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

Yuhao Li

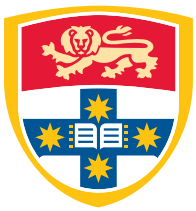
Supervisor: A/Prof. John Ormerod

A thesis submitted in partial fulfillment of

the requirements for the degree of

Bachelor of Science(Honour)(Data Science)

Mathematics and Statistics



THE UNIVERSITY OF SYDNEY

June 2023

Statement of originality

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

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

Yuhao Li

Abstract

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

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

On the other hand, obtaining the posterior of the Bayesian Lasso model involves using Monte Carlo Markov Chain method such as Gibbs sampler for exact posterior distribution. Even though MCMC is famous for its oracle property such as generating arbitary exact target samples if burn-in period is enough, MCMC is slow with high computational cost.

Meanwhile, Variational Approximation: as a deterministic approximation algorithm for intractable posterior distribution, has been applied prevalently for fast Approximate Bayesian Inference(ABI) among the Bayesian Statistical community, while it is also a faster alternative to Monte Carlo methods such as Markov Chain and Monte Carlo(MCMC).

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

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

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

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

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

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

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

Acknowledgements

Thanks to Supervisor and family

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

Contents

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 one of the shrinkage methods. As one of the most traditional shrinkage methods, Lasso regression has been proven for his success in Statistical Community over the years.

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

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

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

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

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

The Lasso estimate of regression coefficient is based on Equation (1.3), where the main distinction of Lasso is adding a penalty term of absolute value of regression coefficients β in

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

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

Larger penalization leads to more sparse solution of regression coefficient due to a square constraint set results from L_1 penalty function, so that achieve variable selection of parameter, generating a higher prediction accuracy as well as interpretable model since we could drop the estimated regression coefficient that has 0 and state they have weak effect for prediction according to Tibshirani (1996). However, due to non-existence of derivative of absolute value of regression coefficient β, alternative improved algorithm have been purposed and deployed such as Least Angle Regression(LARS), iterative soft-thresholding, subgradient method, and iteratively reweighted least square(IRLS) by Efron et al. (2004), Beck and Teboulle (2009), Zhang and Zeng (2005) and Friedman et al. (2010).

Bayesian Lasso One of the detriments of the ordinary lasso is that the variation of inferential quantity can't be captured properly. To resolve this issue, Tibshirani (1996) also suggests that the lasso estimate can also be extended under the Bayesian framework, which can be described as posterior mode Equation(1.4) if independent and identically distributed Laplacian prior from Equation (1.5) is assigned,together with the likelihood form on Equation(1.6).

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

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

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

Introduction of Bayesian Lasso Problem(Why Bayesian Lasso) Nevertheless, there is no tractable integration form for the Bayesian Lasso posterior until Park and Casella (2008) further explore the Lasso model under the setting of Bayesian framework, where the choice of a conditional Laplace prior distribution over the regression coefficient β conditioning by standard error σ^2 is added to the Lasso penalty formulation in the frequentist framework to ensure unimodality of full posterior distribution. Based on closed form of tractable posterior distribution, a three-step Gibbs sampler is proposed to draw approximate samples from Bayes Lasso posterior distribution, which can be utilized for further

inference of parameter of interest. There are several benefits for using the Bayesian Lasso model. Firstly, it has easier implementation than the traditional Lasso, although more computation of conditional distribution form is demanded. Secondly Bayesian credible interval of parameters can be generated simultaneously for modelling uncertainty and therefore can also guide variable selection based on the interpretation that lasso estimated is regarded as the mode of posterior distribution of β. Thirdly, Park and Casella (2008) also state that the Bayesian Lasso model could also be a potential solution for addressing the issue of attaining optimal tuning parameter λ by marginal maximum likelihood method together with a sue of an suitable hyperprior such as gamma prior on the square of tuning parameter λ: λ^2. It could support a more stable automatic tuning process of choosing the most appropriate tuning parameter λ for the Lasso problem, as opposed to inefficient K-fold cross validation approach under the ordinary Lasso model that is time consuming and computational demanding. Lastly, the three-step Bayesian Lasso Gibbs Sampler proposed by Park and Casella (2008) would yield an exact posterior distribution that can be sampled given exact form of conditional distribution of each model parameter conditioning on rest of the other parameters. Theoretically, Khare and Hobert (2013) has demonstrated a Bayesian Lasso Gibbs sampler version of central limit theorem(CLT), indicating Bayesian Lasso Gibbs Sampler satisfy geometriclly ergodicity if any values of sample size $n \geq 3$ and arbitary number of regression coefficient p, data matrix X, tuning parameter λ are assigned. This means, the Bayesian lasso Gibbs Sampler is able to achieve asymptotically uncertainty of posterior estimation. To address this issue, Rajaratnam and Sparks (2015a) invented a reduced step Gibbs sampler instead, successfully accelerating the sampling procedure.

