and the worst plot will be placed on the right.

Regarding the density plots for the Kakadu, Prostate, Bodyfat, and Credit datasets, there are no significant deviations observed between the density plots generated by MFVB, Local-Global Global (LG-Global) approximation, and Local-Global Local (LG-Local) approximation with MCMC. The predictors in these datasets generally demonstrate a roughly normal density pattern. As a result, the approximation accuracy for MFVB remains relatively high, whereas LG-Local and LG-Global may experience a more noticeable decrease in approximation accuracy due to the distinctive characteristics of these datasets.

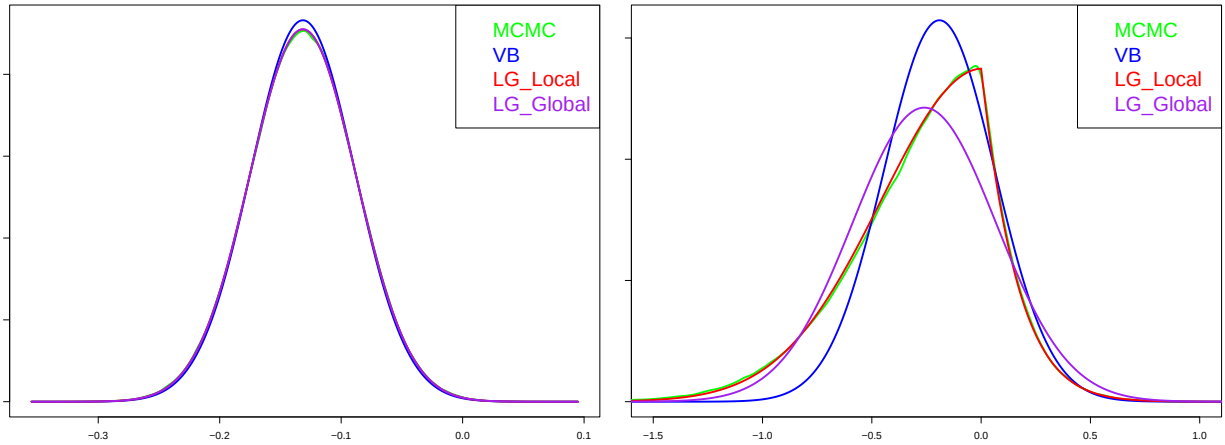


Figure 4.1: Part of Approximation Density for Hitters dataset; Left: best case, Right: worst case

In Figure 4.1, the density plot of this particular predictor reveals significant overlap with the MCMC plot, as shown on the left-hand side. This overlap indicates a high level of accuracy in the approximation performance.

In contrast, when examining the worst density plot, it is observed that MFVB tends to deviate from the left-skewed MCMC (considered as the gold standard). Particularly in scenarios where the actual distribution possesses a sharp tuning point, both the local and global algorithms, particularly the local approximation line, demonstrate a closer approximation to the gold standard. This improved performance can be attributed to the inclusion of the absolute term $c|x|$ in the lasso distribution formula, which allows for accommodating various potential density functions.

