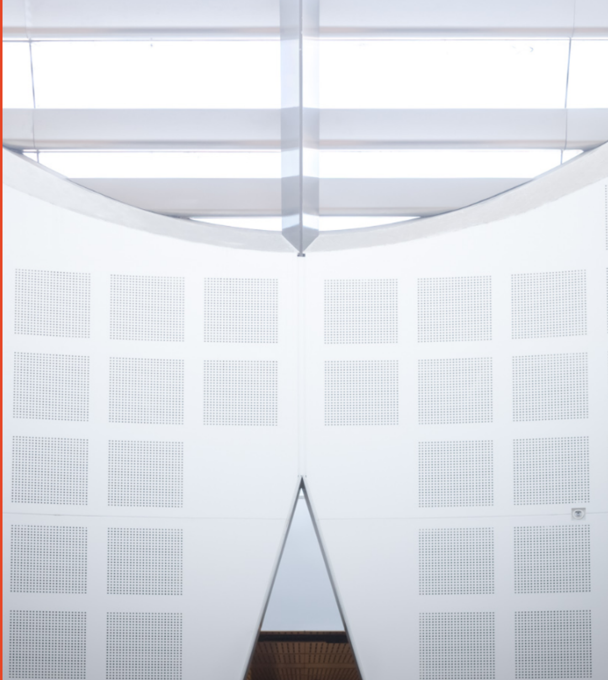
A novel algorithm for the Bayesian Lasso: A local approximation adjustment approach

Presented by: Yuhao Li

Supervised by: A/Prof. John Ormerod and Dr. Mohammad Javad Davoudabadi

May 5, 2023

THE UNIVERSITY OF SYDNEY



Motivation

- Lasso problem is important for coefficient estimation and regularization
- Uncertainty quantification of the Bayesian Lasso is important for inference
- Variational Approximation can also be a faster alternative to MCMC methods
- Variational Approximation is not accurate in certain cases
- Urgent need to find a more accurate and efficient algorithm for the Bayesian Lasso

Overview

- Lasso
- Bayesian Lasso
- Proposed Algorithm
  - Local-Global-Algorithm
  - Lasso Distribution
- Experiment
- Limitation and Future Work
- Conclusion

Lasso: Formulation

- Lasso (Least Absolute Shrinkage and Selection Operator), Invented by [1]
  regression analysis technique used for variable selection and regularization in linear regression models
- X is data matrix
- y is the standardized response variable
- λ is penalty parameter
- β is regression coefficient

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I). \tag{1}$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (y - X\beta)^T (y - X\beta) + \lambda ||\beta||_1, \quad \lambda \geq 0. \tag{2}$$

Lasso: Importance and Shortage

- Advantages
  - Feature selection: introduce sparsity for the model
  - Prevent overfitting: equivalent to l_1 regularization to produce a more generalized model
- Disadvantages
  - Can't capture the variance of the inferential quantity
  - No reliable method for obtaining suitable penalizing parameter λ

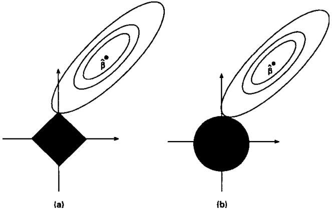


Figure 1: Graphical comparison between lasso regression and ridge regression

From Ordinary Lasso to Bayesian Lasso

- The Bayesian lasso uses a Laplace distributed prior on the β_j's to mimic the Lasso penalty:

$$f(\beta|\lambda) = \left(\frac{\lambda}{2}\right) \exp\left(-\lambda|\beta_j|\right). \tag{3}$$

- Park and Casella[2] introduce an hierarchical representation with the use of auxiliary variables τ_j^2 and an unimodal conditional prior that facilitates Gibbs Sampling

$$\pi(\beta|\sigma^2, \lambda) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \tag{4}$$

$$\int f(\beta|\tau^2) f(\tau|\lambda^2, \sigma^2) d\tau^2 = f(\beta|\lambda^2, \sigma^2). \tag{5}$$