Approxiamte Bayesian Inference Intuition

Review of Bayesian inference intuition can be referred from Section 2.1, but challenges of Bayesian Inference motivates the approximate bayesian inference method thereafter. To be specific, Bishop (2006) states three main challenges of obtaining posterior distribution. Firstly, the dimension of target parameter might be high, which results in heavy computational cost for estimating posterior distribution. Secondly, the exact posterior distribution form might be too complicated to be tractable,. Thirdly, the computation of the posterior mean parameter $\int_\theta \theta p(\theta|\mathcal{D})$ has high probability without having an simple calculation,

resulting in the fact that there might not exist an closed form analytical solution for integration. Much efforts have been paid over the years, there are two main types of sampling approach that are effective currently, which are stochastic sampling algorithms and deterministic approximation algorithms.

Introduction of ABI method There are two genres of Approximate Bayesian Inference method, which includes stochastic approach such as MCMC, where an exact result can be obtained if infinite computational resource is assigned. Another category lies in deterministic approach that as an faster substitution compared with stochastic approximation approaches. As stated above, an approximate inference method such as MCMC is used for posterior distribution estimation. The necessity of approximate bayesian inference method is due to its' expensive and infeasible exact posterior distribution parameter estimate.

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

Challenges of Bayesian Lasso model Returning back to the Bayesian Lasso model, the three-step Gibbs sampler belongs to the class of MCMC algorithm, which produce samples demonstrates gradually decreasing correlated samples as the burn-in period is increased. Burn-in period refers to the time the sample before that period have to be abandoned. On the other hand, the three-step Gibbs sampler from Markov Chain Monte Carlo(MCMC) under the class of stochastic approximation algorithm is time-consuming and computational challenging, given the fact that it normally requires a long time to converge inside the interval of an acceptable tolerance, especially when n is small and p is large derivation according to the upper bound of convergence rate by Rajaratnam and Sparks (2015b).

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

for the distribution from this family that is the closest to the target posterior distribution with some discrepancy metric such as the Kullback–Leibler divergence etc. An optimization based system in Equation (1.8) is established by iteratively updating variational parameter with an appropriate optimization algorithm such as Coordinate Ascent to obtain approximated posterior distribution that is in the family of Q, while the most common choice is the Normal Distribution due to its simple form and adaptability to other distribution. To further illustrate the intuition, Figure 1.1 provides further explanation of aforementioned intuition, goal and procedure of Variational Inference.

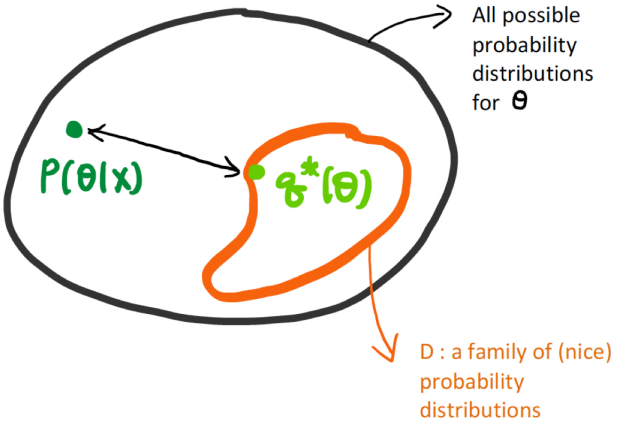


Figure 1.1: Variational Inference intuition, where X is data $\mathcal{D}$, D is equivalent to Q defined above

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

$$\mathrm{KL}(q||p(.|\mathcal{D})) = -\int q(\theta) log(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)}) d\theta + \log p(\mathcal{D}). \tag{1.9}$$

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

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

$$q^*(\theta) = \underset{q_\theta \in Q}{\text{argmax}} \text{ELBO}(q(\theta)), \tag{1.10}$$

$$\begin{aligned} \text{ELBO}(q(\theta)) &= \int q(\theta) \log(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)} d\theta = \mathbb{E}_{q(\theta)} \log(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)}). \\ &= \mathbb{E}_{q(\theta)}[\log p(\theta, \mathcal{D})] - \mathbb{E}_{q(\theta)}[\log[q(\theta)]] \\ &= \mathbb{E}_{q(\theta)}[\log p(\theta, \mathcal{D})] - \text{KL}(q(\theta)||p(\theta)) \end{aligned} \tag{1.11}$$

Mean Field Variational Bayes, History and introduction

The most traditional Variational Inference algorithm is known as Mean-Field Variational Bayes motivated by mean-field Theory in statistical physics yielded by Parisi and Shankar (1988), which assume the approximated distribution is from independent product of parameter distribution from set Q as in Equation (1.12) if assuming there are k sub-parameter of target parameter θ.