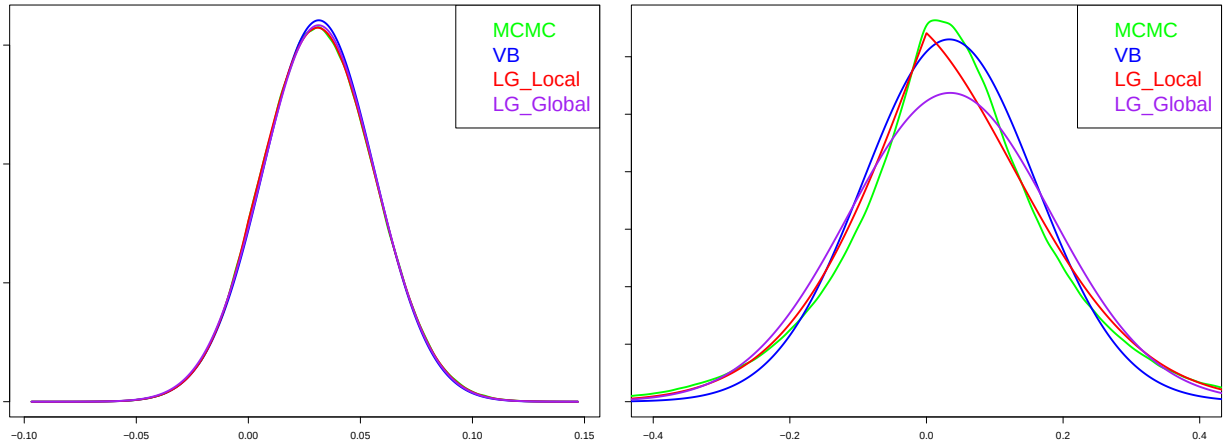


Figure 4.2: Part of Approximation Density for Kakadu dataset; Left: best case, Right: worst case

In Figure 4.2, the best density plot obtained from the Local-Global algorithms exhibits significant overlap with both MCMC and MFVB, similar to the findings in Figure 4.1. In the worst-case scenario, LG-Local continues to demonstrate the best approximation performance. Despite the gold standard density displaying high variance, LG-Local manages to fit a density with a comparable variance, even in the presence of a sharp tuning point.

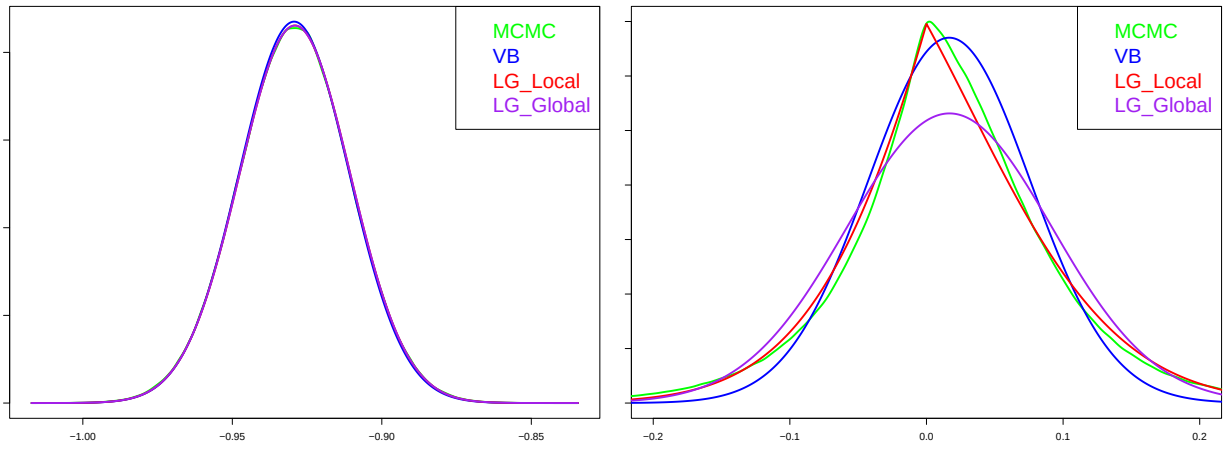


Figure 4.3: Part of Approximation Density for Bodyfat dataset; Left: best case, Right: worst case

In Figure 4.3, the best density plot obtained from the Local-Global algorithms exhibits significant overlap with both MCMC and MFVB. This overlap is attributed to the dataset

itself, which demonstrates an asymptotically normal data distribution, similar to the case mentioned earlier. Furthermore, under the worst-case scenario, LG-Local continues to showcase the best approximation performance, as shown on the right side of Figure 4.3. The LG-Local line aligns closely with the density plot generated by the gold standard, represented by the green line.