Bayesian Lasso: Formulation

- Hierarchical representation

$$y|\mu, X, \beta, \sigma^2 \sim N_n(\mu + X\beta, \sigma^2 I)$$

$$\beta|\tau_1^2, ..., \tau_p^2 \sim N_p(0, \sigma^2 D_\tau)$$

$$D_\tau = diag(\tau_1^2, ..., \tau_p^2)$$

$$\tau_1^2 .., \tau_p^2 \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \tau_1^2, .., \tau_j^2 > 0$$

$$\sigma^2 \sim \pi(\sigma^2) = 1/\sigma^2, \sigma^2 > 0$$

Bayesian Lasso: Three-step Gibbs Sampler

- Gibbs sampling: A Markov chain Monte Carlo (MCMC) method that generates samples from the joint posterior distribution by iteratively sampling from the full conditional distributions of each parameter, given the current values of the other parameters.

$$p(\beta|y,\sigma^2,\tau^2): \text{MVN}((X^TX+\lambda^2A^{-1})^{-1}X^Ty,(X^TX+\lambda^2A^{-1})^{-1}\sigma^2), A=\text{diag}(\tau^2).$$

$$p(\sigma^2|y,\beta,\tau_j^2): \text{Inverse-Gamma}\left(\frac{n}{2}+\frac{p}{2}+a, \frac{||y-X\beta||_2^2}{2}+\frac{\lambda^2\sum_j\beta_j^2}{2\tau_j}+b\right).$$

$$p(\tau_j^2|y,\beta,\sigma^2): \text{GIG}\left(1, \frac{\beta_j^2\lambda^2}{\sigma^2}, 1/2\right).$$

Bayesian Lasso: Comparison

- Advantages
  - Improved coefficient estimates
  - better prediction accuracy
  - More reliable uncertainty quantification compared to the standard Lasso
  - Automatic selection of regularization parameter: λ
- Disadvantages
  - **Computational Expensive**

Bayesian Lasso: Mean Field Variational Bayes

- Mean-Field Variational Bayes assumption [3]

$$q(\theta) = q(\beta)q(\sigma^2)\prod_{i=1}^{p} q(\tau_j^2). \tag{6}$$

- Variational Inference Method

$$q^*(\theta) = \underset{q_\theta \in Q}{\operatorname{argmin}}\ \mathsf{KL}(q(\theta)||p(\theta|\mathcal{D})) \tag{7}$$

$$\mathsf{KL}(q||p(.|\mathcal{D})) = -\int q(\theta)\log\left(\frac{p(\theta)p(\mathcal{D}|\theta)}{q(\theta)}\right)d\theta + \log p(\mathcal{D}). \tag{8}$$

- Optimal solution

$$q_j^*(\theta_j) = \frac{\mathbb{E}_{i\neq j}[\log p(\mathcal{D},\theta)]}{\int \mathbb{E}_{i\neq j}[\log p(\mathcal{D},\theta)d\theta_j} \propto \mathbb{E}_{i\neq j}[\log p(\mathcal{D},\theta)]. \tag{9}$$

Bayesian Lasso: MFVB Update Procedure

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

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

  - The update for σ^2 leads to

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

MFVB Method: Comparison with MCMC

- Advantages
  - Fast approximation
  - Scalability
- Disadvantages
  - Tends to underestimate variance if predictors have high correlation, leads to low approximation accuracy

Local-Global-Algorithm: Definition

- Local Approximation
  - Focuses on the approximation of the **marginal** posterior distribution
- Global Approximation
  - Aims to capture the overall shape of the **joint** posterior distribution
- Improvement based on MFVB:
  - MFVB: Global Approximation
  - Our Algorithm: Local Approximation $\rightarrow$ Global Approximation

Local-Global-Algorithm: Basic Setting

- Assumption:
  - Continue Mean Field Assumption

$$p(\beta, \sigma^2|\mathcal{D}) \approx q(\beta, \sigma^2) = q(\beta)q(\sigma^2). \tag{10}$$

  - $q(\beta) \sim N(\mu, \Sigma)$ approximates $p(\theta|\mathcal{D})$.
- Initial Input: Posterior parameter for MFVB: $(\widetilde{a}, \widetilde{b}, \widetilde{\mu}, \widetilde{\Sigma})$, λ from the posterior mean from Gibbs Sampler.
- Target parameter: β, Current parameter: β_j Other parameter β_{-j}, assuming independence of σ^2
- Goal: Corrected Global Posterior Parameters: $\widetilde{\mu}, \widetilde{\Sigma}$ for the Gaussian Approximation: $q^*(\beta) \sim N(\widetilde{\mu}, \widetilde{\Sigma})$

