Understanding Hierarchical Representation of Bayesian Lasso[*]

Submitted by:

Arkonil Dhar [†‡]

Arkaprova Saha [§†]

Rachita Mondal [¶†]

Souvik Bhattacharyya [‖†]

Shreya Pramanik [**†]

Supervised by:

Dr. Arnab Hazra [†]

Submitted on:

21st april, 2022

[*]This report has been prepared towards the partial fulfillment of the requirements of the course *MTH535A: An Introduction to Bayesian Analysis.*

[†]Department of Mathematics & Statistics, Indian Institute of Kanpur, India.

[‡]201279, M.Sc. Statistics (Final year).

[§]201278, M.Sc. Statistics (Final year).

[¶]201374, M.Sc. Statistics (Final year).

[‖]201433, M.Sc. Statistics (Final year).

[**]201415, M.Sc. Statistics (Final year).

Abstract

Penalized regression is a popular approach for variable selection and model parameter estimation especially in high-dimensional regression problem. In this report we discuss the Bayesian hiearchical representation of Penalized regression. Using Gibbs sampler the final LASSO estimates has been obtained. We validate the method by some simulation studies and real data analysis.

Contents

1 Introduction **4**

2 Penalized Regression **4**
2.1 Generalization of lasso 5

3 Bayesian LASSO and its Hierarchical representations **5**
3.1 Hierarchical models 6
3.1.1 Original Lasso 6
3.1.2 Grouped Lasso 7
3.1.3 Fused Lasso 8
3.1.4 Elastic Net 10
3.2 Tuning Parameter Selection 12
3.3 Grouped LASSO 12
3.4 Fused LASSO 12
3.5 Elastic Net 12

4 Simulations **13**
4.1 Example 1 13
4.2 Example 2 14
4.3 Example 3 16

5 Real Data Analysis with Prostate Cancer Data **16**

6 Concluding Remark **18**

7 Supplementary Material **18**

8 Acknowledgements **19**

1 Introduction

Penalized regression is a popular approach for variable selection and model parameter estimation. When the sample size is small as compared to the number of predictor variables (i.e when $n << p$) then in most of the cases the problem of multicollinearity arises. Another important problem is to select a smaller subset of regressor from a large set so that only the relevant predictors are included in the model and also it is possible possible to have a better fit to the data. In other words, the main focus has been to select a sparse model with higher prediction accuracy. For this purpose several modifications have been introduced in Method of Ordinary Least Square.

Let us consider the linear regression model given by,

$$y = \mu 1_n + X\beta + \varepsilon \tag{1}$$

Here y is an $n \times 1$ vector, X is an $n \times p$ matrix. $\beta = (\beta_1, \ldots, \beta_p)'$. We consider $\varepsilon \sim N_p(0, \sigma^2 I_p)$.When $p >> n$, OLS fails to estimate β uniquely since the design matrix X has rank less than p. Specifically for this type of scenario Penalized regression has become popular.

In this report we will discuss the Bayesian hiearchical representation of Penalized regression. Our report is organized as follows:

In section (2) we discuss various penalization introduced in the literature of regression analysis.In Section (3) the bayesian hierarchies will be presented in details. Section (4) will be focusing on performing simulations to assess the performance of Bayesian LASSO in case of parameter estimation. After assessing the performances we will make applications on real data in Section (5). Finally, we conclude in Section (6).

2 Penalized Regression

Ridge Regression proposed by Hoerl and Kennard (1970), is a type of penalized regression that has been successful to remove multicollinearity. However, it can not produce a model with important predictors. Later in Frank and Friedman (1993) Bridge regression has been introduced. In this method no explicit form of parameter estimates are available. Again, here the Sum of Square due to Error (SSE) is minimized subject to $\sum_{i=1}^{p} \mid \beta_i \mid^{\gamma} \leq t$. Hence, in addition to choosing the tuning parameter it is important to choose an optimal γ to get reasonable parameter estimates.

One of the most popular and effective penalization technique is Least Absolute Shrinkage and Selection Operator (LASSO) which is able to perform both shrinkage and variable selection. LASSO was first proposed by Tibshirani (1996). This is a method to minimize SSE subject to a constraint which is non-differentiable and is expressed in terms of L_1 norm of the coefficient. LASSO has shown excellent performances in many situations. But Tibshirani (1996) mentioned that in $p > n$ case LASSO can not select more than n predictors. Also, if there exist an ordering of the feature variables LASSO fails to consider it. Next we discuss some generalizations and improvisations of LASSO.

2.1 Generalization of lasso

Let us suppose that $\hat{\beta}_L$ is the original LASSO estimate of the model (1), and it is given by ,

$$\hat{\beta}_L = \arg\min_{\beta}(y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i|$$

Here, X is a matrix of standardized regressor amd $\lambda > 0$ is the tuning parameter. As mentioned earlier LASSO fails to take care of the ordering in the feature variables. To recover the ordering limitation Tibshirani et al. (2005) proposed Fused LASSo. Fused LASSO estimate $\hat{\beta}_F$ is given by ,

$$\hat{\beta}_F = \arg\min_{\beta}(y - X\beta)'(y - X\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} |\beta_i - \beta_{i-1}|$$

Here, λ_1 and λ_2 are the two tuning parameters.
Yuan and Lin (2006) has proposed Grouped LASSO for the grouped variables. The corresponding estimator $\hat{\beta}_G$ is given by,

$$\hat{\beta}_G = \arg\min_{\beta}(y - \sum_{k=1}^{k} X_k\beta_k)'(y - \sum_{k=1}^{k} X_k\beta_k) + \lambda \sum_{k=1}^{K} ||\beta_k||_{G_k}$$

Here, K is the number of groups, β_k is the vector of βs corresponding to k-th group. $G_k = I_{m_k}$, where m_k is the number of coefficient vectors present in group G_k and thus $||\beta_k||_{G_k} = \sqrt{\beta' G_k \beta}$. This method is able to perform variable selection under group level.

