Understanding Hierarchical Representation of Bayesian Lasso[1]

A. Dhar* A. Saha * R. Mondal* S. Bhattacharyya* S. Pramanik *

*Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

April 20, 2022

[1] Main Reference: Casella et al. (2010)

Contents

Motivation

- Parameter estimation

Motivation

- Parameter estimation
- Model selection

Motivation

- Parameter estimation
- Model selection
- Multicollinearity in high-dimensional regression

Motivation

- Parameter estimation
- Model selection
- Multicollinearity in high-dimensional regression

Main Goal: Selecting a sparse model with higher prediction accuracy.

Problems with OLS Estimates in high dimensions

- Consider the regression model,

$$y = \mu 1_n + X\beta + \varepsilon$$

- y is an $n \times 1$ vector, X is an $n \times p$ matrix. $\beta = (\beta_1, \ldots, \beta_p)'$.

Problems with OLS Estimates in high dimensions

- Consider the regression model,

$$y = \mu 1_n + X\beta + \varepsilon$$

- y is an $n \times 1$ vector, X is an $n \times p$ matrix. $\beta = (\beta_1, \ldots, \beta_p)'$.

- The OLS estimate of β is given by,

$$\hat{\beta} = (X'X)^{-1}X'y$$

Problems with OLS Estimates in high dimensions

- Consider the regression model,

$$y = \mu 1_n + X\beta + \varepsilon$$

- y is an $n \times 1$ vector, X is an $n \times p$ matrix. $\beta = (\beta_1, \ldots, \beta_p)'$.

- The OLS estimate of β is given by,

$$\hat{\beta} = (X'X)^{-1}X'y$$

- When $p >> n$, OLS fails to estimate β uniquely since the design matrix X has rank less than p and so is $X'X$.

Penalized Regression

- Popular approach in high-dimensional regression.
- Creates a linear model by imposing penalization for having a large number of predictors.

Penalized Regression

- Popular approach in high-dimensional regression.
- Creates a linear model by imposing penalization for having a large number of predictors.
- Adds a constraint to the objective function, known as shrinkage.
- Some coefficients are reduced to zero.

Penalized Regression

- Popular approach in high-dimensional regression.
- Creates a linear model by imposing penalization for having a large number of predictors.
- Adds a constraint to the objective function, known as shrinkage.
- Some coefficients are reduced to zero.
- Useful for Model Selection.

Penalized Regression

Ridge Regression: $\hat{\beta}_R = \arg\min_\beta (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{p} \beta_i^2$

Bridge Regression: $\hat{\beta}_B = \arg\min_\beta (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{p} \beta_i^\gamma$

Penalized Regression

Ridge Regression: $\hat{\beta}_R = \arg\min_\beta (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{p} \beta_i^2$

Bridge Regression: $\hat{\beta}_B = \arg\min_\beta (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{p} \beta_i^\gamma$

- Ridge Regression can not produce a model with important predictors.
- In Bridge Regression no explicit form of parameter estimates are available.
- In addition to choosing the tuning parameter it is important to choose an optimal γ to get reasonable parameter estimates in Bridge method.

LASSO

- LASSO is able to perform both shrinkage and variable selection.
- LASSO estimate is given by,

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

LASSO

- LASSO is able to perform both shrinkage and variable selection.

- LASSO estimate is given by,

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

- In $p > n$ case LASSO can not select more than n predictors.

- If there exists an ordering of the feature variables LASSO fails to consider it.

- **Fused LASSO** estimate is given by,

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

Generalizations of LASSO

- **Fused LASSO** estimate is given by,

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

- **Grouped LASSO** estimate 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 $||\beta_k||_{G_k} = \sqrt{\beta' G_k \beta}$.

Generalizations of LASSO

- **Fused LASSO** estimate is given by,

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

- **Grouped LASSO** estimate 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 $||\beta_k||_{G_k} = \sqrt{\beta' G_k \beta}$.

- **Elastic Net** estimator is given by,

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

Generalizations of LASSO

- Fused LASSO allows sparsity of the coefficients and also that of their differences.
- Grouped LASSO is specifically useful for grouped variables.
- Elastic Net has been mentioned as the stabilized version of LASSO.
- All the aforementioned methods work well in $n << p$ case and they are also capable of model slection.

Bayesian LASSO

- 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.

Next, we present the construction of Group Lasso, Fused Lasso and Elastic Net along with the Original LASSO using the hierarchical representation.

