On Bayesian Variable Selection Using Lasso

Anastasia Lykou[1] and Ioannis Ntzoufras[2]

[1]Department of Mathematics and Statistics, Lancaster University, UK
[2]Department of Statistics, Athens University of Economics and Business, Greece

September 25, 2010

Outline

Lasso (least absolute shrinkage and selection operator)

The Lasso (Tibshirani, 1996) performs variable selection and shrinkage on the linear regression problems by imposing the L_1 norm.

$$\widehat{\beta}_{\text{lasso}} = \text{argmin}_{\beta} \{ \text{var}(Y - \mathbf{X}\beta) + \lambda ||\beta||_1 \}$$

- The L_1 norm properties shrink the coefficients towards zero and exactly to zero if λ is large enough.
- The Lasso estimates can be considered as the posterior modes under independent double exponential priors.

Motivation

Due to the Lasso's advantages and its apparent Bayesian perspective, there are various methods for the Bayesian Lasso regression in the literature, such as

- Yuan and Lin (2005)
- Park and Casella (2008)
- Hans (2009a, 2009b)

However, some of the proposed methods

- perform shrinkage and lack of direct variable selection,
- fail to propose an effective method to specify the shrinkage parameter.

Bayesian Lasso

We use the following formulation

$$\begin{aligned} Y|\beta,\tau,\gamma &\sim \mathrm{N}_n(\mathbf{XD}_\gamma\beta,\tau^{-1}I_n), \quad \text{where} \quad \mathbf{D}_\gamma = \mathrm{diag}(\gamma_1,\ldots,\gamma_p), \\ \beta_j &\sim \mathrm{DE}\left(0,\frac{1}{\tau\lambda}\right), \quad \text{for} \quad j=1,\ldots,p, \\ \gamma_j &\sim \mathrm{Bernoulli}(\pi_j), \\ \tau &\sim \mathrm{Gamma}(a,d), \end{aligned} \qquad (1)$$

where λ is the shrinkage parameter which controls the prior variance given by $2/(\lambda\tau)^2$.

- We estimate $f(\beta|Y,\cdot)$ and $f(\gamma|Y,\cdot)$ with Kuo and Mallick (1998) gibbs sampler for variable selection.
- Any equivalent such as GVS (Dellaportas et al , 2002) or RJMCMC (Green, 1995) will provide similar results.
- Inference is based on the posterior medians of $\beta_j^* = \gamma_j\beta_j$ for $j=1,\ldots,p$.

A Gibbs Sampler for Bayesian Lasso

If $\gamma_j = 1$, then

$$f(\beta_j | Y, \sigma^2, \beta_{\setminus j}, \gamma_{\setminus j}, \gamma_j = 1)$$
$$= \quad w_j \, f_{TN}(\beta_j; m_j^-, s_j^2, \beta_j < 0) + (1 - w_j) \, f_{TN}(\beta_j; m_j^+, s_j^2, \beta_j \geq 0)$$

with

- $f_{TN}(x; \mu, \sigma^2, A)$ is the density distribution evaluated at x of the usual normal distribution truncated in the subset $A \subset \Re$
- $w_j = \dfrac{\Phi(-m_j^-/s_j)/f_N(0\,; m_j^-, s_j^2)}{\Phi(-m_j^-/s_j)/f_N(0\,; m_j^-, s_j^2) + \Phi(m_j^+/s_j)/f_N(0\,; m_j^+, s_j^2)}$.
- $m_j^- = \dfrac{c_j + \lambda}{||X_j||^2}, \; m_j^+ = \dfrac{c_j - \lambda}{||X_j||^2}, \; c_j = X_j^T(e + \beta_j X_j), \; s_j^2 = \dfrac{1}{\tau ||X_j||^2}$.
- X_j is the jth column of matrix $\mathbf{X}$ and $e = Y - \eta$ is the vector of residuals.

A Gibbs Sampler for Bayesian Lasso (cont.)