Elastic Net has been introduced by Zou and Hastie (2005). The elasic net estimator is given by,

$$\hat{\beta}_{EN} = \arg\min_{\beta}(y - X\beta)'(y - X\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} |\beta_i|^2$$

This is a stabilized version of LASSO. It is also useful when $p >> n$. It has the ability to select a sparse model and it is also useful for multicollinear predictors.

3 Bayesian LASSO and its Hierarchical representations

Tibshirani (1996) has pointed out that the L_1 penalty in LASSO can be viewed as a Bayes posterior model under suitable setup of Laplace priors for $\beta_i's$. Park and Casella (2008) had proposed that the hierarchical representation of the full model using Laplace prior can be written as a mixture of normal with exponential mixing densities. In this report we will understand the construction of Group Lasso, Fused Lasso and Elastic Net along with the Original LASSO using the hierarchical representation.

3.1 Hierarchical models

3.1.1 Original Lasso

LASSO in Regression Analysis, is a method that uses regularization techniques to improve the model when we face overfitting problem. "LASSO" stands for Least Absolute Shrinkage and Selection Operator. For the original LASSO model, we take the choice of $h_1(\beta)$ and $h_2(\beta)$ as $\sum_{j=1}^{p}|\beta_j|$ and 0. The hierarchical model is as follows.

$$
\begin{aligned}
y\,|\mu, X, \beta, \sigma^2 &\sim N_n(\mu I_n + X\beta, \sigma^2 I_n)\\
\beta\,|\sigma^2, D_\tau &\sim N_p(0_p, \sigma^2 D_\tau),\\
\tau_1^2, \tau_2^2, ..., \tau_p^2 &\sim \prod_{j=1}^{p}\frac{\lambda^2}{2}e^{-\lambda\tau_j^2/2}d\tau_j^2,\ \tau_1^2, ..., \tau_p^2 > 0\\
\sigma^2 &\sim \pi(\sigma^2)d\sigma^2\ , \sigma^2 > 0
\end{aligned}
$$

To find $\pi(\beta\,|\sigma^2)$ we will need to multiply two pdfs and integrate out τ_1, ... τ_p as below.

$$
\pi(\beta\,|\sigma^2) = \int \pi(\beta\,|\sigma^2, \tau^2)\pi(\tau^2)d\underset{\sim}{\tau}
$$

and

$$
\begin{aligned}
\pi(\beta\,|\sigma^2, \underset{\sim}{\tau}^2) &= \frac{1}{\sigma^p\sqrt{\tau_1\tau_2\ldots\tau_p}^2}e^{-\frac{1}{2}\underset{\sim}{\beta}^T(D_\tau\sigma^2)^{-1}\underset{\sim}{\beta}}\\
&= \frac{1}{\sigma^p(\prod_{j=1}^{p}\tau_j)}e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{1}{\tau_j^2}\beta_j^2}\\
&= \frac{1}{\sigma^p\prod_{j=1}^{p}\tau_j}e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2}}
\end{aligned}
$$

$$
\begin{aligned}
\therefore \pi(\beta\,|\sigma^2) &= \int \frac{1}{\sigma^p\prod_{j=1}^{p}\tau_j}e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2}}\prod_{j=1}^{p}\frac{\lambda^2}{2}e^{-\lambda^2\tau_j^2/2}d\tau_j\\
&\propto \prod_{j=1}^{p}\int\frac{1}{\sqrt{2\pi}\sigma\tau_j}e^{-\frac{\beta_j^2}{2\sigma^2\tau_j^2}}e^{-\frac{\lambda^2\tau_j^2}{2}}\\
&= \prod_{j=1}^{p}\frac{\lambda}{2}e^{-\lambda|\beta_j|} \qquad (2)
\end{aligned}
$$

Note that, (2) has been obtained using the identity $\int_0^\infty\frac{1}{\sqrt{2\pi s}}exp\left(-\frac{z^2}{2s}-\frac{a^2 s}{2}\right)\frac{a^2}{2}ds = \frac{a}{2}exp\left(-a|z|\right)$.

