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

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

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 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

Bayesian Inference approaches shares numerous advantages in statistic community, particularly for the circumstance when there is lack of data. An appropriate prior choice can be benefical in aforementioned case. Additionally, unlike frequentist inference approaches which treat parameter esimtate 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 parameter estimates.

Bayesian inference approach stems from the Bayes rule, which is defined as Equation (1.1) 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 $\mathcal{D}$.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{1.1}$$

Incorporating information from current data and prior knowledge, posterior distribution can be then inferred and simplified to (1.2) since $p(\mathcal{D})$ is equal to constant and is also insignificant to the overall posterior distribution equation.

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta), \tag{1.2}$$

Nevertheless, due to potential complexity for most of the posterior distribution form, intractable nature of exact posterior distribution form becomes one of the major obstacles in the progress of Bayesian Inference. 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.

The most typical stochastic sampling algorithm is Monte Carlo Markov Chain(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 when amount of data is large, and we will expand more the property in the following subsection 2.4.

On the other hand, determinstic approach such as Variational Bayes has also arised as an faster substitution compared with stochastic approximation approaches. 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. The most traditional Variational Inference algorithm is known as Mean-Field Variational Bayes motivated by mean-field Theory in statistical physics produced by Jordan et al. (1998) and Attias (1999), which assume the approximated distribution from independent product of parameter distribution Meanwhile, disadvantages of Variational Bayes include inexact approximation result under some circumstances. 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.

Motivated by the intention of furthernenhancing 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:

- Invention 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 univariate lasso distribution and multivariate lasso distribution

- 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 will briefly illustrate the motivation and background of variational approximation. Section 2 will briefly introduce basic definition and methodlogy in previous work such as MCMC(Monte Carlo Method) 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 Least Absolute Shrinkage and Selection Operator(LASSO) penalized regression

2.1.1 Lasso penalty formulation

2.1.2 Bayesian Lasso regression

2.2 Bayesian Paradigm

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

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.