- If $\gamma_j = 1$,
  - Generate ω_j from Bernoulli(w_j)
  - Generate β_j from $\begin{cases} TN(m_j^-, s_j^2, \beta_j < 0), & \text{if } \omega_j = 1 \\ TN(m_j^+, s_j^2, \beta_j \geq 0), & \text{if } \omega_j = 0 \end{cases}$
- If $\gamma_j = 0$, generate β_j from its prior, that is

$$\beta_j | Y, \sigma^2, \beta_{\setminus j}, \gamma_{\setminus j}, \gamma_j = 0 \quad \sim \quad \mathrm{DE}\left(0, \frac{1}{\tau\lambda}\right)$$

- Generate σ^2 from $IG\left(\frac{n}{2} + p + \alpha, \frac{||Y - \mathbf{XD}_\gamma \beta||^2}{2} + \lambda||\beta|| + d\right)$.
- Generate γ_j from Bernoulli with probability $O_j/(1 + O_j)$ with

$$O_j = \frac{f(Y|\beta, \tau^2, \gamma_{\setminus j}, \gamma_j = 1)}{f(Y|\beta, \tau^2, \gamma_{\setminus j}, \gamma_j = 0)} \frac{\pi(\gamma_{\setminus j}, \gamma_j = 1)}{\pi(\gamma_{\setminus j}, \gamma_j = 0)}.$$

Example

A simulated dataset of Dellaportas et al. (2002), consists of $n = 50$ observations and $p = 15$ covariates generated from a standardised normal distribution and the response from

$$Y_i \sim \mathrm{N}(X_{i4} + X_{i5}, 2.5^2), \quad \text{for} \quad i = 1, \ldots, 50.$$

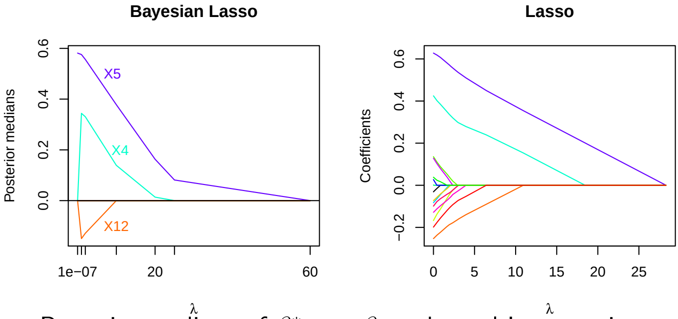

Figure: Posterior medians of $\beta_j^* = \gamma_j \beta_j$ and usual Lasso estimates against λ.

Bayes factors for simple Lasso Regression

The Bayes factors for the comparison between two simple models:

$$m_j : Y|\beta,\tau,m_j \sim \mathrm{N}_n(X_j\beta_j,\tau^{-1}I_n), \qquad m_0 : Y|\beta,\tau,m_0 \sim \mathrm{N}_n(0,\tau^{-1}I_n).$$

Assuming standardized data, $BF_{j\,0}$ can be expressed

$$BF_{j\,0} = \frac{\lambda}{n-1} c \left\{ \left(1+\frac{t_-^2}{df}\right)^{\frac{df}{2}} F_{t_{df}}\left(t_-\right) + \left(1+\frac{t_+^2}{df}\right)^{\frac{df}{2}} F_{t_{df}}\left(t_+\right) \right\}$$

where $F_{t_{df}}$ is the cdf of a Student's t random variable and

$$c = \sqrt{\pi}\frac{\Gamma\left(\frac{df}{2}\right)}{\Gamma\left(\frac{df-1}{2}\right)}, \; t_- = -\frac{M_{j-}\sqrt{df}}{\sqrt{1-M_{j-}^2}}, \; t_+ = \frac{M_{j+}\sqrt{df}}{\sqrt{1-M_{j+}^2}}, \quad (2)$$

$$M_{j-} = \rho_j + \frac{\lambda}{n-1}, \; M_{j+} = \rho_j - \frac{\lambda}{n-1}, \; df = n+2a+1,$$

where ρ_j is the sample Pearson correlation between Y and X_j and without loss of generality we assume that is positive.