Now to obtain samples, we need the full conditional posterior distributions to implement Gibbs Sampler.The posterior densities are given by:

$$\beta \mid \mu, \sigma^2, \tau_1^2, ..., \tau_p^2, \mathbf{X}, \mathbf{y} \sim N_p\big((\mathbf{X}'\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}\mathbf{X}'\tilde{\mathbf{y}}, \sigma^2(\mathbf{X}'\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}\big),$$

$$\frac{1}{\tau_j^2} = \gamma_j \mid \mu, \beta, \sigma^2, \mathbf{X}, \mathbf{y} \sim \text{inverse Gaussian}\left(\frac{\lambda^2\sigma}{|\beta_j|}, \lambda^2\right)\mathbf{I}(\gamma_j > 0), \text{ for } j = 1, ...p$$

$$\sigma^2 \mid \mu, \beta, \tau_1^2, ..., \tau_p^2, \mathbf{X}, \mathbf{y} \sim \text{inverse Gamma}\left(\frac{n-1+p}{2}, \frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\beta)'(\tilde{\mathbf{y}} - \mathbf{X}\beta)\right)$$

3.1.2 Grouped Lasso

For Group Lasso model, we take choices of $h_1(\beta)$ and $h_2(\beta)$ as $\sum_{k=1}^K ||\beta||_G$ and 0 respectively. In order to get conditional prior distribution $\pi(\beta \mid \sigma^2)$ we consider the hierarchical model as follows:

$$\begin{aligned} y \mid \mu, X, \beta, \sigma^2 &\sim N_n(\mu 1_n + X\beta, \sigma^2 I_n) \\ \beta_{G_k} \mid \sigma^2, {\tau_k}^2 &\overset{ind}{\sim} N_{m_k}(0, \sigma^2\tau_k^2 I_{m_k}) \\ \tau_k^2 &\overset{ind}{\sim} gamma\left(\frac{m_k+1}{2}, \frac{\sigma^2}{2}\right), k = 1, 2, \ldots, K. \end{aligned}$$

where we partition the β vector into K groups $G_1, G_2, ..., G_K$ of sizes $m_1, m_2, ..., m_K$ such that $\sum_{k=1}^K m_k$ =p. The vector of β_js in group $k (k = 1, 2, \ldots, K)$ is denoted by β_{G_k}.
To find $\pi(\beta \mid \sigma^2)$ we need to multiply two pdfs and integrate out $\tau_1^2, \tau_2^2, ..., \tau_K^2$ as below:

$$\begin{aligned} \pi(\beta \mid \sigma^2) &= \prod_{k=1}^K \int_0^\infty \frac{1}{(2\pi\sigma^2\tau_k^2)^{m_k/2}} \exp\left(-\frac{1}{2}\beta_k^T(\sigma^2\tau_k^2 I_{m_k})^{-1}\beta_k\right) \frac{(\frac{\lambda^2}{2})^{\frac{m_k+1}{2}}}{\Gamma(\frac{m_k+1}{2})}(\tau_k^2)^{\frac{m_k+1}{2}-1}\exp\left(-\frac{\lambda^2\tau_k^2}{2}\right)d\tau_k^2 \\ &= \prod_{k=1}^K \int_0^\infty \frac{1}{(2\pi\sigma^2\tau_k^2)^{m_k/2}} exp\left(-\frac{||\beta_{G_k}||^2}{2\sigma^2\tau_k^2}\right) \frac{(\frac{\lambda^2}{2})^{\frac{m_k+1}{2}}(\tau_k^2)^{\frac{m_k+1}{2}-1}}{\Gamma(\frac{m_k+1}{2})} \exp\left(-\frac{\lambda^2\tau_k^2}{2}\right)d\tau_k^2 \\ &= \prod_{k=1}^K \exp\left(-\frac{\lambda}{\sigma}||\beta_{G_k}||\right) \qquad (3) \\ &= \exp\left(-\frac{\lambda}{\sigma}\sum_{k=1}^K ||\beta_{G_k}||\right) \end{aligned}$$

The equation (3) can be easily obtained using the identity $\int_0^\infty \frac{1}{\sqrt{2\pi s}} exp\left(-\frac{z^2}{2s}\right) \frac{a^2}{2} exp\left(-\frac{a^2 s}{2}\right) 2ds = \frac{a}{2} exp\left(-a|z|\right)$. Now to obtain samples, we need the full conditional posterior distributions to implement Gibbs Sampler.The posterior densities are given by:

$$\beta_{G_k}\,|\beta_{-G_k},\sigma^2,\tau_1^2,.......,\tau_K^2,\lambda,X,\tilde{y}\sim N_p\left(A_k^{-1}X_k^T\left(\tilde{y}-\frac{1}{2}\sum_{k'\neq k}X_k'\beta_{G_{k'}}\right),\sigma^2A_k^{-1}\right)$$

$$1/\tau_k^2=\gamma_k\,|\beta,\sigma^2,\lambda,X,\tilde{y}\sim inverse\ Gaussian\left(\sqrt{\frac{\lambda^2\sigma^2}{||\beta_{G_k}^2||^2}},\lambda^2\right)\mathbf{I}(\gamma_k>0),$$

$$\text{for } k=1,2,...,K$$

$$\sigma^2\,|\beta,\tau_1^2,....,\tau_K^2,\lambda,X,\tilde{y}\sim inverted\ gamma\left(\frac{n-1+p}{2},\ \frac{1}{2}||\tilde{y}-X\beta||^2+\frac{1}{2}\sum_{k=1}^{K}\frac{1}{\tau_k^2}||\beta_{G_k}||^2\right),$$

where $\beta_{-G_k}=(\beta_{G_1},....,\beta_{G_{k-1}},\beta_{G_{k+1}},....,\beta_{G_k})$ and $A_k=X_k^TX_k+\left(\frac{1}{\tau_k^2}\right)\mathbf{I}_{m_k}$.