Original LASSO I

Hierarchical Model:

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

$$\beta \,|\sigma^2, D_\tau \sim N_p(0_p, \sigma^2 D_\tau),\ D_\tau = diag(\tau_1^2, \ldots, \tau_p^2)$$

$$\tau_1^2, \tau_2^2, ..., \tau_p^2 \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda \tau_j^2/2} d\tau_j^2,\ \tau_1^2, ..., \tau_p^2 > 0$$

$$\sigma^2 \sim \pi(\sigma^2) d\sigma^2\ , \sigma^2 > 0$$

Conditional Prior:

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

Original LASSO II

Full Conditional Posterior

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

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

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

Grouped LASSO I

Hierarchical Model:

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

Conditional Prior:

$$
\pi(\beta|\sigma^2) = \exp(-\frac{\lambda}{\sigma}\sum_{k=1}^{K}||\beta_{G_k}||)
$$

Grouped LASSO II

Full Conditional Posterior

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

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

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

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

where β_{-G_k} = $(\beta_{G_1},....,\beta_{G_{k-1}},\beta_{G_{k+1}},....,\beta_{G_k})$ and A_k = $X_k^TX_k$ + $(\frac{1}{\tau_k^2})I_{m_k}$.

Fused LASSO I

Hierarchical Model:

$$
\begin{aligned}
y \mid X, \beta, \sigma^2 &\sim N_n(X\beta, \sigma^2 I_n) \\
\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2 &\sim N_p(0, \sigma^2 \Sigma_\beta) \\
\tau_1^2, ..., \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda_1^2}{2} e^{-\lambda_1^2 \tau_j^2/2} d\tau_j^2, \;\; \tau_1^2, \ldots, \tau_p^2 > 0 \\
\omega_1^2, ..., \omega_{p-1}^2 &\sim \prod_{j=1}^{p-1} \frac{\lambda_2^2}{2} e^{-\lambda_2^2 \omega_j^2/2} d\omega_j^2, \;\; \omega_1^2, \ldots, \omega_{p-1}^2 > 0
\end{aligned}
$$

Here $\tau_1^2, ..., \tau_p^2, \omega_1^2, ..., \omega_{p-1}^2$ are mutually independent.

Conditional Prior:

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

Fused LASSO II

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 $\frac{1}{\omega_0^2} = \frac{1}{\omega_p^2} = 0$.

Full Conditional Posterior

$$\beta \mid \sigma^2, \tau_1^2, ..., \tau_p^2, \omega_1^2, ...\omega_{p-1}^2, X, \tilde{y} \sim N_p\left((X^TX + \Sigma_\beta^{-1})^{-1}X^T\tilde{y}, \sigma^2(X^TX + \Sigma_\beta^{-1})^{-1}\right)$$

$$1/\tau_j^2 \mid \beta, \sigma^2, \omega_1^2, ...\omega_{p-1}^2, X, \tilde{y} \sim \textit{inverse Gaussian}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \lambda_1^2\right)$$

$$\text{, for } j = 1, 2, ..., p$$

$$1/\omega_j^2 \mid \beta, \sigma^2, \tau_1^2, ...\tau_{p-1}^2, X, \tilde{y} \sim \textit{inverse Gaussian}\left(\sqrt{\frac{\lambda_2^2\sigma^2}{(\beta_{j+1} - \beta_j)^2}}, \lambda_2^2\right)$$

$$\text{, for } j = 1, 2, ..., p-1$$

$$\sigma^2 \mid \beta, \tau_1^2, ...\tau_{p-1}^2, \omega_1^2, ...\omega_{p-1}^2, X, \tilde{y} \sim \textit{inverted gamma}\left(\frac{n-1+p}{2}, \right.$$

$$\left.\frac{1}{2}(\tilde{y} - X\beta)^T(\tilde{y} - X\beta) + \frac{1}{2}\beta^T\Sigma^{-1}\beta\right),$$

Elastic Net I

Hierarchical Model

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

The matrix D_τ is a diagonal matrix given by,

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

Elastic Net II

Conditional Prior:

$$\pi(\beta \mid \sigma^2) = exp\left(-\frac{\lambda_2}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\right) exp\left(-\sum_{i=1}^{p}\frac{\lambda_1|\beta_i|}{\sigma}\right)$$

Full Conditional Posteriors:

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

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

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

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

Tuning Parameter Selectiom

- 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.

Simulation Study I

we draw samples of size $n = 20$ and $n = 200$. The true β has been chosen as $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$. The error variance has been considered as $\sigma = 3$. There are 8 explanatory variables and for each pair (x_i, x_j)) the pairwise correlation is taken as $(1/2)^{|i-j|}$.

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: Performances of MSE for different methods for Example 1

Simulation study II

In this case the set up is exactly same as Example 1, except $\forall j, \beta_j = 0.85$.

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: Performances of MSE for different methods for Example 2

Simulation study III

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: Performances of MSE for different methods for Example 3

Simulation Study IV

- Bayesian LASSO is at least as good as LARS-LASSO.
- Bayesian lassos provide reasonable standard errors for the zero-estimated coefficients.

Real Data Analysis I

Prostate Cancer Data:

- This data has been collected from Stamey et. al. (1989).
- The data consists of seven independent variables.
- 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.

Real Data Analysis II

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

Table: MSE for different methods

Thank You!

Reference

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

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