Local Likelihood Derivation

Similarly, under the Variational Inference setting, the local marginal log-likelihood can be written as:

$$\log(\mathcal{D}, \beta_j) = \mathbb{E}_{q(\beta_{-j}|\beta_j)}\left[\log\left(\frac{p(\mathcal{D}, \beta_j, \beta_{-j})}{q(\beta_{-j}|\beta_j)}\right)\right] + \mathsf{KL}(q(\beta_{-j}|\beta_j), p(\beta_{-j}|\mathcal{D}, \beta_j)) \quad (11)$$

where $s = \mu_{-j} - \Sigma_{-j,j}\Sigma_{j,j}^{-1}\mu_j$ and $t = \Sigma_{-j,j}\Sigma_{j,j}^{-1}$, $\widetilde{a}$ and $\widetilde{b}$ are VB parameters for σ^2, μ, Σ are VB parameters for β. Using (11) leads to:

$$p(\beta_j|\mathcal{D}) \propto p(\beta_j, \mathcal{D}) \sim \mathsf{Lasso}\left(\frac{\tilde{a}}{\tilde{b}}(y - X_{-j}s), \frac{\tilde{a}}{2\tilde{b}}(X_j^T X_j + X_j^T X_{-j}t), \frac{\lambda\Gamma(\tilde{a}+1/2)}{\Gamma(\tilde{a})\sqrt{\tilde{b}}}\right). \quad (12)$$

Univariate Lasso Distribution: Probability Density Function

Theorem

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

$$p(x) = Z^{-1} \exp(-\frac{1}{2}ax^2 + bx - c|x|) \tag{13}$$

where $a \geq 0, b \in \mathbb{R}, c \geq 0$, Z is normalizing constant.

- a and c can't be 0 at the same time
- When $a = 0$, lasso distribution will become a asymmetric Laplace distribution
- When $c = 0$, lasso distribution will become a normal distribution

Univariate Lasso Distribution: Graphical illustration

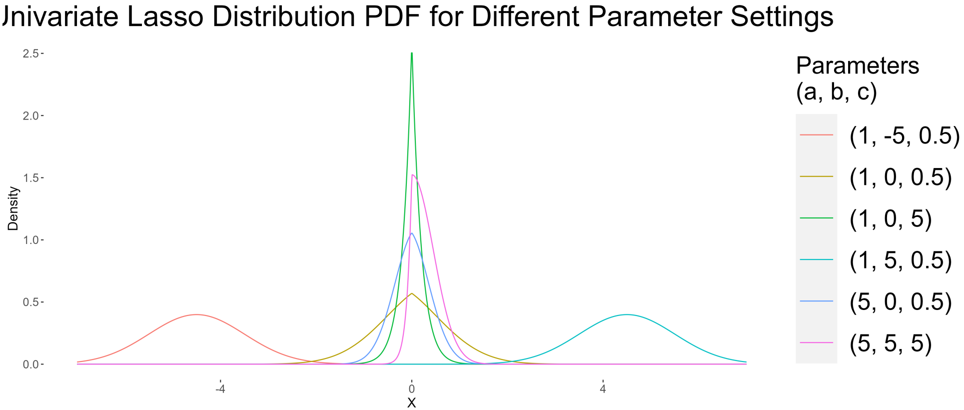


Figure 2: Visualization of Lasso Distribution PDF for different parameter setting

Univariate Lasso Distribution: Properties

- Normalizing constant: $Z(a,b,c) = \sigma\left[\frac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)} + \frac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)}\right]$, where $\mu_1 = (b-c)/a$, $\mu_2 = -(c+b)/a$ and $\sigma^2 = 1/a$.
- Moments: $E(x^r) = \frac{\sigma}{Z}\left[\frac{\Phi(\mu_1/\sigma)}{\phi(\mu_1/\sigma)}\mathbb{E}(A^r) + (-1)^r\frac{\Phi(\mu_2/\sigma)}{\phi(\mu_2/\sigma)}\mathbb{E}(B^r)\right]$, where $A \sim TN_+(\mu_1,\sigma^2)$, $B \sim TN_+(\mu_2,\sigma^2)$ and TN_+ is denotes the positively truncated normal distribution.
- Variance of univariate lasso distribution can be computed by: $\mathbb{V}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$.

