Combining ridge parameter with the g-prior of Zellner.

Denys Pommeret

Université Aix Marseille - Institut de Mathématiques de Luminy
Marseille - France

Thematic month in statistics - Bayesian Statistics and Algorithms - 29 February / 4 March, 2016

Outline

- A short review on g-prior
- Problem of ill-conditioned matrix $(X'X)^{-1} \hookrightarrow$ A ridge-g-prior approach
- An illustration

The g-prior in linear model

Consider the model

$$Y|X,\beta,\sigma^2 \sim \mathcal{N}(X\beta,\sigma^2 I)$$

where
Y is an n-vector of response
X a $n \times p$ design matrix (without constant)
$\beta \in \mathbb{R}^p$, the coefficient regression
$\sigma^2 > 0$, I the identity matrix.

Zellner's g-prior (1986) is given by

$$\beta|X, \sigma^2, g \sim \mathcal{N}(\beta_0 = 0, g\sigma^2(X'X)^{-1})$$

$$\sigma^2 \sim 1/\sigma^2$$

$g > 0$ is called the constant of Zellner.

Zellner's g-prior (1986) is given by

$$\beta|X, \sigma^2, g \sim \mathcal{N}(\beta_0 = 0, g\sigma^2(X'X)^{-1})$$

$$\sigma^2 \sim 1/\sigma^2$$

$g > 0$ is called the constant of Zellner.

Advantages :
$\hookrightarrow$ Simplicity : simple structure $\beta|Y, X, \sigma^2$ is Gaussian with variance $\dfrac{g\sigma^2}{g+1}(X'X)^{-1}$
$\hookrightarrow$ Automatic : using the structure of the variables (Fisher's Information Matrix)

But some cons : Consider the null model M_0 with $p_0 = 0$ and another model M_1 with $p_1 > 0$ covariates. Using g prior for M_1 we have closed forms for marginal likelihood and

- $BF[M_1 : M_0] = (1+g)^{(n-1-p_1)/2}[1+g(1-R_1^2)]^{-(n-1)/2}$
- If $g \to \infty$ (n and p_1 fixed) then $BF[M_1 : M_0] \to 0$ (Bartlett or Lindley's Paradox)
- If $R_1 \to 1$ (n and p_1 fixed) $BF[M_1 : M_0] \to constant$ (Information paradox)

Choice for the parameter g

- g can be fixed arbitrarly (Smith and Kohn, 1997)
- $g = n$ (Kass and Wasserman, 1995) $\hookrightarrow$ the BF is close to the BIC
- $g = p^2$ (Foster and George, 1994) $\hookrightarrow$ related to the Risk Inflation Criteria
- $g = \max(n, p^2)$ (Fernandez et al. 2001)
- Global or local empirical Bayes estimate $\hookrightarrow$ avoid the Information paradox

Prior on the hyperparameter g : mixture of g-priors

- Zellner and Siow (1980) prior

  $$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}$$

  which is an Inv-Gamma(1/2, n/2) prior

- Hyper g prior (Liang et al. 2008)

  $$\frac{1}{1+g} \sim Beta(a/2-1,1)$$

  with $2 < a < 4$ ($a = 2$ is the Jeffrey's prior, it is uniform for $a = 4$)
  Rmk : $\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}$
  $\hookrightarrow$ closed form of posterior distribution of g in terms of Gaussian hypergeometric function.