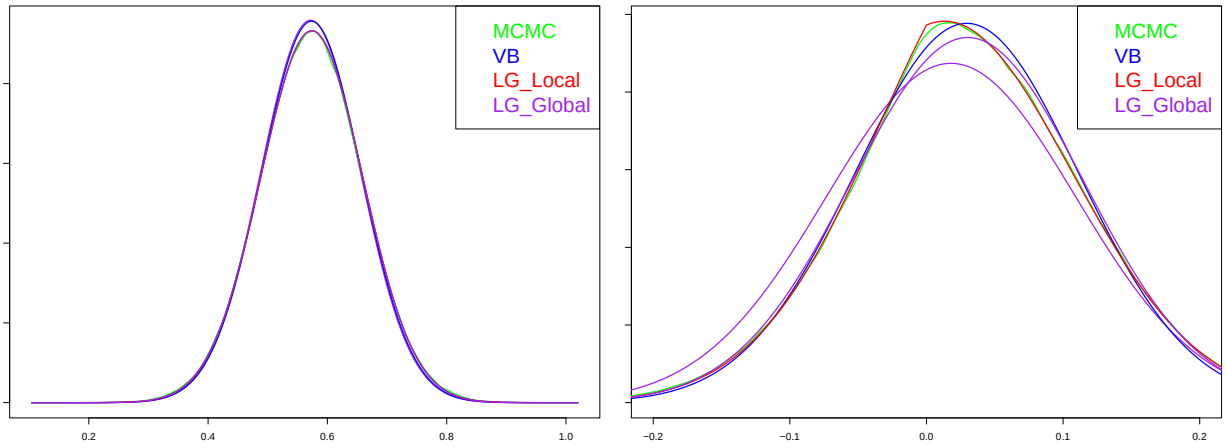


Figure 4.4: Part of Approximation Density for Prostate dataset; Left: best case, Right: worst case

In the best-case scenario depicted in Figure 4.4, MFVB slightly overestimates the posterior variance. However, both LG-local and LG-global exhibit high robustness against the target distribution, accurately capturing its characteristics.

On the other hand, in the worst-case scenario, while LG-global moderately underestimates the posterior variance, LG-local demonstrates once again its consistency with MCMC, accurately reflecting the distribution's properties.

In a similar fashion to previous observations, the Local-Global approximation exhibits significant overlap with both MFVB and MCMC, as illustrated on the left side of Figure 4.5.

In the worst-case plot, MFVB tends to overestimate the posterior variance. However, both LG-local and LG-global demonstrate a remarkable alignment with the MCMC density plot, indicating their ability to accurately capture the distribution's characteristics.

In the best scenario depicted on the left side of Figure 4.6, the most significant distinction is observed. This is due to the fact that, for a majority of predictors in the Eyedata dataset, the actual distribution, as represented by the MCMC plot, does not follow an

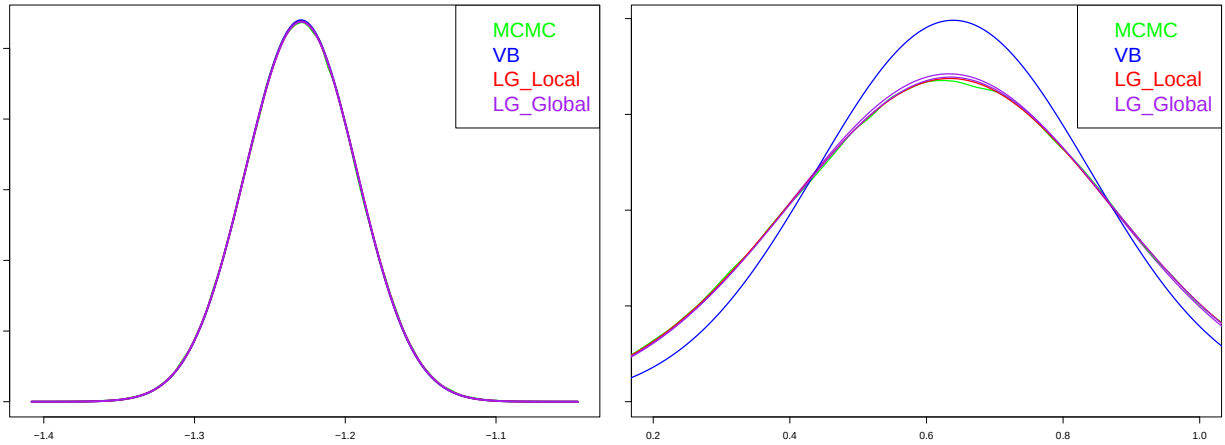


Figure 4.5: Part of Approximation Density for Credit dataset; Left: best case, Right: worst case