Local Approximation Adjustment

Lastly, for each variable β_j we calculate the mean and variance of its corresponding Lasso distribution:

$$\mu_j^* = \mathbb{E}[\beta_j|\mathcal{D}]. \tag{14}$$

$$\Sigma_{jj}^* = \mathbb{V}[\beta_j|\mathcal{D}]. \tag{15}$$

The conditional distribution $q(\beta_{-j}|\beta_j)$ for any jth variable can be derived by $q(\beta_{-j}|\beta_j) \propto q(\beta)$, resulting another multivariate normal distribution with dimension of $p-1$ as shown in (16)

$$q(\beta_{-j}|\beta_j) = N_{p-1}(\mu_{-j} + \Sigma_{-j,j}\Sigma_{j,j}^{-1}(\beta_j - \mu_j), \Sigma_{-j,j}\Sigma_{-j,-j}^{-1}\Sigma_{j,j}). \tag{16}$$

Global Approximation Propagation

Using (14), (15), and (16), using a normal pdf with lasso mean and lasso variance to propagate the global mean and global covariance via:

$$q^*(\beta) = q(\beta_{-j}|\beta_j)\phi(\beta_j; \mu_j^*, \Sigma_{jj}^*). \tag{17}$$

Using (16) leads to $q^*(\beta) = N(\widetilde{\mu}, \widetilde{\Sigma})$ where $\widetilde{\mu}$ and $\widetilde{\Sigma}$ can be updated via:

$$\widetilde{\mu} = \begin{bmatrix} \mu_j^* \\ \tilde{\mu}_{-j} + \tilde{\Sigma}_{-jj}\tilde{\Sigma}_{jj}^{-1}\left(\tilde{\mu}_j^* - \tilde{\mu}_j\right) \end{bmatrix}, \quad \text{and} \tag{18}$$

$$\widetilde{\Sigma} = \begin{bmatrix} \Sigma_{jj}^* & \Sigma_{jj}^*\tilde{\Sigma}_{jj}^{-1}\tilde{\Sigma}_{j-j} \\ \tilde{\Sigma}_{-jj}\tilde{\Sigma}_{jj}^{-1}\Sigma_{jj}^* & \tilde{\Sigma}_{-j-j} + \tilde{\Sigma}_{-jj}\tilde{\Sigma}_{jj}^{-1}(\Sigma_{jj}^* - \Sigma_{jj})\tilde{\Sigma}^{-1}\tilde{\Sigma}_{j-j} \end{bmatrix}. \tag{19}$$

1

Univariate Local-Global-Algorithm

- Input: data X, response variable y, parameter from MFVB $(\tilde{a}, \tilde{b}, \tilde{\mu}, \tilde{\Sigma})$, Penalizing parameter: λ

Algorithm 1 Univariate-Local-Global-Algorithm

```
while μ̃ is changing less than ϵ do
    for j = 1 to p do
        Get current Lasso Distribution Parameter (a, b, c)
        Update local mean and local variance: μ*_j = E[β_j|D], Σ*_jj = V[β_j|D]
        Correct global mean and global variance: μ̃, Σ̃
    end for
end while
return μ̃, Σ̃
```

Line 4: $\mu_j^* = \mathbb{E}[\beta_j|\mathcal{D}], \Sigma_{jj}^* = \mathbb{V}[\beta_j|\mathcal{D}]$; line 5: $\widetilde{\mu}$, $\widetilde{\Sigma}$; line 8: $\tilde{\mu}, \tilde{\Sigma}$

Experiment Setup: Dataset Description

Dataset Name	n: number of samples	p: number of predictors
Hitters	263	20
Kakadu	1828	22
Bodyfat	250	15
Prostate	97	8
Credit	400	11
Eyedata	120	200

Table 1: Number of observations and predictors of different datasets

Experiment Setup: Evaluation Metric

- l_1 norm accuracy

$$l_1(f,g) = \int |f(x) - g(x)| dx \tag{20}$$

$$\text{Acc}(f,g) = 1 - \frac{1}{2} l_1(f,g) \tag{21}$$

  - Emphasize the accuracy measuring the center of distribution rather than the tail distribution
- Running Speed: Total number of time (second) used for generating posterior density

Experiment Result: Approximation Accuracy

Mean Accuracy(%)	MCMC	VB	LG_Local	LG_Global
Hitters	100	94.2	**99.3**	97.1
Kakadu	100	98.6	**99.4**	98.8
Bodyfat	100	97.0	**99.2**	97.2
Prostate	100	97.5	**99.6**	98.7
Credit	100	97.9	**99.7**	99.6
Eyedata	100	88.9	**98.7**	91.8