- Hyper g/n prior

- Truncated gamma prior (Wang et George, 2007)

$$\text{Density of } u = \frac{1}{1+g}: \qquad \pi(u) = \frac{s^a}{\gamma(a,s)} u^{a-1} e^{-su} \mathbb{I}_{(0,1)}(u)$$

- Beta prime prior (Maruyama and George, 2011)

$$\frac{1}{1+g} \sim Beta(1/2, (n-p-1.5)/2)$$

- Robust prior (Bayarri et al. 2012) which can be reduced to a particular truncated gamma prior

$$\text{Density of } u = \frac{1}{1+g}: \qquad \pi(u) = u^{-1/2} \mathbb{I}_{(0,(p+1)/(n+1))}(u)$$

- CH-g prior (Li and Clyde, 2016)

g-prior in mixture of linear models

- Mixture of linear models (Gupta and Ibrahim, 2006)

$$\beta(m)|X,\sigma^2 \sim \mathcal{N}(\beta_0(m), g\sigma^2(m)(X'X)^{-1})$$

- Mixture of linear models (Lee et al, 2016)

$$\beta(m)|X,\sigma^2 \sim \mathcal{N}(\beta_0(m), g(m)\sigma^2(m)(X'(m)X(m))^{-1})$$

g-prior and variable selection

Stochastic Search Variable Selection

γ vector indicating which variables are active
$\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ otherwise.

- $p_\gamma = \sum_{i=1}^{n} \gamma_i$
- X_γ $n \times p_\gamma$ design matrix with active variables
- β_γ p_γ vector with non-null elements.

g-prior and variable selection

Stochastic Search Variable Selection

γ vector indicating which variables are active
$\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ otherwise.

- $p_\gamma = \sum_{i=1}^{n} \gamma_i$
- X_γ $n \times p_\gamma$ design matrix with active variables
- β_γ p_γ vector with non-null elements.

Choice of γ_i : Bernoulli
$P(\gamma_i = 1) = \pi_i$

g-prior and variable selection

Stochastic Search Variable Selection

γ vector indicating which variables are active
$\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ otherwise.

- $p_\gamma = \sum_{i=1}^{n} \gamma_i$
- X_γ $n \times p_\gamma$ design matrix with active variables
- β_γ p_γ vector with non-null elements.

Choice of γ_i : Bernoulli
$P(\gamma_i = 1) = \pi_i$

Another choice : $\gamma|\omega \sim Bernoulli(\omega)$ and $\omega \sim Beta(a, b)$

g-prior in GLM and variable selection

$$h(\mathbb{E}(Y_i|U,\beta)) = X_i'\beta + Z_i'U,$$

where

- h is the link function
- $U = (U_1, \cdots, U_k)$ are the random effect, with U_i of size q_i
- $\beta_\gamma|\gamma \sim \mathcal{N}(0, \Sigma_\gamma)$
  with $\Sigma_\gamma = \tau(X_\gamma' X_\gamma)^{-1}$.
- τ is the variable selection coefficient (Bottolo and Richardson, 2010)

Different applications of variable selection with g-prior in GLM or GL2M :

- Probit model : Lee et al. (2003), Sha et al. (2004), Zhou et al. (2004)
- GLM : Chen and Ibrahim (2003), Marin and Robert (2007), Wang and George (2007), Gupta and Ibrahim (2009), Bové and Held (2011), Li and Clyde (2016)
- Probit mixed model : Yang and Song (2010), Baragatti and P. (2010), Baragatti (2011), Baragatti and P. (2012)
- Logistic model : Hanson et al. (2014)

New g-prior in GLM

Li and Clyde (2016) proposed the truncated Compound Confluent Geometric Hyperbolic prior (tCCGH)

- Let $u = 1/(1+g)$
- $u \sim tCHHH(a/2, b/2, r, s/2, v, \theta)$,
  which generalizes all previous cited mixtures of g priors.

Another method proposed by Li and Clyde (2016) in GLM

Based on the following idea in linear model (Zellner, 1980, Maruyama and George, 2011) :

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

- Less certain about β_1
- More certain about β_2

Write $V = (I - P_{X_1})X_2$, where $P_{X_1} = X_1(X_1'X_1)^{-1}X_1'$ is the projection matrix on X_1. Then

$$Y = X_1\xi + V\beta_2 + \epsilon$$

with $\xi = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$.