3.1.3 Fused Lasso

In this case to get the conditional prior distribution $\pi(\beta\,|\sigma^2)$ we can consider the following hierarchical model,

$$y\,|X,\beta,\sigma^2\sim N_n(X\beta,\sigma^2\mathbf{I}_n)$$

$$\beta\,|\sigma^2,\tau_1^2,...,\tau_p^2,\omega_1^2,...,\omega_{p-1}^2\sim N_p(0,\sigma^2\Sigma_\beta)$$

$$\tau_1^2,...,\tau_p^2\sim\prod_{j=1}^{p}\frac{\lambda_1^2}{2}e^{-\lambda_1^2\tau_j^2/2}d\tau_j^2,\ \ \tau_1^2,\dots,\tau_p^2>0$$

$$\omega_1^2,...,\omega_{p-1}^2\sim\prod_{j=1}^{p-1}\frac{\lambda_2^2}{2}e^{-\lambda_2^2\omega_j^2/2}d\omega_j^2,\ \ \omega_1^2,\dots,\omega_{p-1}^2>0$$

Here $\tau_1^2,...,\tau_p^2,\omega_1^2,...,\omega_{p-1}^2$ are mutually independent. Also the matrix Σ_β is given by,

$$\Sigma_\beta=\begin{pmatrix} d_1 & -\frac{1}{\omega_1^2} & 0 & 0 & \dots & 0 & 0\\ -\frac{1}{\omega_1^2} & d_2 & -\frac{1}{\omega_2^2} & 0 & \dots & 0 & 0\\ 0 & -\frac{1}{\omega_2^2} & d_3 & -\frac{1}{\omega_3^2} & \dots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & 0 & \dots & d_{p-1} & -\frac{1}{\omega_{p-1}^2}\\ 0 & 0 & 0 & 0 & \dots & -\frac{1}{\omega_{p-1}^2} & d_p \end{pmatrix}$$

here, $d_i=\frac{1}{\tau_i^2}+\frac{1}{\omega_{i-1}^2}+\frac{1}{\omega_i^2}$, for $i=1,2,...,p$ and $\frac{1}{\omega_0^2}=\frac{1}{\omega_p^2}=0$. Before getting the conditional prior let us first consider the pdf $\pi(\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2)$.

$$\pi(\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2) = \frac{1}{2\pi\sqrt{\mid\Sigma_\beta\mid}} exp\Big(-\frac{1}{2\sigma^2}\beta^T\Sigma_\beta^{-1}\beta\Big) \tag{4}$$

Here the Σ_β^{-1} has the form,

$$\Sigma_\beta^{-1} = \begin{pmatrix} d_1 & -\frac{1}{\omega_1^2} & 0 & 0 & \dots & 0 & 0 \\ -\frac{1}{\omega_1^2} & d_2 & -\frac{1}{\omega_2^2} & 0 & \dots & 0 & 0 \\ 0 & -\frac{1}{\omega_2^2} & d_3 & -\frac{1}{\omega_3^2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & d_{p-1} & -\frac{1}{\omega_{p-1}^2} \\ 0 & 0 & 0 & 0 & \dots & -\frac{1}{\omega_{p-1}^2} & d_p \end{pmatrix}$$

Hence we have,

$$\begin{aligned}\beta^T\Sigma_\beta^{-1}\beta &= \beta_1\left(d_1\beta_1 - \frac{\beta_2}{\omega_1^2}\right) + \beta_2\left(-\frac{\beta_1}{\omega_1^2} + d_2\beta_2 - \frac{\beta_3}{\omega_2^2}\right) + \beta_3\left(-\frac{\beta_2}{\omega_2^2} + d_3\beta_3 - \frac{\beta_4}{\omega_3^2}\right) + \dots + \beta_p\left(-\frac{\beta_{p-1}}{\omega_{p-1}^2} + d_p\beta_p\right) \\ &= \sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2} + \sum_{j=1}^{p-1}\frac{(\beta_{j+1}-\beta_j)^2}{\omega_j^2}\end{aligned}$$

Hence we have from equation (4),

$$\pi(\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2) = \frac{1}{2\pi\sigma\tau_1...\tau_p\omega_1...\omega_{p-1}} exp\Big(-\frac{\lambda_1^2}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2}\Big) exp\Big(-\frac{\lambda_2^2}{2\sigma^2}\sum_{j=1}^{p-1}\frac{(\beta_{j+1}-\beta_j)^2}{\omega_j^2}\Big)$$

Now we can write the joint distribution conditioning only on σ^2 as,

$$\pi(\beta, \tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2 \mid \sigma^2) = \pi(\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2)\pi(\tau_1^2, ..., \tau_p^2)\pi(\omega_1^2, ..., \omega_{p-1}^2) \tag{5}$$

We have to integrate (5) with respect to $\tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2$ to get $\pi(\beta \mid \sigma^2)$.
Note that,

$$\int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\omega_j^2}} exp\Big(-\frac{(\beta_{j+1}-\beta_j)^2}{2\sigma^2\omega_j^2}\Big)\frac{\lambda_2^2}{2} exp\Big(\frac{\lambda_2^2\tau_j^2}{2}\Big) d\tau_j^2 = \frac{\lambda_2}{\sigma} exp\Big(\frac{\lambda_2}{\sigma}\mid\beta_{j+1}-\beta_j\mid\Big) \tag{6}$$

$$\int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} exp\Big(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\Big)\frac{\lambda_1^2}{2} exp\Big(\frac{\lambda_1^2\tau_j^2}{2}\Big) d\tau_j^2 = \frac{\lambda_1}{\sigma} exp\Big(\frac{\lambda_1}{\sigma}\mid\beta_j\mid\Big) \tag{7}$$