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

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

Advantages of Variational Inference Variational approach is more superior and appropriate at present for several reasons according to Blei et al. (2017). One is that Variational Inference algorithm shows a descent computation cost and scalability, and thus Variational inference can be adaptive when the amount of data is huge. For instance, if there exists a billion image that requires to be fitted into probabilistic machine learning model, then exact precision method such as MCMC will be computataional demanding, while Variational Inference would sacrifice tiny accuracy with hundreds of time faster speed as return. Secondly, the descent time-efficiency of Variational Inference becomes another significant factor why it is popular, given the fact it only involves updating variational parameters iteratively until convergence, as opposed to MCMC that produce correlated samples that limit the ideal behavior of MCMC algorithm.

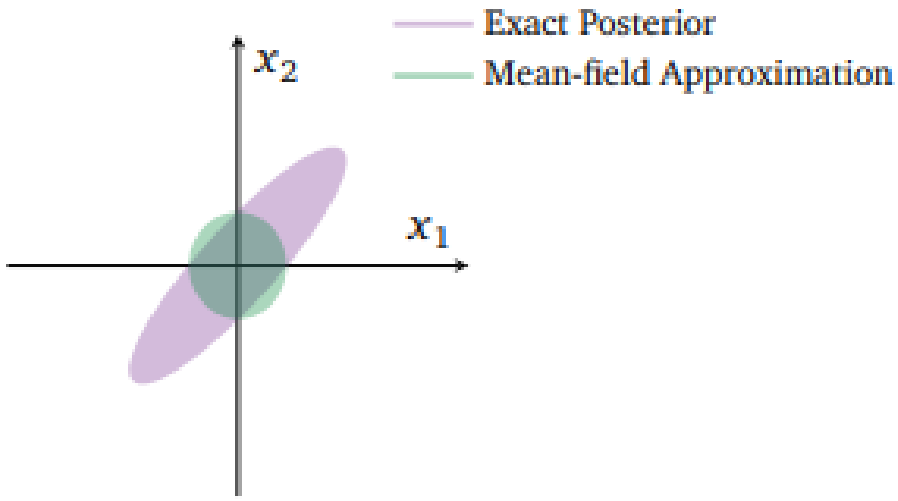


Figure 1.2: Visualization of Mean-Field Variational Approximation compared with exact posterior when correlation is large

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

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

Motivation Motivated by the intention of further enhancing the approximation ac-

curacy of Variational Bayes, as well as the oracle property of Bayesian Lasso regression coefficient estimation for variable selection and standard error estimation leads to more desired demand for fast approximate inference. We would like to design new Variational Inference based algorithms for the Bayesian Lasso regression problem, for the purpose of obtaining bayesian lasso posterior distribution in a much faster and more accurate manner based on the Mean-Field Variational Bayes assumption.

Feeding initial value of mean μ and covariance parameter Σ and writing out the marginal likelihood form of regression coefficient β_i for j th variable, $\log(p(\mathcal{D})|\beta_j)$, we've invented a new distribution called univariate lasso distribution for matching the marginal likelihood. Mixing each marginal likelihood with a Gaussian Approximation, we've shown its' approximation accuracy have surpassed every existing algorithm such as Mean-Field Variational Bayes. Even though the speed of our algorithm is slightly slower than MFVB, the approximation accuracy illustrates a tiny gap between exact estimation from MCMC, with hundred times faster time complexity. Nevertheless, there is a drawback of this method, given the fact that the global covariance matrix would remain diagonal if the initial global covariance matrix is diagonal. To remedy this issue, we've also purposed another algorithm based on marginal likelihood estimation by a bivariate lasso distribution. Instead of updating corresponding mean, covariance matrix for each variable for each iteration, the marginal likelihood of each pair of variables would be matched, so that further generalize our algorithm. Our conclusion are the univariate lasso algorithm would be faster with a lower accuracy while bivariate lasso algorithm would be slower with a higher accuracy, since it updates each pair of variables at a time resulting in $\binom{p}{2}$ of unique parirs. We will show the full intuition and idea later in Chapter 3. By utilizing and fitting a both univariate and multivariate Lasso distribution to each of the marginal distribution, an improved estimate for global Gaussian Approximation can be obtained as defined in Equation (1.13). Finally, we will show our experiment result in 4 using various accuracy metrics.

Our contribution have been listed in the following subsection 1.2.

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

1.2 Contribution

Our main contribution could be concluded as the following part:

- Introduction of Univariate and Multivariate Lasso Distribution
- Derivation of properties for Univariate Lasso Distribution, such as the expectation, variance, Cumulative Density Function form etc.
- Derivation of properties for Multivariate Lasso Distribution such as the expectation, variance, Cumulative Density Function form etc.
- Implementation of Univariate Lasso Distribution and Multivariate Lasso Distribution property in R.
- Design of two new Variational Inference approaches based on local approximation by univariate lasso distribution and multivariate lasso distribution respectively.
- Conduct of experiment to testify two algorithms under dataset by several evaluation metrics for approximation accuracy such as Hitters dataset etc.