These decomposition allow the use of two independent g-priors

- $\xi \sim \mathcal{N}(\xi_0, g_1\sigma^2(X_1'X_1)^{-1})$
- $\beta_2 \sim \mathcal{N}(\beta_0, g_2\sigma^2(V'V)^{-1})$

If little information is available on $X_1 \hookrightarrow$ large value for g_1

A numerical problem with the inversion of $(X'X)$

$$\beta | X \sim \mathcal{N}(0, \tau (X'X)^{-1})$$

$(X'X)^{-1}$ appears in g-prior (for linear models, mixture of linear models, GLM, GL2M). It can be ill-conditioned

- If $p > n$
- If there are linear dependence between regressors

A numerical problem with the inversion of $(X'X)$

$$\beta|X \sim \mathcal{N}(0, \tau(X'X)^{-1})$$

$(X'X)^{-1}$ appears in g-prior (for linear models, mixture of linear models, GLM, GL2M). It can be ill-conditioned

- If $p > n$
- If there are linear dependence between regressors

$(X'_\gamma X_\gamma)^{-1}$ also appears in g-prior with SSVS. It can also be ill-conditioned

- If $p_\gamma > n$
- If there are linear dependence between regressors of X_γ.

Some solutions

- Using the decomposition of Li and Clyde (2016) :

$$\xi \sim \mathcal{N}(\xi_0, g_1\sigma^2(X_1'X_1)^{-1})$$
$$\beta_2 \sim \mathcal{N}(\beta_0, g_2\sigma^2(V'V)^{-1})$$

- Using bounds for p_γ (as done in Baragatti, 2011)

Using generalized inverse

- In Probit mixed model Yang and Son (2010) replaced $(X'_\gamma X_\gamma)^{-1}$ by $(X'_\gamma X_\gamma)^+$ its Moore Penrose 's inverse.
  $\hookrightarrow$ drawback in the MCMC algorithm
- Wang et al. (2014) used also the generalized singular g-prior in linear model with

$$\beta \sim \mathcal{N}(\beta_0, g\sigma^2 (X'_\gamma X_\gamma)^+)$$

Changing the prior

- Bayesian Lasso (Park and Casella, 2008, Hans, 2009)

$$\begin{aligned}
\beta|\Lambda, \sigma^2, \gamma &\sim \mathcal{N}(0, \sigma^2\Lambda) \\
\Lambda &= diag(\lambda_1, \cdots, \lambda_p) \\
\lambda_1, \cdots, \lambda_p|\delta &\sim \prod_{i=1}^{p} \frac{\delta}{2} \exp\{-\delta\lambda_i/2\} \\
\delta &\sim \gamma(a, b) \\
\sigma^2 &\sim 1/\sigma^2
\end{aligned}$$

- Bayesian Lasso and SSVS : Lykou and Ntzoufras (2013)
  $\hookrightarrow$ introducing the vector γ.
- Bayesian ElasticNet (Li and Lin, 2010)

Ridge-g-prior

Gupta and Ibrahim (2009), Baragatti and P. (2012), Lee et al (2016), Li and CLyde (2016)

$$\beta_\gamma|\lambda,\gamma \sim \mathcal{N}(0,\Sigma_\gamma(\lambda))$$

where

$$\Sigma_\gamma(\lambda) = (\tau^{-1}X'_\gamma X_\gamma + \lambda I)^{-1}$$

τ is the variable selection coefficient, λ is the ridge parameter

Choice of τ and λ

Following the idea of the g-prior, where the variance-covariance structure is preserved, we replicate the total variance of the data as follows :

- Write $\Sigma_\gamma(0) = \tau_0 (X_\gamma' X_\gamma)^{-1}$ the classical g-prior (without ridge parameter).
- The constraint used is : $tr(\Sigma_\gamma(0)^{-1}) = tr(\Sigma_\gamma(\lambda)^{-1})$
- We choose $\lambda = 1/p$
- We get $\tau = \tau_0 \left[1 + \dfrac{\tau_0}{tr(X'X) - \tau_0} \right]$