Using the two identities (6) and (7), we get,

$$\pi(\beta\,|\sigma^2) \propto exp\Big(-\frac{\lambda}{2\sigma}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda}{2\sigma}\sum_{j=1}^{p-1}|\beta_{j+1}-\beta_j|\Big)$$

Lastly we need the full conditional posterior distributions to implement Gibbs Sampler to get samples. The posterior densities are given by,

$$\beta\,|\sigma^2,\tau_1^2,...,\tau_p^2,\omega_1^2,...\omega_{p-1}^2,X,\tilde{y} \sim N_p\left((X^TX+\Sigma_\beta^{-1})^{-1}X^T\tilde{y},\sigma^2(X^TX+\Sigma_\beta^{-1})^{-1}\right)$$
$$1/\tau_j^2\,|\beta,\sigma^2,\omega_1^2,...\omega_{p-1}^2,X,\tilde{y} \sim inverse\ Gaussian\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}},\lambda_1^2\right)\ \text{,for } j=1,2,...,p$$
$$1/\omega_j^2\,|\beta,\sigma^2,\tau_1^2,...\tau_{p-1}^2,X,\tilde{y} \sim inverse\ Gaussian\left(\sqrt{\frac{\lambda_2^2\sigma^2}{(\beta_{j+1}-\beta_j)^2}},\lambda_2^2\right)\ \text{,for } j=1,2,...,p-1$$
$$\sigma^2\,|\beta,\tau_1^2,...\tau_{p-1}^2,\omega_1^2,...\omega_{p-1}^2,X,\tilde{y} \sim inverted\ gamma\left(\frac{n-1+p}{2},\frac{1}{2}(\tilde{y}-X\beta)^T(\tilde{y}-X\beta)+\frac{1}{2}\beta^T\Sigma^{-1}\beta\right),$$

3.1.4 Elastic Net

In the case of Elastic Net model, we take the choice of $h_1(\beta)$ and $h_2(\beta)$ as $\sum_{j=1}^{p}|\beta_j|$ and $\sum_{j=1}^{p}|\beta_j|^2$. The hierarchical model is as follows.

$$y\,|\mu,X,\beta,\sigma^2 \sim N_n(\mu\mathbf{I}_n+X\beta,\sigma^2\mathbf{I}_n)$$
$$\beta\,|\sigma^2,D_\tau \sim N_p(0_p,\sigma^2D_\tau),$$
$$\tau_1^2,\tau_2^2,...,\tau_p^2 \sim \prod_{j=1}^{p}\frac{\lambda_1^2}{2}e^{-\lambda_1\tau_j^2/2}d\tau_j^2,\ \tau_1^2,...,\tau_p^2>0$$

The matrix D_τ is a diagonal matrix given by,

$$D_\tau = \begin{pmatrix} (\lambda_2+\tau_1^{-2})^{-1} & 0 & 0 & \dots & 0 \\ 0 & (\lambda_2+\tau_2^{-2})^{-1} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & (\lambda_2+\tau_p^{-2})^{-1} \end{pmatrix}$$

As the covariance matrix contains λ_2, β is not conditionally independent of λ_2.

The form of D_τ^{-1} is given by,

$$D_\tau^{-1} = \begin{pmatrix} (\lambda_2+\tau_1^{-2}) & 0 & 0 & \dots & 0 \\ 0 & (\lambda_2+\tau_2^{-2}) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & (\lambda_2+\tau_p^{-2}) \end{pmatrix}$$

We have,

$$\beta^T D_\tau^{-1} \beta = \sum_{i=1}^{p} \beta_i^2 (\lambda_2 + \tau_i^{-2})$$

Now using the above expression we get,

$$\pi(\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^p exp\left(-\frac{1}{2}\sum_{i=1}^{p}\beta_i^2(\lambda_2+\tau_i^{-2})\right)$$

To find $\pi(\beta \mid \sigma^2)$ we need to multiply two pdfs and integrate out τ_1, ... τ_p as below.

$$\begin{aligned} \pi(\beta \mid \sigma^2) &= \int_0^\infty \cdots \int_0^\infty \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^p exp\left(-\frac{1}{2}\sum_{i=1}^{p}\beta_i^2(\lambda_2+\tau_i^{-2})\right) \times \prod_{j=1}^{p}\frac{\lambda_1^2}{2} exp\left(-\lambda_1\tau_j^2/2\right) d\tau_1^2 ... d\tau_p^2 \\ &\propto exp\left(-\frac{\lambda_2}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\right) \times \prod_{i=1}^{p}\int_0^\infty exp\left(\frac{\beta_i^2}{2\sigma^2\tau_i^2} - \frac{\lambda_1^2\tau_i^2}{2}\right) d\tau_i^2 \\ &= exp\left(-\frac{\lambda_2}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\right) exp\left(-\sum_{i=1}^{p}\frac{\lambda_1|\beta_i|}{\sigma}\right) \end{aligned} \tag{8}$$

The equation (8) can be easily obtained using the identity $\int_0^\infty \frac{1}{\sqrt{2\pi s}} exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2} exp\left(-\frac{a^2 s}{2}\right) 2ds = \frac{a}{2} exp\left(-a|z|\right)$.