1.3 Thesis Organization

This paper will be divided up into 5 chapters. Chapter 1 briefly illustrate the motivation and background of the Lasso problem, Bayesian Lasso Problem, the motivation for Approximate Bayesian Inference, with a specific focus on determinstic Variational Approximation. Chapter 2 will briefly review and explain the details of the methodlogy in previous work such as the lasso problem, Approximate Bayesian Inference algorithm, MCMC(Monte Carlo Method), Bayesian Expectation Maximization algorithm and their variants and Mean-Field Variational Bayes(MFVB). We will present our main methodology of variational algorithm in Chapter 3, followed by a comprehensive experiment for testing the effectiveness of algorithm in Chapter 4.

Chapter 2

Literature Review

2.1 Bayesian Inference Paradigm

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

Bayesian Inference Intuition Bayesian inference approach stems from the Bayes rule, which is defined as Equation (1.7) based on theory developed by Beech et al. (1959). Suppose θ is our model parameter of interest, $\mathcal{D}$ is data, then $p(\theta)$ is known as prior distribution, which offers pre-existing knowledge or information about θ. Posterior distribution $p(\theta|\mathcal{D})$ refers to the likelihood conditioning on the data $\mathcal{D}$. Incorporating information from current data and prior knowledge, posterior distribution can be then inferred and simplified to Equation (2.1) since $p(\mathcal{D})$ is equal to constant and is also insignificant to know overall posterior distribution.

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

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

2.2.1 Lasso penalty formulation

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

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

In order to transform constraint form of lasso to penalty form, Lagranage multiplier method, as a pivotal technique from transforming a constraint optimization system into an unconstrained penalty formulation of system has been used. The Lagrangian function for constrained Lasso Regression is constructed by Equation 2.3

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

Since the objective function contains a quadratic term $||y - X\beta||_2$ with a linear term $\lambda||\beta||_1 - \lambda t)$, leading to a convex optimization problem. Due to strong duality theorem in convex optimization system, therefore the penalty formulation of lasso regression can be deduced as Equation 1.3, is equivalent to constraint form 2.2 after ignoring the unaffected constant $-\lambda t$.

Graphical demonstration of the lasso for Equation 2.2 and Equation 1.3 can also be found on the left hand side of the Figure 2.1, where the squared constraint set is drawn, in addition to the contour line of regression coefficient. Given that λ is set as penalty term that control the strength of penalization, larger penalization facilitate a more sparse solution, so that further enclosing the estimated coefficient to lies on the axis of each parameter as shown in Figure 2.1. Lasso regression coefficient would have higher chance to render the contour line of β intersect with the corner of the squared constraint set, causing the occurrence of sparse estimated regression coefficient. Compared with the ridge regression

