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

Yuhao Li

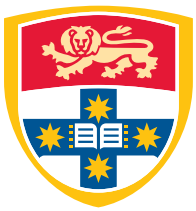
Supervisor: A/Prof. John Ormerod

A thesis submitted in partial fulfillment of

the requirements for the degree of

Bachelor of Science(Honour)(Data Science)

Mathematics and Statistics



THE UNIVERSITY OF SYDNEY

June 2023

## Statement of originality

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

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

Yuhao Li

# Abstract

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

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

On the other hand, obtaining the posterior of the Bayesian Lasso model involves using Monte Carlo Markov Chain method such as Gibbs sampler for exact posterior distribution. Even though MCMC is famous for its oracle property such as generating 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

**Contents** v

**List of Figures** 1

**List of Tables** 2

**1 Introduction** 1

1.1 Background and Motivation . . . 1

1.2 Contribution . . . 4

1.3 Thesis Organization . . . 5

**2 Definition and Literature Review** 6

2.1 Bayesian Inference Paradigm . . . 6

2.2 Least Absolute Shrinkage and Selection Operator(LASSO) penalized regression 7

2.2.1 Lasso penalty formulation . . . 7

2.2.2 Bayesian Lasso regression . . . 9

2.3 Expectation Maximization . . . 9

2.3.1 Bayesian Expectation Maximization . . . 9

2.4 Markov Chain Monte Carlo(MCMC) . . . 9

2.4.1 Metropolis–Hastings (MH) Algorithm . . . 9

2.4.2 Gibbs Sampler . . . 9

2.5 Variational Inference . . . 9

2.5.1 Mean Field Variational Bayes(MFVB) . . . 9

**3 Methodlogy** 10

3.1 Lasso distribution . . . 10

3.1.1 Univariate Lasso Distribution . . . 10

3.1.2 Multivariate Lasso Distribution . . . 10

3.2 Local-Global Algorithm . . . 10

**4 Experiment Result and Analysis** **11**

4.1 Experimental Setting . . . . . 11

4.1.1 Parameter selection . . . . . 11

4.1.2 Evaluation metric . . . . . 11

4.1.3 Experimental datasets . . . . . 11

4.2 Experimental Result . . . . . 11

**5 Discussion and Conclusion** **12**

**Bibliography** **13**

# List of Figures

2.1 Graphical comparison between lasso regression and ridge regression . . . . 8

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background and Motivation

**Introduction of Lasso Problem**

The Least Absolute Selection and Shrinkage Operator(Lasso) regression proposed by Tibshirani (1996) belongs to one of the shrinkage methods. As one of the most traditional shrinkage methods, Lasso regression has been proven for his success in Statistical Community over the years.

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

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

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

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 $\beta$ in

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

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

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 $\beta$, alternative improved algorithm have been purposed and deployed such as Least Angle Regression(LARS), iterative soft-thresholding, subgradient method, and iteratively reweighted least square(IRLS) by Efron et al. (2004), Beck and Teboulle (2009), Zhang and Zeng (2005) and Friedman et al. (2010).

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

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

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

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

**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 $\beta$ conditioning by standard error $\sigma^2$ is added to the Lasso penalty formulation in the frequentist framework to ensure unimodality of full posterior distribution. Based on closed form of

tractable posterior distribution, a three-step Gibbs sampler is proposed to draw approximate samples from Bayes Lasso posterior distribution, which can be utilized for further inference of parameter of interest.

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \tag{1.7}$$

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

**Challenges of Bayesian Lasso Problem** 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 inisde the interval of an acceptable tolerance, especially when $n$ is small and $p$ is large Rajaratnam and Sparks (2015).

**Approximation Algorithm: Deterministic type** Despite of the Bayesian Lasso

Gibbs Sampler, alternative methods such as determinstic Approximate Bayesian Infernece methods has becoming popular due to its fast speed and simple computaation. There are two genres of Approximate Bayesian Inferenece method, which includes stochastic approach such as MCMC, where an exact result can be obtained if infinte computataional resource is assigned. Another category lies in deterministic approach that 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. Deterministic 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 shows 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.

**Drawbacks of VI**

**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 methodlogy 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 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 (2.1) based on theory developed by Beech et al. (1959). Suppose $\theta$ is our model parameter of interest, $\mathcal{D}$ is data, then $p(\theta)$ is known as prior distribution, which offers pre-existing knowledge or information about $\theta$. Posterior distribution $p(\theta|\mathcal{D})$ refers to the likelihood conditioning on the data $\mathcal{D}$.