Table 2: Average approximation accuracy result on 6 datasets

Experiment Result: Approximation Speed

Running Speed(s)	MCMC	VB	LG_Local	LG_Global
Hitters	453.75	**0.17**	0.17	0.17
Kakadu	6696.56	**0.14**	0.19	0.19
Bodyfat	398.59	**0.14**	0.17	0.17
Prostate	336.31	**0.11**	0.12	0.12
Credit	359.92	**0.10**	0.11	0.11
Eyedata	18144.7	**1.21**	1.72	1.72

Table 3: Average approximation speed (in seconds) result on 6 datasets

Approximation Density Visualization: Hitters

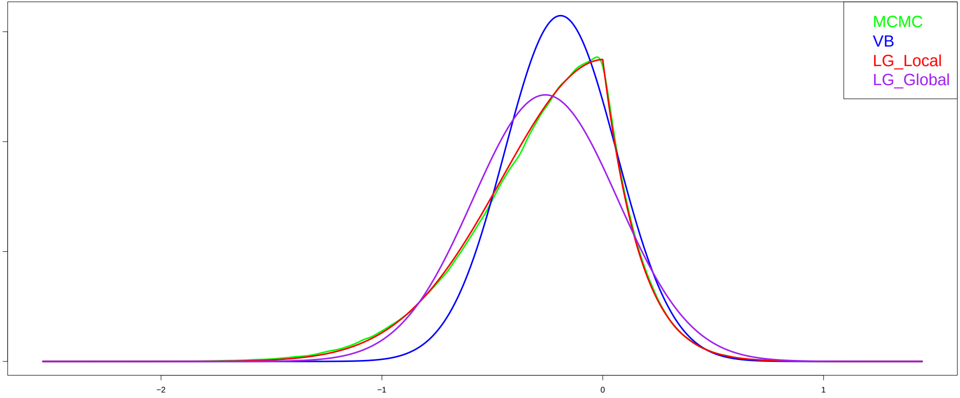


Figure 3: Part of Approximation Density for Hitters dataset

Approximation Density Visualization: Eyedata

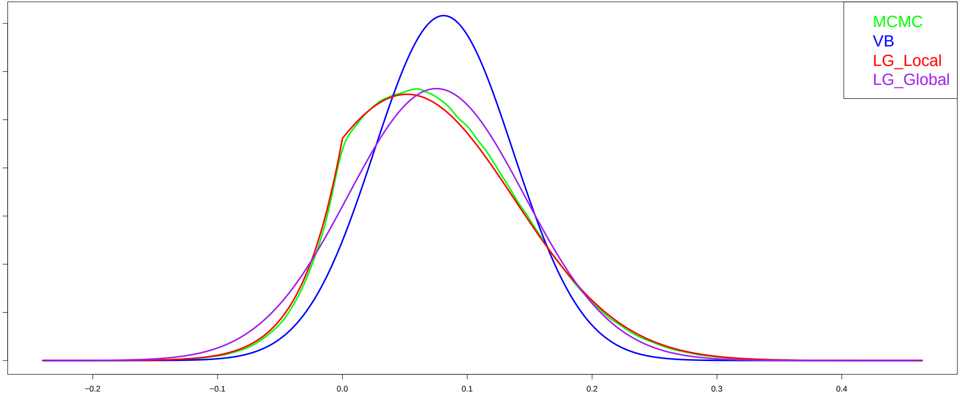


Figure 4: Part of Approximation Density for Eyedata dataset

Experiment Result: Discussion

- MFVB tends to produce a density with less variance, making it more concentrated at the center.
- Local-Global Algorithm The global posterior distribution is more accurate compared with MFVB.
- Local-Global Algorithm is highly accurate even when there is a high correlation between predictors.
- Local-Global Algorithm is highly accurate even when there are more predictors than a number of samples.

Limitation and Future Work

Limitations

- Automatic choice of λ is still obtained by Gibbs Sampling.
- More evaluation metrics can be used to further examine the superiority of the proposed method.
- The Univariate Local-Global Algorithm can't deal with the case when initial covariance is a diagonal matrix.

Future work

- Propose a Bivariate-Local-Global Algorithm to address the problem when the initial covariance is a diagonal matrix
- Derive the update formula of Σ

Contribution

Results

- Demonstrate superiority in Approximation Accuracy for surpassing all existing algorithms
- Demonstrate superiority in execution time of approximation efficiency even though a bit slower than MFVB
- Invent an univariate lasso distribution for better fitting the Bayesian Lasso posterior distribution

Reference

[1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[2] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[3] Giorgio Parisi and Ramamurti Shankar. Statistical field theory. 1988.