Feasible values of λ

From Equation (2) quantities $1 - M_{j-}^2$ and $1 - M_{j+}^2$ must be positive and therefore a range of feasible values for λ is specified.

$$0 < \lambda < (n - 1)(1 - \rho_j).$$

Active values of λ

- We examine the sensitivity of the Bayes factors (BF) on different values of λ.
- **We focus on** the shrinkage values that provide sufficient evidence in favour of either of the two competing models.
- We use the Kass and Raftery (1995) interpretation tables to discard
  a) extremely low values of BF that fully support the null model (due to Bartlett-Lindley's paradox) and
  b) BF values which cannot separate between the two competing models (i.e. when BF provides weak evidence in favour of the supported model).
- We therefore define as active values of λ the following area

$$\Lambda_{act} = \{\lambda : \log(BF) > 1\} \cup \{\lambda : -5 < \log(BF) < -1\}.$$

Graphical representation

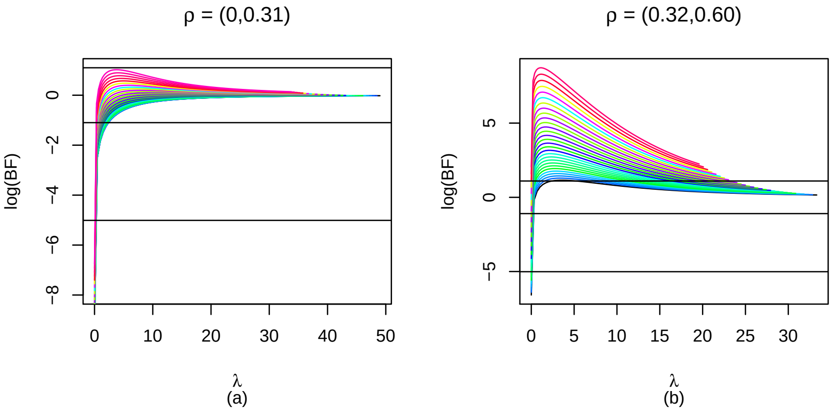

Figure: Bayes factor BF_{j0} of model m_j versus model m_0 against the values of λ, ρ; sample size is fixed to $n = 50$.

Graphical representation (cont.)

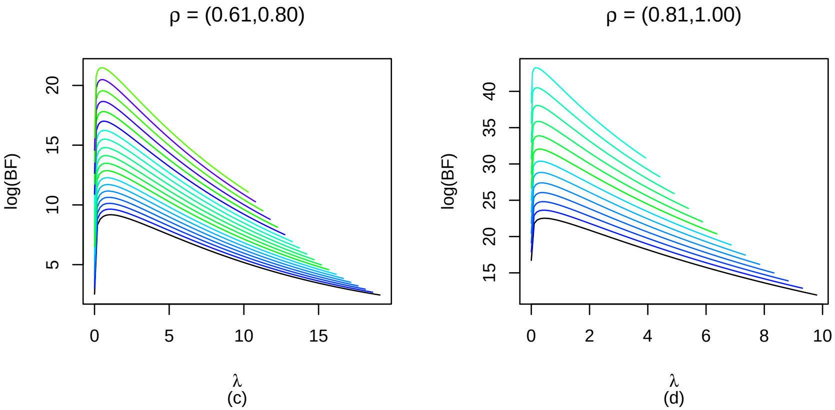

Figure: Bayes factor BF_{j0} of model m_j versus model m_0 against the values of λ, ρ; sample size is fixed to $n = 50$.

Specification of λ

Figure 1a shows that

- There is a range of values of ρ where the BF cannot separate between the two competing models:
  $$BF < 3 \quad \text{for all the values of} \quad \lambda.$$
- For $n = 50$, the BF never provides evidence in favour of the simple regression model for X_j which are associated with Y with $\rho \leq 0.31$.
- For any given sample size n, there is a range of sample correlations which can be considered as "non-important" for all values of λ.

  We use the upper limit of this range as the benchmark to tune the shrinkage parameter.

Specification of λ (cont.)

In order to define the prior shrinkage parameter λ, we set a threshold value ρ_t for the correlation ρ and we select the corresponding BF value for this threshold value.

For example, for $n = 50$, we select that any covariate X_j which is correlated with Y with $\rho_t = 0.4$ must have BF=1 $\Rightarrow \lambda = 0.067$.