Now to get the full conditional posterior distributions to implement Gibbs Sampler to get samples the posterior densities are given by,

$$\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2, X, \tilde{y} \sim N_p\left((X^TX + D_\tau^{-1})^{-1}X^T\tilde{y}, \sigma^2(X^TX+D_\tau^{-1})^{-1}\right)$$

$$1/\tau_j^2 \mid \beta, \sigma^2, X, \tilde{y} \sim inverse\ Gaussian\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \lambda_1^2\right) \text{ ,for } j = 1,2,...,p$$

$$\sigma^2 \mid \beta, \tau_1^2, ... \tau_{p-1}^2, X, \tilde{y} \sim inverted\ gamma\left(\frac{n-1+p}{2}, \frac{1}{2}(\tilde{y}-X\beta)^T(\tilde{y}-X\beta) + \frac{1}{2}\beta^T D_\tau^{-1}\beta\right)$$

where D_τ is a diagonal matrix with diagonal elements $(\lambda_2+\tau_i^{-2})^{-1}, i = 1, ..., p$.

3.2 Tuning Parameter Selection

Previously we have discussed the hiearchical representation of Bayesian LASSO for given values of tuning parameters viz. λ_1 and λ_2. In section we will discuss how this tuning parameters can be obtained for each model. On of the approaches is Cross Validation. Park and Casella (2008) suggested alternative methods using the Gibbs Samplers. In this approach λ_1 and λ_2 are given appropriate hyperprior. For all the cases it is assumed that the tuning parameters will have a Gamma priors with the density,

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma r}(\lambda^2)^{r-1}e^{-\delta\lambda^2}, \quad (r > 0, \delta > 0)$$

Only exception is Elastic Net, where we assume two different Gamma priors for two hyper-parameters, $Gamma(r_1, \delta_1)$ and $Gamma(r_2, \delta_2)$. Next we can have the full conditional posterior for λ^2 and add it to the Gibbs Sampler.

3.3 Grouped LASSO

With a gamma(r, δ) prior, the full conditional distribution of λ^2 is

$$\pi(\lambda^2 \,|\beta, \sigma^2, \tau_1^2, \tau_K^2, \mathbf{X}, \tilde{\mathbf{y}}) \sim \text{gamma}\left(\frac{p+K}{2} + r, \frac{1}{2}\sum_{k=1}^{K} \tau_k^2 + \delta\right)$$

3.4 Fused LASSO

λ_1 and λ_2 are estimated with gamma(r, δ) priors the full conditional distributions of λ_1^2 and λ_2^2 are given by

$$\pi(\lambda_1^2 \,|\beta, \sigma^2, \tau_1^2, \tau_p^2, \omega_1, \ldots \omega_{p-1}, \lambda_1, \mathbf{X}, \tilde{\mathbf{y}}) \sim \text{gamma}\left(p + r, \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + \delta\right),$$

$$\pi(\lambda_2^2 \,|\beta, \sigma^2, \tau_1^2, \tau_p^2, \omega_1, \ldots \omega_{p-1}, \lambda_2, \mathbf{X}, \tilde{\mathbf{y}}) \sim \text{gamma}\left(p - 1 + r, \frac{1}{2}\sum_{j=1}^{p} \sigma_j^2 + \delta\right),$$

3.5 Elastic Net

In this case λ_1 and λ_2 are estimated with gamma(r_h, δ_h) $(h = 1, 2)$ priors, the full conditional distributions of λ_1^2 and λ_2^2 are given by

$$\pi(\lambda_1^2 \,|\beta,\sigma^2,\tau_1^2,\tau_p^2,\lambda_1,\mathbf{X},\tilde{\mathbf{y}}) \sim \text{gamma}\left(p+r_1,\frac{1}{2}\sum_{j=1}^{p}\tau_j^2+\delta_1\right),$$
$$\pi(\lambda_2^2 \,|\beta,\sigma^2,\tau_1^2,\tau_p^2,\lambda_2,\mathbf{X},\tilde{\mathbf{y}}) \sim \text{gamma}\left(\frac{p}{2}+r_2,\frac{1}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2+\delta_2\right),$$

4 Simulations

In this section we will perform simulations to assess the performances of the above discussed method. We have considered three different models and for each we have reported the following:

- avg. model MSE : The average Mean squared Error for the model
- SE MSE : Standard Error of MSE
- avg. est : Average estimated β
- beta MSE : Mean Squared error for the estimates of β

Each characteristic has been reported for Original LASSO , Fused LASSO and Elastic Net for Gibbs Sampling, Original LASSO and Elastic Net for LARS Algorithm.

4.1 Example 1

Here we draw samples of size $n = 200$. The true β has been chosen as $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. The error variance has been considered as $\sigma^2 = 9$. There are 8 explanatory variables and for each pair (x_i, x_j) the pairwise correlation is taken as $(1/2)^{|i-j|}$.