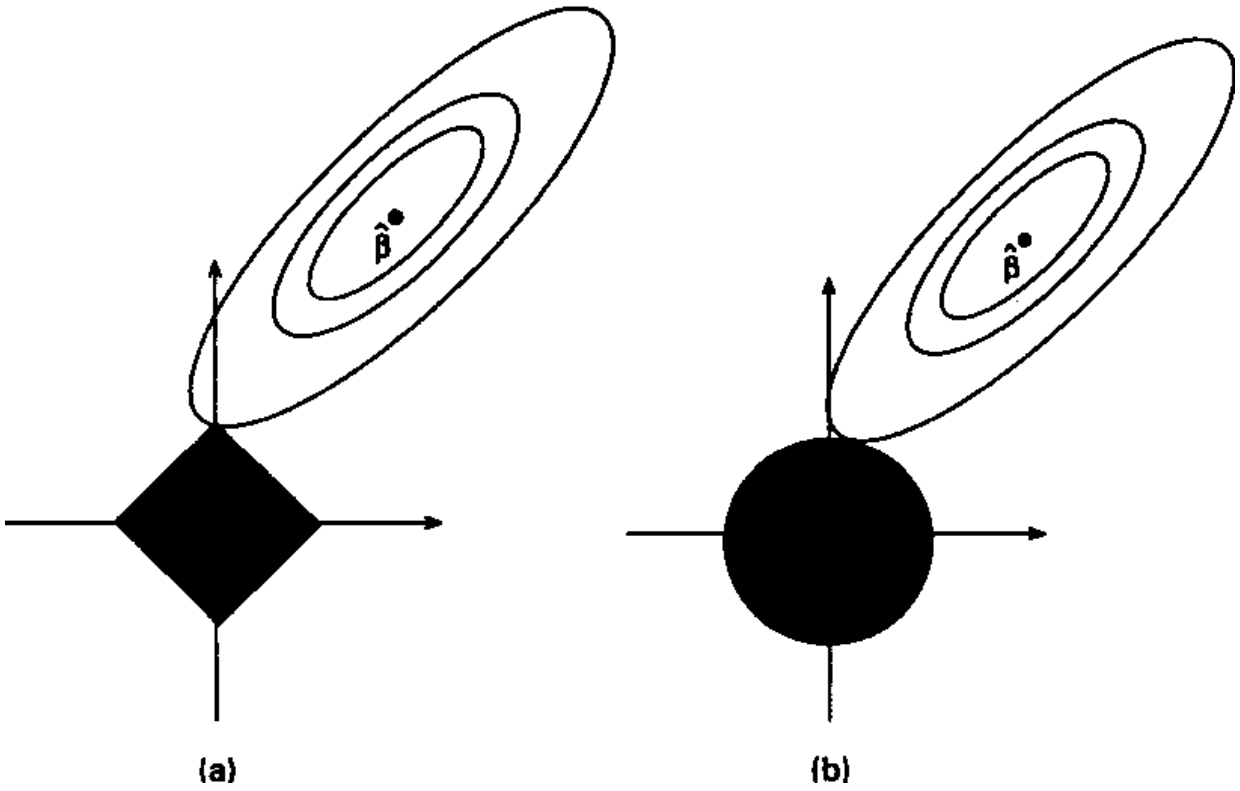


Figure 2.1: Graphical comparison between lasso regression and ridge regression

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

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

2.3 Bayesian Lasso

2.3.1 Bayesian Lasso model

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

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

For a given variance, the mode of posterior form in Equation(2.5) is consistent with the estimate of lasso equation in 1.3, but it will hinder the bayesian interpretation, inference

and variable selection since the bayesian predictive distribution make future inference via a posterior mean instead of posterior mode. In addition, if a variance is unknown, the posterior will be a multimodal distribution, the derivation has been provided by the appendeix from Park and Casella (2008).

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

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

To remedy this issue, a conditional Laplacian prior from 2.6 with respect to Equation(2.4) has been designed, ensuring the unimodality of the posterior for β, and the current prior with respect to β, σ^2 can be written in

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

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 2.7, given an improper prior selection for $\pi(\sigma^2) = \frac{1}{\sigma^2}$ and $\lambda \geq 0$ Additionally, an additional latent variable τ are introduced as a scale mixture of Gaussians for reformulation of conditional prior 2.6 as 2.8, which can be regarded as corresponding weight assigned to each regression coefficient. If τ_j goes to 0 then the corresponding regression coefficient will be shrunk towards zero accordingly.

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

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

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

.

2.3.2 Bayesian Lasso Gibbs Sampler

Gibbs Sampler

? introduced an special case of Metropolis-Hastings algorithm called Gibbs sampler, as a Markov Chain Monte Carlos sampling algorithm it can used for efficient sampling of any probability density function, given the posterior form from corresponding conditional distribution. For each iteration, each parameter of interest will be sampled once by conditional distribution for the current iteration. After reasonable burn-in period, it will return posterior distribution samples with descent approximation accuracy after discarding samples generated before burn-in period. After completion, the combination of each component of samples can formulate full samples of posterior distribution. As a result, the functional form of conditional distribution given any other parameter of interest has to be acquired. For the bayesian lasso model, given the parameter of interest: (β, σ^2, τ). This means the following exact functional form has to be derived

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

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

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

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

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

as

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

Formulation of β:

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

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

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

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

Formulation of σ^2:

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

Formulation of τ_j^2:

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

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

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

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

The gibbs sampler can be established by the following algorithm.

Algorithm 1 Gibbs Sampler for the Bayesian Lasso

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

Automatic selection of the penalty parameter λ Common choice of penalty parameter λ in the non-bayesian paradigm involves cross-validation approach, which is time-consuming and computational challenging especially for large datasets. Park and Casella (2008) arrange a hyperprior to λ^2 from Gamma distribution as 2.17, instaed of λ to facilitate conjugacy. According to Park and Casella (2008), there are some additional notification of choosing prior, which involves: firstly, to avoid mixing issue, the prior distribution for λ^2 should reach zero asymptotically with a descent speed as λ^2 goes to infinity, Secondly, the density at maximum likelihood estimate should be assigned with enough probability density with a overall flat distribution.

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

The penalty parameter is the extent of penalization of non-zero coefficient, which is also a compromise between model simplicity and fitting capability to data in the frequentist lasso setting. According to the posterior form of τ_j, λ controls the shape of generalized inverse-Gaussian posterior distribution of τ_j as shown in the 2.16.

To obtain the posterior form of λ, we need to incorpoate a proper hyperprior distribution to the joint distribution $p(y, \beta, \sigma^2, \tau)$ first, assuming the gamma distribution has shape θ and rate parameter γ.

$$\begin{aligned} p(\lambda|\tilde{y}, \beta, \sigma^2, \tau^2) &\propto p(\tilde{y}, \beta, \tau^2, \sigma^2, \lambda) \\ &= (\prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2})(\lambda^2)^{\gamma-1} e^{-\delta\lambda^2} \\ &= (\lambda^2)^{p+\gamma-1} e^{-\lambda^2(\frac{1}{2}\sum_{j=1}^{p}\tau_j^2+\delta)} \end{aligned} \tag{2.18}$$

Thus, the posterior distribution of λ is still following a Gamma distribution, with a shape parameter $p+\gamma-1$ and rate parameter $\sum_{j=1}^{p}\tau_j^2+\delta$. The λ^2 can be sampled by 2.18, based on using an augmented Gibbs sampler.