This procedure provides with a value of λ (here 0.067) that gives 50% posterior probability to the simple regression model with covariate X_j and 50% to the constant model for a selected level of correlation ρ_t (here 0.4).

n	50	100	500
ρ_b	0.31	0.22	0.10
$BF = 1$	$\rho_t = 0.35,\ \lambda = 0.218$	$\rho_t = 0.25,\ \lambda = 0.335$	$\rho_t = 0.15,\ \lambda = 0.060$
$BF = 1$	$\rho_t = 0.40,\ \lambda = 0.067$	$\rho_t = 0.30,\ \lambda = 0.069$	$\rho_t = 0.20,\ \lambda = 7 \times 10^{-4}$
$BF = 1$	$\rho_t = 0.50,\ \lambda = 0.004$	$\rho_t = 0.40,\ \lambda = 0.001$	$\rho_t = 0.30,\ \lambda = 5 \times 10^{-6}$
$BF = \frac{1}{150}$	$\rho_t = 0.01,\ \lambda = 0.038$	$\rho_t = 0.01\ \lambda = 0.053$	$\rho_t = 0.01\ \lambda = 0.116$

Table: Shrinkage levels that correspond to BF=1 for various values of rho_t and n, ρ_b denotes the upper limit of the non-important correlations.

Bayes Factor for multiple Lasso regression

To understand and interpret the behavior of our procedure we also facilitate comparisons involving multiple regression setups including or not a specific covariate.

Hence, we compare the full model m_f with the model $m_{f\setminus j}$, where the j_{th} variable is excluded.

We use the Laplace approximation to derive the corresponding BF.

$$BF_m \approx kc\left[1 - pr_j^2\right]^{-df/2} \frac{1}{\sqrt{\left(1 - R^2_{X_j|\mathbf{X}_\gamma}\right)\left(1 - R^{(lasso)2}_{Y|\mathbf{X}_\gamma}\right)}}.$$

where

- $k = \lambda/(n-1)$
- $pr_j = \text{corr}^{(lasso)}(Y, X_j|\mathbf{X}\gamma)$ is the LASSO version of the partial correlation,
- $R^2_{X_j|\mathbf{X}_\gamma}$ is the multiple correlation coefficient when regressing X_j on $\mathbf{X}_\gamma$,
- $R^{(lasso)2}_{Y|\mathbf{X}_\gamma}$ is the LASSO version of the multiple correlation coefficient when regressing Y on $\mathbf{X}_\gamma$.

Relation between the partial correlation and Pearson correlation

To complete the interpretation of our selected λ, we identify the threshold value $pr_{j,t}$ of pr_j which provides the same BF as the corresponding one for simple LASSO regression model.

Using the above specification, we find that

$$(1 - pr_{j,t}^2)\left[(1 - R^2_{X_j|\mathbf{X}_\gamma})(1 - R^{(lasso)2}_{Y|\mathbf{X}_\gamma})\right]^{1/(n+2a+1)} = 1 - (\rho_t - k)^2.$$

From the above it is deduced that

- the threshold value for the LASSO partial correlation is upper bounded by a penalized expression of the corresponding threshold value of the Pearson correlation.

$$pr_{j,t}^2 \leq (\rho_t - k)^2,$$

- the two thresholds are approximately equal as $n \to \infty$ (i.e. for large sample sizes).

Simulation study 1

We perform the Bayesian Lasso on the simulated data from Dellaportas et. al. (2002) for specific values of λ that have been chosen through the univariate Bayes factor.

			Var. incl.	Post. Incl. Prob	Prob. of model	
ρ_t	BF	λ		X_4, X_5, X_{12}	MAP	true
0.35	1	0.217	X_4, X_5	$0.96, 1.00, 0.38$	26.22%	
0.40	1	0.067	X_4, X_5	$0.85, 1.00, 0.15$	55.49%	
0.50	1	0.004	X_5	$0.45, 0.96, 0.01$	50.45%	43.33%
0.01	$\frac{1}{150}$	0.038	X_4, X_5	$0.78, 1.00, 0.09$	61.39%	

Table: Posterior summaries for various choices of λ.

The absolute values of the lasso partial correlations of the variables X_4, X_5, X_{12} in this data set are: $(0.51, 0.68, 0.34)$.