asymptotically normal pattern. As a result, the optimal approximation case displays a Laplacian-like distribution. This deviation from the assumed distribution can directly lead to inaccurate results for both MFVB and LG-Global.

In the worst-case scenario shown on the right side of Figure 4.6, the marginal LG-local approximation slightly underestimates the data distribution, but overall, it still outperforms the other two plots. MFVB approximation, on the other hand, considerably overestimates the posterior variance once again.

Although the assumption of normal distribution exists for LG-Global, it tends to correct the overestimated posterior variance of the regression coefficient when compared with MFVB.

4.2.2 Approximation Accuracy Result

The following six tables present the experimental results for the approximation accuracy of three different algorithms. LG-Local represents the marginal approximation of the Local-Global Algorithm using the lasso distribution, while LG-Global represents the global approximation of the Local-Global Algorithm using a univariate normal distribution. The global mean for LG-Global is set at the j_{th} component, and the variance is determined by the j_{th} row and j_{th} column of $\tilde{\Sigma}$.

Each row in the tables corresponds to a specific quantile of approximation accuracy,

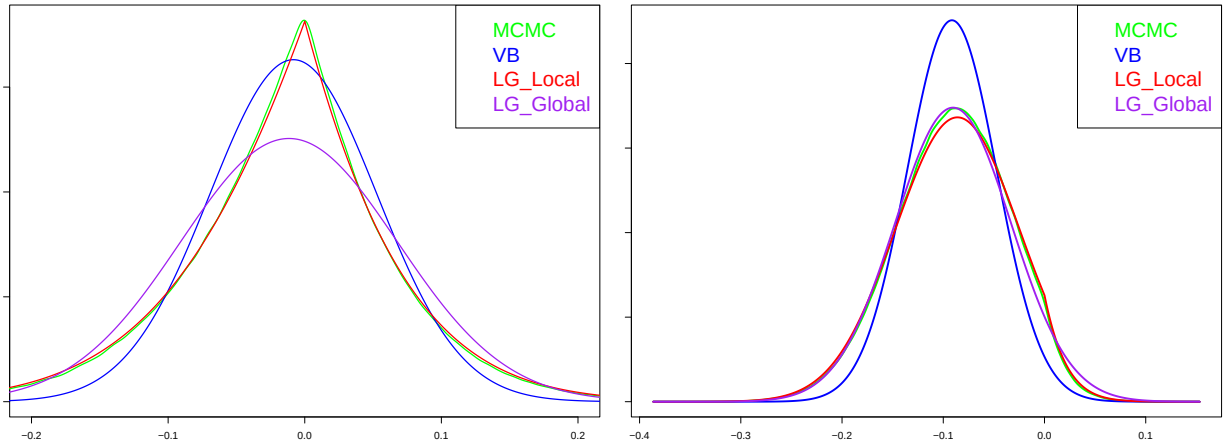


Figure 4.6: Part of Approximation Density for Eyedata dataset; Left: best case from 52nd predictor, Right: worst case from 200th predictor

such as the minimum approximation accuracy, maximum approximation accuracy, and so on. LG-Local reflects the local approximation accuracy achieved by our algorithm, while LG-Global represents the global approximation accuracy. The VB column denotes the Mean-Field-Variational-Bayes algorithm, and the MCMC column refers to the Monte Carlo Markov Chain method, which serves as a gold standard with 100% accuracy for each approximation density.