Illustration with a probit mixed model

The model is

- $P(Y_i = 1 \mid U, \beta) = \Phi(X_i^T \beta + Z_i^T U)$,
  where Φ stands for the standard Gaussian cumulative distribution function.
  Following Albert and Chib (1993) and Lee et al. (2003), a vector of latent variables $L = (L_1, \ldots, L_n)^T$ is introduced, and we assume that that is $L \mid U, \beta \sim \mathcal{N}_n(X\beta + ZU, I_n)$.

$$Y_i = \begin{cases} 1 & \text{if } L_i > 0 \\ 0 & \text{if } L_i < 0. \end{cases}$$

- The γ_j are assumed to be independent Bernoulli(π_j)
- $U|D \sim \mathcal{N}(0, D)$
  with (for simplicity) $D = diag(A_1, \ldots, A_K)$, where $A_l = \sigma_l^2 I$, $l = 1, \ldots, K$
- σ_l^2 are Inverse Gamma $\mathcal{IG}amma(a, b)$

- The full conditional distribution of L is given by :

$$L_i|\beta, U, Y_i = 1 \sim \mathcal{N}(X_i^T\beta + Z_i^T U, 1) \text{ left truncated at } 0$$
$$L_i|\beta, U, Y_i = 0 \sim \mathcal{N}(X_i^T\beta + Z_i^T U, 1) \text{ right truncated at } 0.$$

- Defining $W = (Z^T Z + D^{-1})^{-1}$, the full conditional distribution of U is :

$$U|L, \beta, D \sim \mathcal{N}_q(WZ^T(L - \mathbf{X}\beta), W).$$

- The full conditional distribution of the $\sigma_l^2, l = 1, \dots, K$ are Inverse-Gamma :

$$\sigma_l^2 \mid U_l \sim \mathcal{IG}amma\Big(\frac{q_l}{2} + a, (\frac{1}{2}U_l^T U_l + b)\Big).$$

Only the full conditional distributions of β_γ and γ depend on λ, as follows :

- For β_γ :

  $$\beta_\gamma | L, U, \gamma \sim \mathcal{N}(V_\gamma \mathbf{X}_\gamma^T (L - ZU), V_\gamma),$$

  with $V_\gamma = \left[\frac{(1+\tau)}{\tau}\mathbf{X}_\gamma^T \mathbf{X}_\gamma + \lambda I\right]^{-1}$.
- And for γ :

  $f(\gamma | L, U, \beta_\gamma) \propto \frac{(2\pi)^{-\frac{d_\gamma}{2}}}{|\Sigma_\gamma(\lambda)|^{1/2}} \prod_{j=1}^{p} \pi_j^{\gamma_j}(1 - \pi_j)^{1-\gamma_j}$
  $\times \exp\left[-\frac{1}{2}\left(\beta_\gamma^T V_\gamma^{-1} \beta_\gamma - (L - ZU)^T \mathbf{X}_\gamma \beta_\gamma - \beta_\gamma^T \mathbf{X}_\gamma^T (L - ZU)\right)\right]$

  with $\Sigma_\gamma(\lambda) = (\tau^{-1} X_\gamma' X_\gamma + \lambda I)^{-1}$

MCMC

Simulations from all the full conditional distributions can be easily obtained, except for γ which does not correspond to a standard multivariate one.
We use a Metropolis-within-Gibbs algorithm.
Following Lee et al. (2003) combined with the grouping technique of Liu (1994), we consider γ and β_γ jointly. We have

$$f(\gamma|L,U) \propto \frac{|V_\gamma|^{1/2}}{|\Sigma_\gamma(\lambda)|^{1/2}} \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}$$

$$\times \exp\Big[-\tfrac{1}{2}(L-ZU)^T(I-\mathbf{X}_\gamma V_\gamma \mathbf{X}_\gamma^T)(L-ZU)\Big]$$

