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

The Lasso problem

The Bayesian Lasso Problem

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 under-estimating the variance when the correlation of variables becomes large.

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

In this thesis, we will firstly introduce lasso distribution, which is an invented distribution that could be matched for facilitating local parameter estimate, 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 univariate or multivariate lasso distribution so that global distribution would be obtained by product of local distribution and a 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 Shrinkage Operator(Lasso) regression proposed by Tibshirani (1996) belongs to one of the shrinkage methods, the main idea behind shrinkage method is that it will eliminate regression coefficient that are close to zero, by discarding the subset of them, the rest of the model might shares numerous advantages including interpretability and less prediction error than the model fitting by all predictors. As one of the most traditional shrinkage methods, Lasso regression has been proven for his success in Statistical Community over the years. It has also been deployed in Machine Learning Community as well, as another name called L_1 regularization techniques for effectively avoiding over-fitting problems and reducing model complexity. The main idea of lasso is adding a penalty term of absolute value of coefficients in addition to the sum of squared value of residuals. Due to squared constraint set shape of regression coefficient, lasso penalty regression will tend to have a higher chance to shrink the subset of the estimated regression coefficient to be zero, generating a more interprable model if discard them. Tibshirani (1996) has also shown that the effect of new submodel is competitive as opposed to other typical shrinkage methods such as subset selection and ridge regression.

Moreover, he also purpose that the lasso estimate can also be calculated under the Bayesian framework, if independent and identically distributed laplacian prior is assigned to the problem.

Introduction of Bayesian Lasso Problem(Why Bayesian Lasso) Furthermore, Park and Casella (2008) further explore the Lasso problem under the setting of Bayesian framework, where the choice of a conditional Laplace prior distribution over the regression coefficietn β given by standard error σ^2 is equivalent to the Lasso penalty formulation in the frequentist framework. It could also support a more stable automatic tuning process of choosing the most appropriate tuning parameter λ for the Lasso problem.

Challenges of Bayesian Lasso Problem

The use of Variational Inference

Approximation Algorithm: Stochastic type The most typical stochastic approximation algorithm is Markov Chain Monte Carlo(MCMC), where an exact posterior distribution can be sampled given form of conditional distribution of each model parameter conditioning on rest of the other parameters. Theoretically, MCMC will lead to an arbitary precise result of posterior distribution approximation if an arbitarily long burn-in period is allocated. However, MCMC is notorious for suffering from long execution time especially and heavy computational cost, and we will expand more the property in the following subsection 2.4.

Approximation Algorithm: Deterministic type On the other hand, determinstic approach has also arised as an faster substitution compared with stochastic approximation approaches. Numerous algorithms have been designed and utilized widely such as Variational Bayes, Expectation Propagation algorithms etc. Determinstic approaches assume the approximation originates from a tractable distribution first and attempt to search for the distribution from this family that is the closest to the target posterior distribution by optimization techniques, it has been indicated that Variational Inference algorithm demonstrate a descent computation cost and time-efficiency.

Variational Bayes The most traditional Variational Inference algorithm is known as Mean-Field Variational Bayes motivated by mean-field Theory in statistical physics yielded by Jordan et al. (1998) and Attias (1999), which assume the approximated distribution is from independent product of parameter distribution. Meanwhile, disadvantages of Variational Bayes include inexact approximation result under some scenarios. For example, it is suggested by Bishop (2006) that Variational Inference algorithm might underestimate the covariance between parameter of interest, if parameter of interest have a strong correlation. We will expand properties and derivation of Variational Inference more in subsection 2.5.

Motivation Motivated by the intention of further enhancing the approximation accuracy of Variational Bayes, we have designed two new Variational algorithms, particularly for Bayesian Lasso problem. By utilizing and fitting a Lasso distribution to marginal distribution, an improved estimate for global Gaussian Approximation can be obtained. Our contribution have been listed in the following subsection 1.2.

1.2 Contribution

Our main contribution could be concluded as the following part:

- Introduction of Lasso Distribution
- Derivation of properties for Univariate Lasso Distribution.
- Derivation of properties for Multivariate Lasso Distribution.
- 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 .

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 and fast Approximate Bayesian Inference such as Variational Approximation. Chapter 2 will briefly review and explain the details of the methodlgy in previous work such as the lasso problem, MCMC(Monte Carlo Method) and their variants and Mean-Field Variational Bayes(MFVB). We will present our main methodlogy of variational correction algorithm in Chapter 3, followed by a comprehensive experiment for testing the effectiveness of algorithm in Chapter 4. In Chapter 5, we will briefly discuss and explain our result and potential improvement in the future.

Chapter 2

Definition and Literature Review

2.1 Bayesian Inference Paradigm

Why Bayesian Bayesian Inference approaches shares numerous advantages in statistic community and applicationa reas, 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 (2.1) 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 $\mathcal{D}$.

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

Incorporating information from current data and prior knowledge, posterior distribution can be then inferred and simplified to Equation (2.2) since $p(\mathcal{D})$ is equal to constant and is also insignificant to the overall posterior distribution equation.