Table 4.1 presents the approximation results for the Hitters dataset. The global approximation of the Local-Global algorithm outperforms the benchmark approach, MFVB, across all metrics by approximately 1 to 5 percent. The running time of the Local-Global algorithm is 0.17s, which is 0.03s slower than MFVB. This is expected as MFVB only involves the global parameter approximation, without the additional step of local approximation. Furthermore, both methods are 400 times faster than MCMC.

The Local-Global algorithm demonstrates superior approximation accuracy in this particular dataset, which is considered the most challenging among all the datasets, without a significant increase in running speed.

Table 4.2 presents the approximation results for the Kakadu dataset. While the minimum global approximation of the Local-Global algorithm is lower than MFVB, all other metrics show slightly higher values for the Local-Global algorithm. Regarding the local approximation, LG-Local, there is a one to two percentage increase compared to MFVB.

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

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

Additionally, the execution time for the Local-Global algorithm is 0.05s shorter than that of MFVB (0.14s). Both methods achieve a speed that is 6000 times faster than that of

MCMC. This significant speed improvement is due to the high number of samples in the Kakadu dataset, which requires a longer running time for MCMC sampling. Despite the longer execution time required for MCMC, the performance of both the MFVB method and the Local-Global algorithm does not degrade significantly because the data distribution in Kakadu is approximately normal.

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

Table 4.3 presents the approximation results for the Bodyfat dataset. Similar to the Kakadu dataset, the minimum global approximation of the Local-Global algorithm is lower than MFVB, while all other metrics show slightly higher values for the Local-Global algorithm. In terms of the local approximation, LG-Local, there is a 2 to 5 percent increase compared to MFVB. Moreover, the execution time for the Local-Global algorithm is 0.03s shorter than that of MFVB (0.14s). Both methods achieve a speed that is 300 times faster than that of MCMC.

The data distribution in the Bodyfat dataset is similar to that of the Kakadu dataset, resulting in no drastic difference between MFVB and the Local-Global algorithm.

Table 4.4 presents the approximation results for the Prostate dataset. Similar to the Kakadu and Bodyfat datasets, the global approximation of the Local-Global algorithm is

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

slightly higher than MFVB. Additionally, the local approximation accuracy of LG-Local surpasses both the global approximation and MFVB.

In terms of execution time, the Local-Global algorithm is 0.01s faster than MFVB (0.11s), and both methods achieve a speed that is 300 times faster than that of MCMC.

The data distribution in the Prostate dataset is similar to that of the Bodyfat and Kakadu datasets, with an approximately normal distribution. Consequently, there is no significant distinction between the performance of MFVB and the Local-Global algorithm in this dataset.

Table 4.5 showcases the approximation results for the Credit dataset. Similar to the Prostate dataset, the global approximation of the Local-Global algorithm is slightly higher than MFVB, although the minimum value of MFVB is approximately 8% lower than LG-Global. Both LG-Global and LG-Local achieve an approximation accuracy above 99%.

In terms of execution time, the Local-Global algorithm requires 0.11s, which is only 0.01s slower than the Local-Global Algorithm. Both methods achieve a speed that is 300 times faster than that of MCMC.

It is worth noting that the Credit dataset, like the Prostate, Bodyfat, and Kakadu

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

datasets, exhibits properties of a roughly normal distribution. Consequently, the results obtained align with the findings from the previous datasets.

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

Table 4.6 presents the approximation results for the Eyedata dataset. The global ap-

proximation of the Local-Global algorithm surpasses that of MFVB by a significant margin. Furthermore, the local approximation accuracy of LG-Local is notably higher than both the global approximation and MFVB.

In terms of execution time, the Local-Global algorithm is 0.51s faster than MFVB (1.21s). Both methods achieve a speed that is 18000 times faster than that of MCMC.

It is important to highlight the unique characteristic of the Eyedata dataset, namely its high-dimensional sparsity. The curse of dimensionality poses challenges to the approximation process, particularly for MCMC, which requires more steps to converge. In this case, although MFVB offers slightly faster execution, the Local-Global algorithm demonstrates a more robust performance in terms of approximation accuracy.