Metropolis-Hasting step :

$$\rho(\gamma^{(i)}, \gamma^*) = \min\left\{1, \frac{f(\gamma^*|L,U)}{f(\gamma^{(i)}|L,U)}\right\},$$

with $\dfrac{f(\gamma^*|L,U)}{f(\gamma^{(i)}|L,U)}$

$$= \left(\frac{|V_{\gamma^*}\Sigma_{\gamma^{(i)}}(\lambda)|}{|\Sigma_{\gamma^*}(\lambda)V_{\gamma^{(i)}}|}\right)^{1/2}$$

$$\times \exp\left\{-\tfrac{1}{2}(L-ZU)^T(\mathbf{X}_{\gamma^i}V_{\gamma^{(i)}}\mathbf{X}^T_{\gamma^{(i)}} - \mathbf{X}_{\gamma^*}V_{\gamma^*}\mathbf{X}^T_{\gamma^*})(L-ZU)\right\}$$

$$\times \prod_{j=1}^{p}\left(\frac{\pi_j}{1-\pi_j}\right)^{\gamma_j^* - \gamma_j^{(i)}},$$

where γ^* corresponds to $\gamma^{(i)}$ in which r components have been randomly changed (see Chipman et al. 2001, George and McCulloch, 1997).

Post-processing :
The number of iterations of the algorithm is $b+m$, where b corresponds to the burn-in period and m to the observations from the posterior distributions. For selection of variables, the sequence $\{\gamma^{(t)} = (\gamma_1^{(t)}, \ldots, \gamma_p^{(t)}), t = b+1, \ldots, b+m\}$ is used. The most relevant variables for the regression model are those corresponding to the γ components with higher posterior probabilities, and can be identified as the γ components that are most often equal to 1.

The Bayesian Lasso :
For each $\beta_j, j = 1, \ldots, p$ we consider

- $\beta_j \mid \lambda_j \sim \mathcal{N}(0, \lambda_j)$
- $\lambda_j \sim \mathcal{E}xpo(\delta/2)$.

Writing $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$, we have
$\beta \mid \Lambda \sim \mathcal{N}_p(0, \Lambda)$.
δ : $\delta \sim \mathcal{G}amma(e, f)$,

$\beta | L, U, \Lambda \sim \mathcal{N}_p(V_\Lambda \mathbf{X}^T (L - ZU), V_\Lambda)$
with
$V_\Lambda = \Big[\mathbf{X}^T \mathbf{X} + \Lambda^{-1}]^{-1}$
$\lambda_j \mid \beta \sim \mathcal{IG}auss\Big(\frac{\sqrt{\delta}}{\beta_j}, \delta\Big).$
The posterior for the Lasso parameter δ is a gamma distribution :
$\delta \mid \Lambda \sim \mathcal{G}amma\Big(p + e, \big(\frac{\sum \lambda_j}{2} + \frac{1}{f}\big)^{-1}\Big)$

Post-processing for Lasso approach : From the results of the Bayesian Lasso we obtain posterior estimates for the β_js and the λ_js, and the variables can be selected by different ways :

- One can select the variables corresponding to an absolute value $|\beta_j|$ higher than a threshold (Li et al. 2011).
- Bae and Mallick (2004) proposed to select variables corresponding to high values of λ_j.
- Finally, the results of the Lasso enable us to obtain posterior credible intervals (CI) for the β_js (Kyung et al. 2010).