2.4 Expectation Maximization

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

2.4.1 Classical Expectation Maximization

The expectation maximization algorithm was proposed by Dempster et al. (1977), which is an iterative approach for seeking for the maximum likelihood estimate of parameter for probabilistic model that have missing data or the latent variables. The application of EM algorithm includes the inference of the parameters of the Gaussian Mixture model etc. The Expectation-Maximization algorithm involves two main steps, which are E-steps and M-steps. Suppose Z is the set of latent variable X is the set of entire set of observed variables, θ is parameter. t refers to the step during iteration, $\log(P(X, Z|\theta)))$ refers to complete

log-likelihood of data, and $\log(P(X|\theta))$ refers to incomplete log-likelihood of data without considering the hidden variables.

E-steps

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

$$Q(\theta, \theta^{(t-1)}) = E_{Z|X,\theta^{(t-1)}}[\log(P(X, Z|\theta))]. \tag{2.19}$$

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

$$\theta^{(t)} = \text{argmax}_\theta Q(\theta, \theta^{(t-1)}). \tag{2.20}$$

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

2.4.2 Bayesian Expectation Maximization

The Bayesian Expectation Maximization algorithm incorporates the idea of Expectation Maximization algorithm and Bayesian inference for estimation of the probabilistic model

when the data has missing or hidden. As opposed to the traditional EM approach, the Bayesian EM approach incorporate prior knowledge about the parameter, for the estimation of the posterior mode $p(\theta|\mathcal{D})$, considering the prior distribution as $p(\theta)$, the bayes rule can be written in the log scale

$$\ln p(\theta|\mathcal{D}) = \ln(p(\mathcal{D}|\theta)) + \ln(p(\theta)) - \ln(p(\mathcal{D})) \tag{2.21}$$

We can then further expand Equation 2.21 to Equation 2.22

$$\ln p(\theta|\mathcal{D}) = Q(\theta, \theta^{(old)}) + \mathrm{KL}(q||p(Z|X)) + \ln(p(\theta)) - \ln(p(\mathcal{D})) \tag{2.22}$$

, where $\mathrm{KL}(P||Q)$ is defined by Equation 1.9

2.4.3 Bayesian EM for Bayesian Lasso model

In order to deploy Bayesian EM algorithm to the bayesian lasso model for attaining the posterior mode, our purpose is to iteratively calcualte

$$\theta_1^{(t+1)} = \mathrm{argmax}_{\theta_1}[E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\log p(y, \theta_1, \theta_2)]]. \tag{2.23}$$

E-step

Using the same notation as before, firstly, the complete log-likelihood can be written by Equation 2.24

$$\log(p(\theta_1, \theta_2, \tilde{y})) \propto -\frac{n+p}{2}\log(\sigma^2) - \frac{||\tilde{y} - X\beta||_2^2}{2\sigma^2} - \frac{1}{2\sigma^2}\beta^T E_{\theta_2|\tilde{y},\theta_1^{(t)}}[D_\tau]\beta - \frac{b}{\sigma^2} \tag{2.24}$$

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

$$E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\log p(y, \theta_1, \theta_2)] = -\frac{n}{2}\log(\sigma^{2(t)}) - \frac{||y - X\beta^{(t)}||_2^2}{2\sigma^{2(t)}} - E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\sum_{j=1}^{p}\frac{\lambda^2\beta_j^2}{2\sigma^2\tau_j^2}] - (a+1)\log(\sigma^{2(t)})) - \frac{b}{\sigma^{2(t)}}. \tag{2.25}$$

Next, we need to take expectation of hidden variables: $E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\sum_{j=1}^{p}\frac{\lambda^2\beta_j^2}{2\sigma^2\tau_j^2}]$. After extracting constant with respect to θ_2, required formulation is $E_{\theta_2|\tilde{y},\theta_1^{(t)}}[\frac{1}{\tau_j}]$. Given the fact that $\tau_j^2|\sigma^2, \tilde{y}, \beta \sim GIG(1, \frac{\beta_j^2\lambda^2}{\sigma^2}, \frac{1}{2})$ and the special property of Generalized Inverse Gaussian distribution that if $X \sim GIG(a, b, p)$, then $\frac{1}{X} \sim GIG(b, a, -p)$, the distribution of $\frac{1}{\tau_j^2}$ can be

rearranged to $\frac{1}{\tau_j^2}|\sigma^2, \tilde{y}, \beta \sim GIG(\frac{\beta_j^2\lambda^2}{\sigma^2}, 1, -\frac{1}{2})$. However, taking expectation with respect to Generalized Gaussian Distribution is still complicated and require advanced mathematical operation and functional properties such as modified Bessel function. Thus, we can continue converting the distribution into a Inverse Gaussian distribution family, which render $\frac{1}{\tau_j^2}|\sigma^2, \theta_1, \tilde{y} \sim InverseGaussian(b^{-\frac{1}{2}}, 1)$. After continue rewriting Equation 2.25, the final conditional expectation form can be written as the following equation:

$$Q(\theta, \theta^{(t)}) = -(\frac{n}{2}+\frac{p}{2}+a+1)log(\sigma^2) - \frac{b}{\sigma^2} - \frac{||y - X\beta||_2^2}{2\sigma^2} - \frac{\lambda^2}{2\sigma^2}\sum_{j=1}^{p}(\beta_j^2 E[\frac{1}{\tau_j^2}]),\ E[\frac{1}{\tau_j^2}] = \frac{\sigma^{(t)}}{|\beta_j^{(t)}|\lambda} \tag{2.26}$$

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

M-step

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

$$\frac{\partial Q}{\partial \beta} = -\frac{1}{2\sigma^2}(-X^T y + 2X^T X\beta) - \frac{\lambda^2}{2\sigma^2}X^T X\beta = 0 \tag{2.27}$$

Rearranging the Equation 2.27, the updated formula for $\beta^{(t)}$ can be written in the following equation:

$$\beta^{(t)} = (X^T X + \lambda^2 A)^{-1} X^T y, \text{where } A = \text{diag}(\frac{\sigma^{(t-1)}}{|\beta^{(t-1)}|\lambda}) \tag{2.28}$$

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

$$\frac{\partial Q}{\partial \sigma^2} = -\frac{(\frac{n+p}{2} + a + 1)}{\sigma^2} + \frac{b + \frac{||y - X\beta||_2^2}{2} + \frac{\lambda^2(\beta^T A\beta)}{4}}{\sigma^4} = 0 \tag{2.29}$$

Rearranging the Equation 2.29, the updated formula for $\sigma^{2(t)}$ can be written in the following equation:

$$\sigma^{2(t)} = \frac{||y - X\beta^{(t)}||_2^2 + \lambda^2(\beta^{T(t)} A\beta^{(t)}) + 2b}{n + p + 2a + 2} \tag{2.30}$$

After completing the iteration process of bayesian lasso, the posterior mode of Bayesian Lasso posterior distribution can be extracted from β generated by 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 $A = \text{diag}(\frac{\sigma^{(t-1)}}{|\beta^{(t-1)}|\lambda})$ ▷ 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

  $A = diag(\frac{\sigma^{(t)}}{|\beta_j^{(t)}|\lambda})$ ▷ 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 statistician is to approximate over-complex probability density function in fast and efficient manner. Variational Inference severs as an effective alternative to MCMC algorithm especially for large datasets as mentioned in the last chapter. A proxy of exact posterior distribution can be fitted by addressing an optimization-based system. As the indispensable foundation of our proposed methodlogy, 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)

Under the Variational Inference setting, the purpose of it is to find a candidate approxiamation $Q(\theta) \in Q$ after specifying a specific family of posterior distribution Q that minimize the KL divergence to the exact posterior distribution as shwon in Equation 1.8 for each subelement of parameter θ. The complexity for finding optimal distribution relies heavily on the complexity of Q. Nevertheless, due to difficulty for computing marginal logarithm evidence $p(\mathcal{D}) = \int_\theta P(\mathcal{D}, \theta)$, as well as implicity dependency nature of $p(\mathcal{D})$ to KL divergence as explained in Equation 2.31, therefore, additional conversion is required for further

processing this optimization system, which transform Equation 1.8 to Equation 1.10.

$$\begin{aligned} KL(q(\theta||p(\theta|\mathcal{D}))) &= \mathbb{E}_{q(\theta)}[\log(q(\theta))] - \mathbb{E}_{q(\theta)}[\log(p(\theta|\mathcal{D}))] \\ &= \mathbb{E}_{q(\theta)}[\log(q(\theta))] - \mathbb{E}_{q(\theta)}[\log(p(\theta,\mathcal{D}))] + \log[p(\mathcal{D})]. \end{aligned} \tag{2.31}$$

KL divergence

We haven't introduced the KL-divergence formally even though this term is repeated several times. KL-divergence defined by Equation 2.31 is a distance metric for measuring the discrepancy of two probability distribution. Several important theoretical property includes non-negativity, and asymmetric property of $KL(q||p)$ and $KL(p||q)$.