Chapter 5

Discussion and Conclusion

5.1 Discussion

The experiment results presented in the previous chapter provide insights into the performance of different algorithms in terms of approximation accuracy and density plots. The following six bullet points highlight notable observations from the experiments:

- MFVB tends to generate densities with less variance, resulting in higher concentration at the center.
- The marginal posterior distribution obtained from the Local-Global algorithm demonstrates high accuracy and versatility in accommodating various predictor distributions beyond the normal distribution.
- The global posterior distribution obtained from the Local-Global algorithm exhibits greater accuracy compared to MFVB.
- The running time of MFVB is slower than that of the Local-Global algorithms, both for local and global approximation.
- The Local-Global algorithm maintains high accuracy even when there is a high correlation between predictors, as observed in the results from the Hitters dataset.
- The Local-Global algorithm achieves high accuracy even when the number of predictors is higher than the sample size, as evidenced by the results from the Eyedata dataset.

These bullet points effectively summarize the notable phenomena observed in the study. Further explanations reveal that the variance of the density produced by MFVB tends

to be less because of the adaptability of the function form of the univariate Lasso distribution. The addition of local adjustment by capturing the correlation between β_j and β_{-j} makes the global approximation from the Local-Global algorithm more accurate than MFVB. Although the Local-Global algorithm is slightly slower than MFVB due to an extra procedure for the local approximation process and calculation of moments for Lasso distribution, both algorithms are faster than MCMC. Additionally, the Local-Global algorithm's high accuracy in the presence of a high correlation between predictors and more predictors than samples demonstrates the algorithm's adaptability and adjustment, leading to better approximation results.

5.2 Limitations

The two bullet points effectively summarize the limitations:

- The automatic choice of λ is still obtained by Gibbs Sampling.
- The Univariate Local-Global algorithm cannot handle the case when the initial covariance is a diagonal matrix.

The first limitation is explained further, discussing the current choice of λ from the three-step Gibbs sampler. It suggests that using the fully conditional posterior distributions of λ^2 with a gamma prior distribution could lead to an optimal λ posterior distribution due to conjugacy. However, obtaining these posterior estimates for λ takes a significant amount of execution time, similar to sampling other posterior estimates.

The second limitation is explained by discussing the behavior of the update formula for $\tilde{\Sigma}$ in the univariate Local-Global algorithm. It highlights that if the initial covariance $\tilde{\Sigma}$ is diagonal, it remains a diagonal matrix throughout the iterations, which restricts the algorithm's generalizability. The explanation provides an example of how the update formula results in a diagonal matrix, causing a discrepancy when the actual Σ is non-diagonal.

5.3 Future Work

Several improvements can be explored as a future study to address the aforementioned limitations.

- Propose a Bivariate-Local-Global Algorithm to address the problem when the initial covariance is a diagonal matrix.
- Derive the updated formula of σ^2.

Firstly, since the update for each iteration is done via a pair of variables, the covariance matrix of each variable will be updated multiple times for each pair. Secondly, our Local-Global algorithm can only obtain the posterior distribution for the regression coefficient β, while the temporary assumption of independence between σ^2 and β is limiting. In fact, it is necessary to derive and explore an alternative update formula to calculate Σ in the future.

5.4 Conclusion

In conclusion, we have proposed a novel algorithm for the Bayesian Lasso, which utilizes a local approximation correction approach to capture the correlation between the distribution of the target variable and other variables. The introduction of the Lasso distribution enhances the precision of the local approximation, leading to significant improvements over the MFVB approximation.

Looking ahead, our future work will primarily focus on successfully implementing the Bivariate-Local-Global algorithm as mentioned in the last paragraph in Chapter 3. Furthermore, a valid derivation formula for updating the covariance distribution $q(\sigma^2)$ will be a key area of exploration. These efforts will contribute to advancing the accuracy and efficiency of Bayesian variable selection and approximation methods.

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.

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.

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.

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.

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.

Khare, K., Hobert, J.P., 2013. Geometric ergodicity of the Bayesian Lasso. Electronic Journal of Statistics 7. doi:10.1214/13-ejs841.

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.

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.

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.