Numerical study

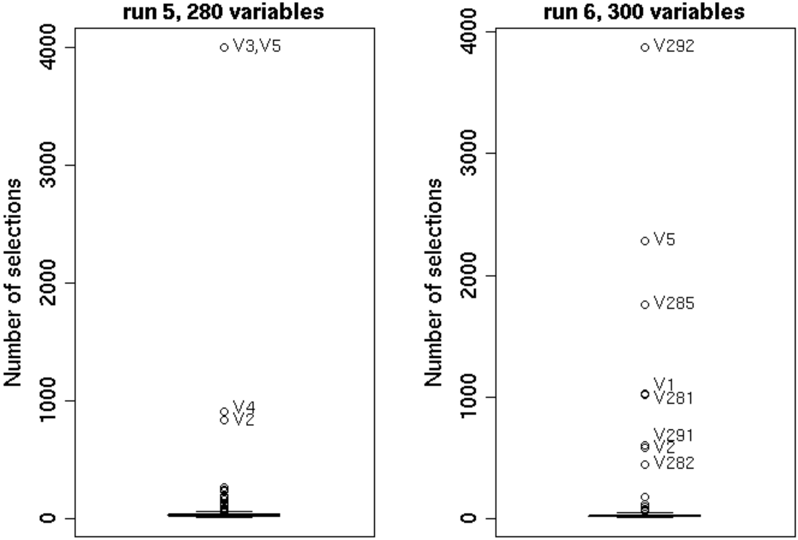
We start with $n = 200$ observations : 100 for training set, 100 for test set. With $p = 300$ variables
The response are obtained using a probit mixed model with only 5 of these variables : $\mathbf{V_1}, \cdots, \mathbf{V_5}$ and one random effect of length 4.
$V_1, \cdots, V280$ are iid $Uniform(0,1)$.
$V281 = 2 \times V1, \cdots, V290 = 2 \times V10$
$V291 = V1 + V2, V292 = V3 - V4, V293 = V5 + V13$
$V_{294} \cdots, V_{300}$ linear combinations of $V6, \cdots, V20$.

$\beta = (1, -1, 2, -2, 3)$

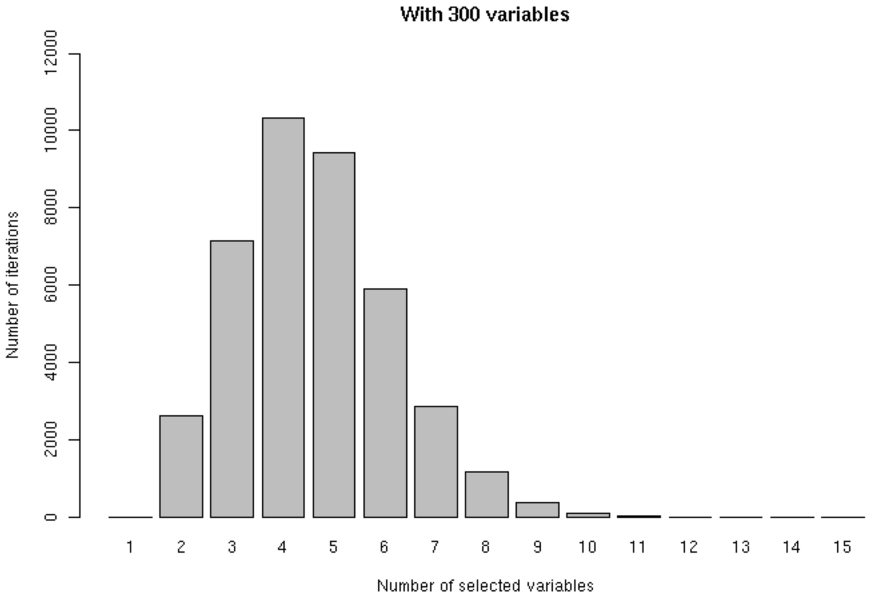
Summary of important variables : $\mathbf{V1}, \cdots, \mathbf{V5}$
$\mathbf{V281 = 2 \times V1}, \cdots, \mathbf{V285 = 2 \times V5}$
$\mathbf{V291 = V1 + V2, V292 = V3 - V4, V293 = V5 + V13}$