ELBO

In addition, the definition of the Evidence Lower bound is defined by Equation 1.11, which is equivalent to the negative KL divergence despite of adding constant $p(\mathcal{D})$ with respect to $q(\theta)$. Apart from the equivalence of optimization system, the explanation of why it is called "Evidence lower bound" can be shown in Equation 2.32.

$$\log(p(\mathcal{D})) = KL(q(\theta)||p(\theta|\mathcal{D})) + ELBO(q(\theta)), \tag{2.32}$$

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

ELBO properties

2.5.3 The Mean-Field Variational Family

Motivated by statistical mean-field theory proposed by Parisi and Shankar (1988), the mean-field variational family is the most common choice that are easier to optimize with a tractable solution form. The mean field variational family represents a group of probability distributions employed in variational inference. Its objective is to estimate intricate, infeasible distributions, such as the genuine posterior distribution within a Bayesian model, by using more tractable and simpler distributions. The generic member of mean field variational family is as described in Equation 1.12, assuming mutual independence of each

target parameter θ_j, and the joint distribution is factorized as a product of individual distributions. Finally, no further assumption has been arranged to each individual distribution $q(\theta_i)$. The two dimensional visualization of mean-field variational family can be observed from Figure 1.2, while the contour of mean-field variational family member forms a concentric contour line over the optimization surface.

2.5.4 Coordinate ascent Variational Inference(CAVI)

After introducing the concept of ELBO and mean field variational family, the current goal is to find a optimum $q^*(\theta)$ to maximize the ELBO, and the most fundamental frequent choice for optimization algorithm is the Coordinate Ascent Variational Inference (CAVII) algorithm due to its simplicity in understanding and computational efficiency. Coordinate ascent is an optimization algorithm used to maximize or minimize a multivariate function. The main intuition behind coordinate ascent is to optimize the function with respect to one variable each time while fixing other variables fixed, iteratively update until convergence. The algorithm won't stop until either the differences between current function value and previous function value is less than a predefined threshold or when a maximum number of iterations is reached. Similarly, the CAVI works by iteratively maximizing each component of $q(\theta)$: $q_i(\theta_i)$, while maintaining other factor of distribution unchanged $q_{-i}(\theta_{-i})$, enforcing the final distribution can achieve a local optimum of the ELBO.

Derivation

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

$$\begin{aligned}
\text{ELBO}(q(\theta)) &= \int q(\theta) \log(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)} d\theta \\
&= \int \prod_i q_i(\theta_i)[\log[p(\mathcal{D},\theta) - \sum_i \log(q_i(\theta_i))]]d\theta \\
&= \int q_j(\theta_j)[\int \log p(\mathcal{D},\theta) \prod_{i\neq j} q_i d\theta_j]d\theta_j - \int q_j(\theta_j)\log[q_j(\theta_j)]d\theta_j + const \\
&= \int q_j(\theta_j)\log[\tilde{p}(\mathcal{D},\theta_j)] - \int q_j(\theta_j)\log[q_j(\theta_j)]d\theta_j
\end{aligned} \tag{2.33}$$

where $\log[\tilde{p}(\mathcal{D}, \theta_j)] = \mathbb{E}_{i \neq j}[\log[p(\mathcal{D}, \theta)]] + const$, and $\mathbb{E}_{i \neq j} = \int \log[p(\mathcal{D}, \theta)] \prod_{i \neq j} q_j(\theta_j) d\theta_j$ assuming optimizing the jth posterior parameter θ_j. Identifying the 2.33 is negative KL-divergence between $q_j(\theta_j)$ and $\tilde{p}(\mathcal{D}, \theta_j)$, the KL-divergence can be minimized when two distribution are same. 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,

$$\log[q_j^*(\theta_j)] = \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)] + const \tag{2.34}$$

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

$$q_j^*(\theta_j) = \frac{\mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)]}{\int \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta) d\theta_j} \propto \mathbb{E}_{i \neq j}[\log p(\mathcal{D}, \theta)] \tag{2.35}$$

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

Algorithm 3 Coordinate Ascent Variational Inference (CAVI)

1: Input: $p(\mathcal{D}, \theta)$, data $\mathcal{D}$, Initialize Variational factor $q_j(\theta_j)$
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 Mean Field Variational Bayes algorithm for the Bayesian Lasso problem.

Log likelihood

Assumption

Derivation

Chapter 3

Methodlogy

3.1 Lasso distribution

3.1.1 Univariate Lasso Distribution

Basic Property

Derivation

3.1.2 Multivariate Lasso Distribution

Basic Property

Derivation

3.2 Local-Global Algorithm

Chapter 4

Experiment Result and Analysis

4.1 Experimental Setting

4.1.1 Parameter selection

4.1.2 Evaluation metric

L1 accuracy

(MORE) Matrix norm Posterior cov and estimated Cov

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

Expectation: posterior mode sparse, posterior mean not sparse

4.1.3 Experimental datasets

toy dataset 3-4 datasets

4.2 Experimental Result

Chapter 5

Discussion and Conclusion

Bibliography

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.

Beech, D.G., Kendall, M.G., Stuart, A., 1959. The advanced theory of statistics. volume 1, distribution theory. Applied Statistics 8, 61. URL: https://doi.org/10.2307/2985818, doi:10.2307/2985818.

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.

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.

Rajaratnam, B., Sparks, D., 2015a. Fast bayesian lasso for high-dimensional regression .

Rajaratnam, B., Sparks, D., 2015b. Mcmc-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. URL: https://arxiv.org/abs/1508.00947.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

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.