β	True β	Gibbs avg. est	Gibbs beta MSE	LARS avg. est	LARS beta MSE
β_1	3	2.88843254	0.06789391	2.958718804	0.05987366
β_2	1.5	1.48942617	0.06271175	1.533524716	0.06666565
β_3	0	0.01146561	0.04881304	-0.003085043	0.06152773
β_4	0	0.04821343	0.05171738	-0.001272813	0.06421990
β_5	2	1.90818802	0.06095377	2.006450930	0.05446258
β_6	0	0.06512050	0.04593849	0.041697334	0.05906808
β_7	0	-0.01785400	0.04538546	-0.013391531	0.05911670
β_8	0	-0.05849326	0.05869510	-0.079821734	0.07536810

Table 1: Performances of estimates of β for Original LASSO for Example 1

β	True β	Gibbs avg. est	Gibbs beta MSE	LARS avg. est	LARS beta MSE
β_1	3	2.9724794335	0.04675897	2.969114899	0.04748702
β_2	1.5	1.548365629	0.07195305	1.553292870	0.07435523
β_3	0	-0.0457361379	0.06741199	-0.043977773	0.06766152
β_4	0	0.0526840585	0.07307603	0.050663775	0.07407653
β_5	2	1.9373684633	0.06078204	1.942562537	0.06014575
β_6	0	0.0270271263	0.06761217	0.024836340	0.06751890
β_7	0	0.0009086687	0.08471850	0.001437407	0.08478037
β_8	0	-0.0189979571	0.08073539	-0.021996995	0.08196564

Table 2: Performances of estimates of β for Elastic Net for Example 1

β	True β	Gibbs avg. est	Gibbs beta MSE
β_1	3	2.950929768	0.05734161
β_2	1.5	1.433319210	0.07527121
β_3	0	0.022451684	0.05252007
β_4	0	0.048418237	0.05787989
β_5	2	1.926099125	0.06846568
β_6	0	0.060955441	0.06733308
β_7	0	-0.053352266	0.06930967
β_8	0	-0.004755552	0.05388037

Table 3: Performances of estimates of β for Fused LASSO for Example 1

Method	avg. MSE	SE MSE
Gibbs Original LASSO	8.719332	0.8148674
Gibbs Elastic Net	8.706339	0.8782833
Gibbs Fused LASSO	8.735206	1.002595
LARS Original LASSO	8.695225	0.8150475
LARS Elastic Net	8.706981	0.8781362

Table 4: Performances of MSE for different methods for Example 1

4.2 Example 2

Here we draw samples of size $n = 200$. The error variance has been considered as $\sigma^2 = 9$. There are 8 explanatory variables and for each pair (x_i, x_j) the pairwise correlation is taken as $(1/2)^{|i-j|}$. In this case the set up is exactly same as Example 1, except $\forall j, \beta_j = 0.85$.

β	True β	Gibbs avg. est	Gibbs beta MSE
β_1	0.85	0.8343420	0.07575995
β_2	0.85	0.8116247	0.07513768
β_3	0.85	0.8791123	0.04091025
β_4	0.85	0.7746213	0.05293517
β_5	0.85	0.8621825	0.04270752
β_6	0.85	0.7834219	0.07177262
β_7	0.85	0.8853076	0.06086652
β_8	0.85	0.7822373	0.06436606

Table 7: Performances of estimates of β for Fused LASSO for Example 2

β	True β	Gibbs avg. est	Gibbs beta MSE	LARS avg. est	LARS beta MSE
β_1	0.85	0.8235382	0.06958458	0.84102894	0.07413627
β_2	0.85	0.8130762	0.07485020	0.8452692	0.08165400
β_3	0.85	0.7842403	0.08507009	0.8034546	0.09137624
β_4	0.85	0.8145922	0.04311277	0.8323793	0.04740126
β_5	0.85	0.8386510	0.062551827	0.8528904	0.06801073
β_6	0.85	0.8516326	0.04538018	0.8631806	0.05006335
β_7	0.85	0.8200812	0.06199767	0.8380034	0.06692997
β_8	0.85	0.7721619	0.05544135	0.8162394	0.05763238

Table 5: Performances of estimates of β for Original LASSO for Example 2

β	True β	Gibbs avg. est	Gibbs beta MSE	LARS avg. est	LARS beta MSE
β_1	0.85	0.8589517	0.05804533	0.8582798	0.05978047
β_2	0.85	0.7819619	0.06395055	0.7832063	0.06649757
β_3	0.85	0.8465887	0.05999471	0.8466644	0.06003347
β_4	0.85	8455640	0.07551359	0.8454907	0.07586268
β_5	0.85	0.8620161	0.05788938	0.8632740	0.05994182
β_6	0.85	0.8662605	0.06213445	0.8664305	0.06223904
β_7	0.85	0.8340455	0.05657243	0.8342288	0.05643050
β_8	0.85	0.8446485	0.06452950	0.8439497	0.06436600

Table 6: Performances of estimates of β for Elastic Net for Example 2

Method	avg. MSE	SE MSE
Gibbs Original LASSO	8.79459	0.8613404
Gibbs Elastic Net	8.792605	0.8015767
Gibbs Fused LASSO	8.505563	0.8788934
LARS Original LASSO	8.783282	0.8604453
LARS Elastic Net	8.792605	0.8015767

Table 8: Performances of MSE for different methods for Example 2

4.3 Example 3

In this case number of predictors is 40. We have generated samples of size 500. Here, $\beta = (\mathbf{0}', \mathbf{2}', \mathbf{0}', \mathbf{2}')'$. $\mathbf{0}'$ and $\mathbf{2}'$ are the vectors of length 10 consisting of $0's$ and $2's$ respectively. $\sigma = 15$ and $corr(i, j) = 0.5$.