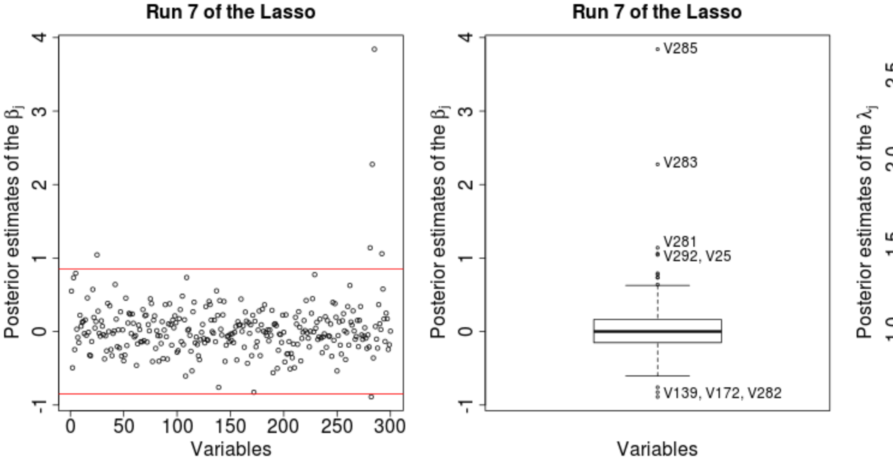
We used 10 runs, starting with $\tau_0 = 50$, $\pi_j = \pi = 5/300$, $b = 2000$ burn-in iterations, $m = 4000$ observations after burn-in, $\lambda = 1/300$.

Boxplot of a run with ridge g-prior



Variables	Number of selections among the 10 runs with 280 variables	Number of selections among the 10 runs with 300 variables
$V1$	0	10
$V2$	9	8
$V3$	10	2
$V4$	5	0
$V5$	10	10
$V281 = 2 \times V1$	Not available	10
$V282 = 2 \times V2$		9
$V283 = 2 \times V3$		3
$V284 = 2 \times V4$		0
$V285 = 2 \times V5$		10
$V291 = V1 + V2$		7
$V292 = V3 - V4$		10


With 300 variables
Number of iterations
0
2000
4000
6000
8000
10000
12000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of selected variables


Run 7 of the Lasso
Posterior estimates of the β_j
4
3
2
1
0
-1
0
50
100
150
200
250
300
Variables
Run 7 of the Lasso
Posterior estimates of the β_j
4
3
2
1
0
-1
V285
V283
V281
V292, V25
V139, V172, V282
Variables
Posterior estimates of the λ_j
2.5
2.0
1.5
1.0

To compare Bayesian Lasso and ridge g prior we keep the five most selected variable at each run.

Variables	Using Bayes Lasso	Using Ridge g prior
selected in 9 runs	$V285$	$V292$
selected in 8 runs		$V5, V285$
selected in 7 runs	$V283, V292$	$V281$
selected in 6 runs		$V282$
selected in 4 runs		$V283, V291$
selected in 3 runs	$V282$	$V2$
selected in 2 runs	$V281$	
selected in 1 runs	None of interest	

Comparison with Bayesian Lasso with SSVS

- we chose a linear model with 300 covariates and with a sample size $n = 50, 100, 200$.
- The $n \times 300$ design matrix $\mathbf{X}$ is first formed by a centered normal vector of size 100, with very high correlations (uniformly distributed between 0.6 and 1). The 200 next covariates are independent and uniformly distributed into $(-5, 5)$.
- The response $\mathbf{y}$ is constructed from the relation :

$$\mathbf{y} = \mathbf{X}\beta^\top + \epsilon,$$

- with $(\beta_1, \cdots, \beta_8) = (1, -1, 2, -2, 3, -3, 5, -5)$ and $\beta_j = 0$, $\forall j > 8$, and ϵ a vector of i.i.d. centered normal variables with variance 4
- In general no more than eight variables was clearly most retained by the algorithms and we then restricted our attention to the 3, 5 and 8 most often selected variables.