Simulation study 2

- A simulated study from Nott and Kohn (2005).
- Consists of 15 covariates and 50 observations.
- The first 10 variables follow independent $N(0,1)$.
- The last 5 are generated using the following scheme

$$(X_{11},\dots,X_{15}) = (X_1,\dots,X_5) \times (0.3,0.5,0.7,0.9,1.1)^T \times (1,1,1,1,1) + \mathrm{E},$$

where E consists of 5 independent $N(0,1)$.
- The response is generated as

$$Y = 2X_1 - X_5 + 1.5X_7 + X_{11} + 0.5X_{13} + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, 2.5^2 I).$$

ρ_t	BF	λ	Var. incl.	Post. Incl. Prob X_1, X_5, X_7, X_{11}, X_{13}	Prob. of model MAP	true
0.35	1	0.217	X_1, X_7, X_{11}	1.00, (0.24), 1.00, 0.97, (0.19)	20.07%	1.47%
0.40	1	0.067	X_1, X_7, X_{11}	1.00, (0.09), 1.00, 0.96, (0.08)	57.38%	0.45%
0.50	1	0.004	X_1, X_7, X_{11}	1.00, (0.01), 0.99, 0.91, (0.04)	88.50%	0%
0.01	$\frac{1}{150}$	0.038	X_1, X_7, X_{11}	1.00, (0.05), 1.00, 0.95, (0.07)	70.65%	0.19%

The absolute values of the lasso partial correlations of the important variables are:(0.50, 0.27, 0.67, 0.49, 0.18)

Simulation study 2 (cont.)

We have simulated 100 samples and we perform the Bayesian Lasso regression for the chosen levels of shrinkage.

The average posterior inclusion probabilities for each variable are:

ρ_t	BF	λ	Post. Incl. Prob X_1, X_5, X_7, X_{11}, X_{13}
$\rho = 0.35$	1	0.218	0.99, 0.37, 0.91, 0.83, 0.40
$\rho = 0.40$	1	0.067	0.98, 0.24, 0.84, 0.78, 0.27
$\rho = 0.50$	1	0.004	0.83, 0.07, 0.51, 0.57, 0.18
$\rho = 0.01$	$\frac{1}{150}$	0.038	0.97, 0.18, 0.79, 0.75, 0.23

The following Table shows the frequency that three selected models are the MAP model in our Bayesian LASSO procedure.

ρ_t	BF	λ	X_1, X_7, X_{11}	X_1, X_5, X_7, X_{11}	$X_1, X_5, X_7, X_{11}, X_{13}$
0.35	1	0.218	27%	11%	6%
0.40	1	0.067	43%	9%	4%
0.50	1	0.004	30%	3%	0%
0.01	$\frac{1}{150}$	0.038	43%	9%	2%

Conclusions

- We propose an approach to select the prior (shrinkage) parameter λ based on its effect on Bayes factors.
- No specific data are utilized in this specification so the approach is purely Bayesian.
- We can interpret the behaviour of our Bayesian LASSO based on levels of practical significance of correlations and partial correlations which are widely understood.
- We have also specified an active area for λ, which truncates the range of λ in order to avoid the Bartlet-Lindley paradox and over-shrinkage.
- We managed to identify non-important covariates in reference to sample correlations that will be never supported by BF for all values of λ. This benchmark correlation can be calculated with simple iterative approaches while an lower bound of it is also available.

Further work

- Extend to hierarchical model, where a hyperprior is imposed on the shrinkage parameter.
- Allow different shrinkage parameters for each covariate (Adaptive Lasso).
- Perform the proposed ideas on the Bayesian Ridge regression
- Extend them for GLMs and categorical regressors.
- Examine the Bayesian implementation of other related methods such as Elastic net.

References

Dellaportas, P. and Forster, J. and Ntzoufras, I. (2002). *On Bayesian Model and Variable Selection Using MCMC*. Statistics and Computing. 12:27-36.

Green, P. (1995) *Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination*. Biometrika. 82:711-732.

Hans, C. (2009) *Bayesian Lasso Regression*. Biometrika. 96:835-845.

Hans, C. (2009) *Model uncertainty and variable selection in Bayesian lasso regression*. Accepted in Statistics and Computing.

Kass, R. and Raftery, A.(1995). *Bayes Factors*. Journal of the American Statistical Association. 90: 773-795.

Kuo, L. and Mallick, B.(1998). *Variable Selection for Regression Models*. The Indian Journal of Statistics. 60: 65-81.

Nott, D. and Kohn, R. (2005) *Adaptive sampling for Bayesian variable selection*. Biometrika. 92:747-763.

Park, T. and Casella, G. (2008). *The Bayesian Lasso*. Journal of the American Statistical Association. 103:681-687.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. J. Royal. Statist. Soc. B. 58:267-288.

Yuan, M. and Lin, Y. (2005) *Efficient Empirical Bayes Variable selection and Estimation in Linear Models*. Journal of American Statistical Association 100.