Method	avg. MSE	SE MSE
Gibbs Original LASSO	211.5532	11.9751
Gibbs Elastic Net	207.6556	13.55423
Gibbs Fused LASSO	209.1848	11.9562
LARS Original LASSO	210.7121	11.74765
LARS Elastic Net	207.6877	13.5559

Table 9: Performances of MSE for different methods for Example 3

From the above tables it can be observed that, the performance of Bayesian LASSO is at least good as LARS-LASSO. The average MSE and standard error of MSE over 50 replications corresponding to bayesian LASSO estimates are reasonably good. Hence this method is able to achieve high prediction accuracy. Also, in the simulations some of the parameters were intentionally set to zero to check the stability of the estimates around zero by a particular method. It has been observed in those cases that, the standard errors parameter estimates by bayesian LASSO is lower than LARS-LASSO. This implies that for the zero parameters bayesian LASSO gives more stable estimates. It makes the method suitable for variable selection problem in high-dimensional scenarios.

5 Real Data Analysis with Prostate Cancer Data

This data has been collected from Stamey et al. (1989). The data consists of seven variables. There are log(Cancer volume), log(prostate weight), log(benign prostatic hyperplasia amount), seminal vesicle invasion, log(capsular penetration), Gleason score and percentage Gleason scores. The independent variable is log(prostate specific antigen). The data set has been divided into training set and test set. The training set has 67 observations while the test set has 30 observations. The model has been fitted for the training data set. We try to assess the performance of Bayesian LASSO in comparison to LARS-LASSO. As a criterion to performance assessment we have selected prediction error i.e. the MSE. Lower the MSE better is the performance. In the following table we have reported the MSE for Gibbs original LASSO, Gibbs Elastic Net and Gibbs Fused LASSO. Additionally we have also reported LARS-Original LASSO and LARS Elastic Net. For each method the MSEs for Training set and that for Test set have been noted.

β	Gibbs Original LASSO	Gibbs Fused LASSO	Gibbs Elastic Net
β_1	0.571216677	0.417812141	0.571317834
β_2	0.647547548	0.256603566	0.643790226
β_3	-0.018192212	0.010063478	-0.017798607
β_4	0.138037225	0.061303117	0.137394252
β_5	0.743531465	0.049929558	0.722938404
β_6	-0.208560312	0.018662220	-0.203265676
β_7	0.011929335	0.0136660758	0.011012494
β_8	0.008765407	0.007980472	0.008727114

Table 10: Estimates corresponding to the Coefficients of Independent variables for Bayesian LASSO

β	LARS Original LASSO	LARS Elastic Net
β_1	0.576543185	0.576543185
β_2	0.614020004	0.614020004
β_3	-0.019001022	-0.019001022
β_4	0.144848082	0.144848082
β_5	0.737208645	0.737208645
β_6	-0.206324227	-0.206324227
β_7	-0.029502884	-0.029502884
β_8	0.009465162	0.009465162

Table 11: Estimates corresponding to the Coefficients of Independent variables for LARS-LASSO

From the above estimates of the parameters it can be observed that the estimates corresponding to Elastic Net for Bayesian LASSO have different structures than all the other estimates for Bayesian LASSO. The values are significantly different than the other two methods. Also note that, the estimates for LARS-LASSO are quiet similar to that of Gibbs Original LASSO and Gibbs Elastic Net. Now from the Train and Test MSE we try to assess how good the methods are in terms of fitting the data. We are following the general criteria that lower the MSE better is the method.

Method	Train MSE	Test MSE
Gibbs Original LASSO	0.4864719	0.5270230
Gibbs Fused Net	0.5929787	0.5558311
Gibbs Elastic Net	0.4398752	0.5181083
LARS Original LASSO	0.6233868	0.7427998
LARS Elastic Net	0.6233868	0.7427998

Table 12: MSE for different methods

In the upper section of the table we have included the results of Train and Test MSEs for Gibbs LASSO techniques and in the lower section we have reported the same for LARS-LASSO. looking at the two section it is clear that the MSEs corresponding to Bayesian LASSO is lower than that of LARS-LASSO. It gives enough indication that the performance of Bayesian LASSO is better than LARS-LASSO. Now, among the MSEs of Bayesian LASSO methods it can be seen that Elastic nets yields the lowest MSE for both Train and Test data. So, based on the above observations it can be concluded that for this data the performance of Gibbs Elastic Net is the best.

6 Concluding Remark

In this report we have considered the bayesian approach of Penalized regression analysis. We have also discussed the hierarchical representation of different modifications of bayesian LASSO. Using Gibbs sampler the final LASSO estimates have been obtained. We have validated the method by some simulation studies and real data analysis. From the simulations in section (4) we have seen that the performance of Bayesian LASSO is reasonably good and it performs at least as good as LARS-LASSO. In the section (5) we have concluded that Bayesian Elastic Net is the most suitable approach for analysing the Prostate cancer data.

7 Supplementary Material

For more details regarding the codes for simulations and real data analysis the readers are directed to the GitHub repository.

8 Acknowledgements

We want to express our gratitude towards Prof. Arnab Hazra for his continuous support and valuable suggestions in our project. The group members Arkaprova Saha, Rachita mondal and Shreya Pramanik have their contributions in understanding and explaining the theoretical point of view of the hierarchical representation of Bayesian LASSO as mentioned in the paper Casella et al. (2010). On the other hand the menbers Arkonil Dhar and Souvik Bhattacharyya have put their substantial efforts in implementing the methods via simulation and real data analysis.

References

Casella, G., Ghosh, M., Gill, J., and Kyung, M. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian analysis*, 5(2):369–411.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076–1083.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.