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

Challenges for Bayesian Inference Nevertheless, several disadvantages inference are still in the progress of Bayesian Inference. 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, 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.

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 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 2.3.

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

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 2.4

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

Since the objective function contains a quadratic term $||y - X\beta||_2$ with a linear term $\lambda||\beta||_1 - \lambda t$), leading to a convex optimization problem. Due to strong duality theorem in convex optimization system, therefore the penalty formulation of lasso regression can be deduced as Equation 2.5, is equivalent to constraint form 2.3 after ignoring the unaffected constant $-\lambda t$.

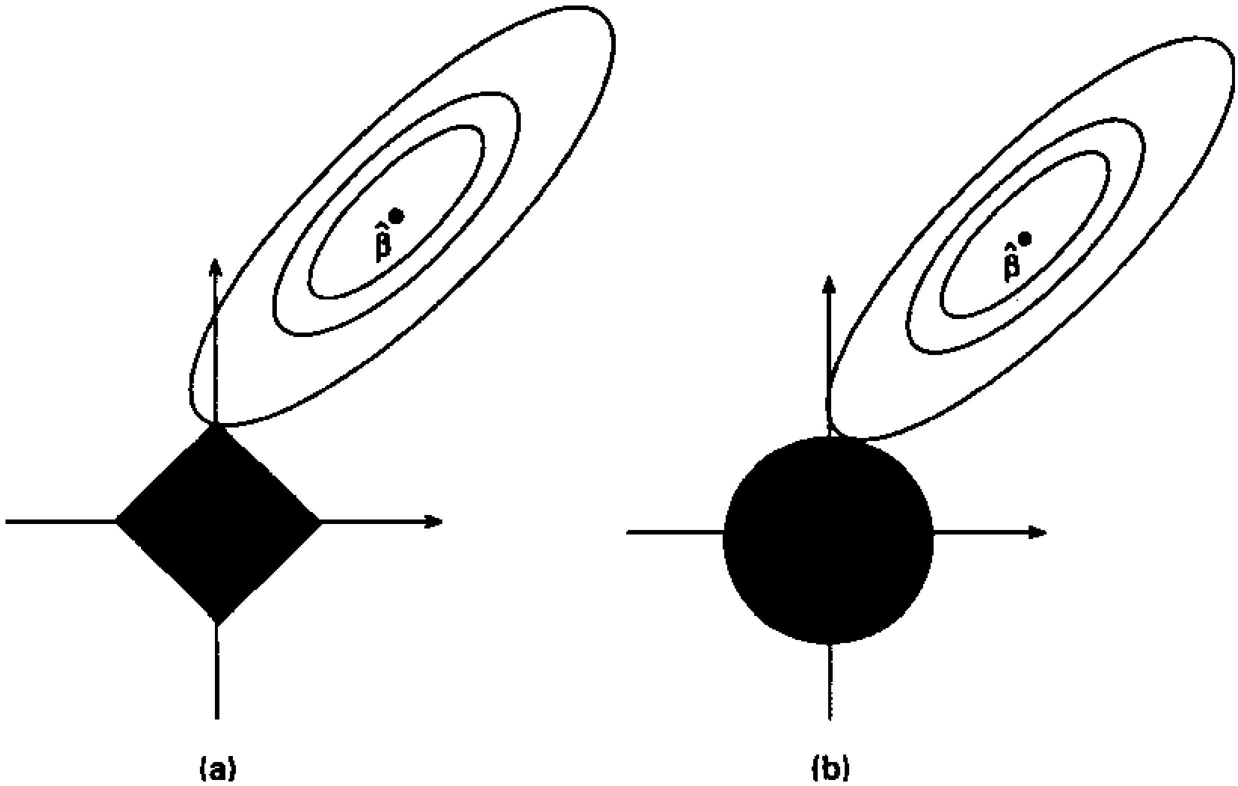

Figure 2.1: Graphical comparison between lasso regression and ridge regression

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2 + \lambda||\beta||_1, \lambda \geq 0. \tag{2.5}$$

Graphical demonstration of the lasso for Equation 2.3 and Equation 2.5 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 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 (2.6). In addition, lasso estimated can be efficiently computed via Least Angle Regression algorithm by

2.2.2 Bayesian Lasso regression

2.3 Expectation Maximization

2.3.1 Bayesian Expectation Maximization

2.4 Markov Chain Monte Carlo(MCMC)

2.4.1 Metropolis–Hastings (MH) Algorithm

2.4.2 Gibbs Sampler

2.5 Variational Inference

2.5.1 Mean Field Variational Bayes(MFVB)

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

$$q(\theta) = \prod_{i=1}^{n} q_i(\theta_i) \tag{2.6}$$

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

$$KL(q(x)||p(x|\mathcal{D})) \tag{2.7}$$

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

Attias, H., 1999. Inferring Parameters and Structure of Latent Variable Models byVariational Bayes. p. 21–30.

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.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1998. An introduction to variational methods for graphical models. Learning in Graphical Models , 105–161doi:10.1007/978-94-011-5014-9_5.

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.