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

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

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

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 1.3, is equivalent to constraint form 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 the Figure 2.1, where the squared constraint set is drawn, in

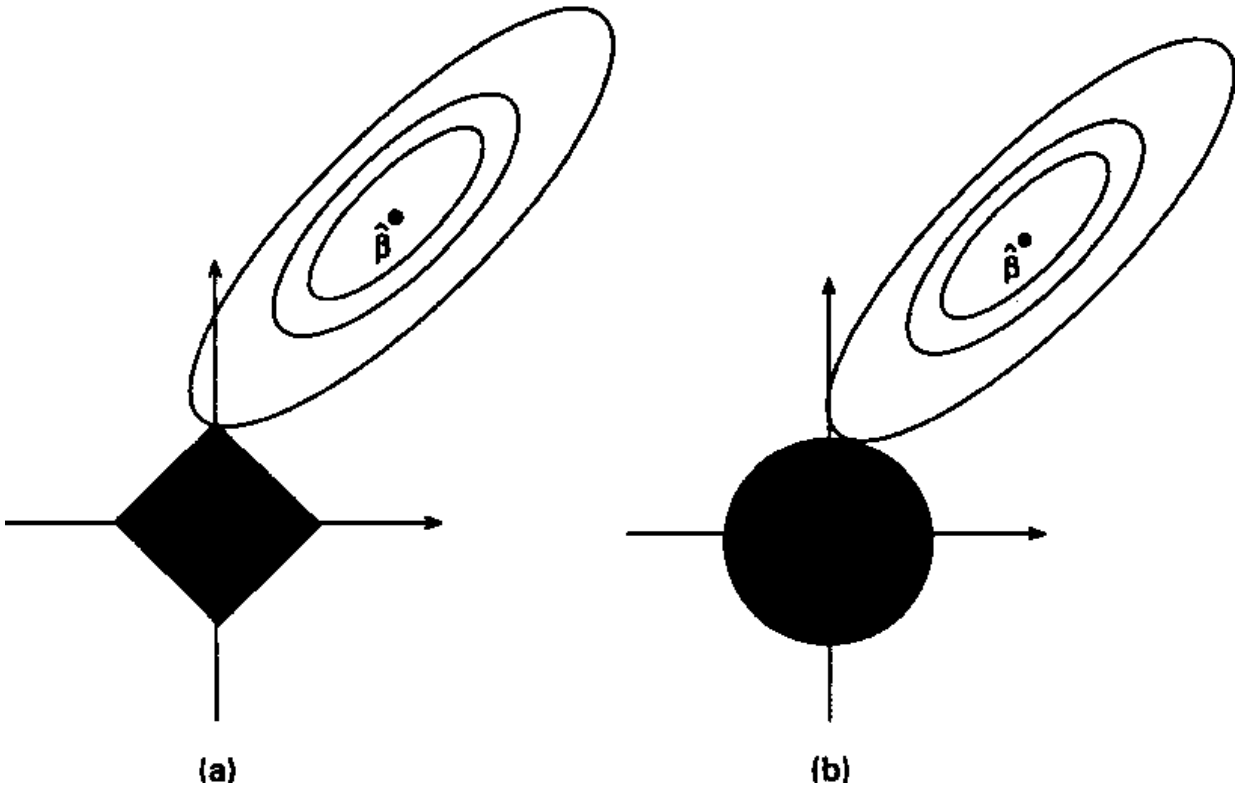



Figure 2.1: Graphical comparison between lasso regression and ridge regression

addition to the contour line of regression coefficient. Given that $\lambda$ is set as penalty term that control the strength of penalization, larger penalization facilitate a more sparse solution, so that further enclosing the estimated coefficient to lies on the axis of each parameter as shown in Figure 2.1. Lasso regression coefficient would have higher chance to render the contour line of $\beta$ intersect with the corner of the squared constraint set, causing the occurrence of sparse estimated regression coefficient. Compared with the ridge regression where a sum of square of penalty term is yielded instead on the right side of Figure 2.1, ridge regression tends to gain a non-sparse solution due to circled constraint set for $\beta$.

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

### 2.2.2 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.

$$q(\theta) = \prod_{i=1}^{n} q_i(\theta_i) \tag{2.5}$$

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

$$KL(q(x)||p(x|\mathcal{D})) \tag{2.6}$$

# 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.

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.

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.

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.

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

Park, T., Casella, G., 2008. The bayesian lasso. Journal of the American Statistical Association 103, 681–686. doi:10.1198/016214508000000337.

Rajaratnam, B., Sparks, D., 2015. 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.