	n	SSVS Bayesian Lasso	SSVS ridge g prior	SSVS g prior
RSS (3)	50	331	**328**	343
RSS (5)	50	232	**224**	231
RSS (8)	50	178	141	**134**
RSS (3)	100	1145	**1066**	1076
RSS (5)	100	**491**	505	508
RSS (8)	100	**340**	345	347
RSS (3)	200	2740	**2607**	2696
RSS (5)	200	1445	**1318**	1400
RSS (8)	200	**784**	795	793

Conclusion

- Ridge g prior is easy to implement
- It seems that it stabilizes the variable selections (in presence of colinearity)
- It works for $p > n$
- Automatic choice
- Not too much sensible wrt the choice of (τ, λ).
- It could be compared to Bayesian ridge regression.

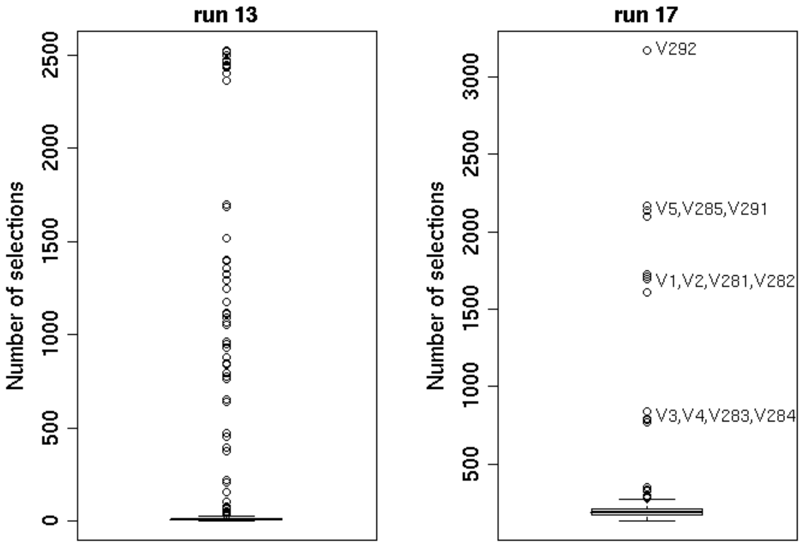


Figure: Boxplot of the number of selections of a variable after the burn-in period, for two runs with 300 variables.

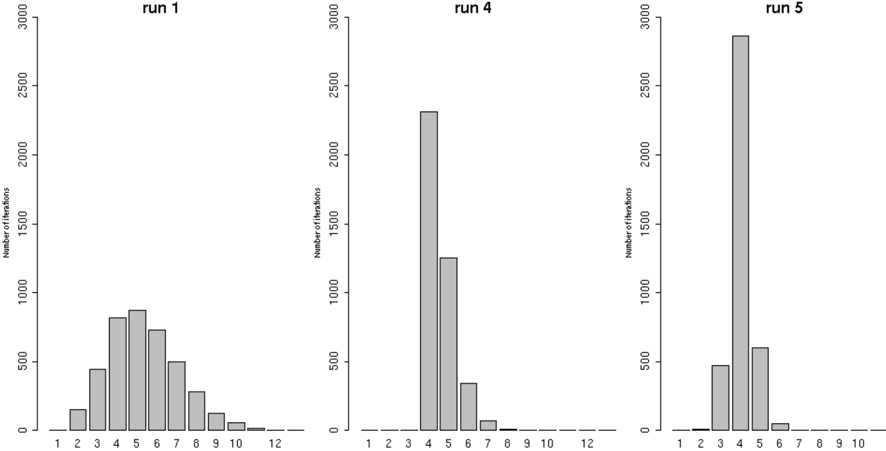


Figure: Number of iterations of the runs 1,4 and 5 associated with a number of selected variables from 1 to 14. For each run, there were 4000 post burn-in iterations.

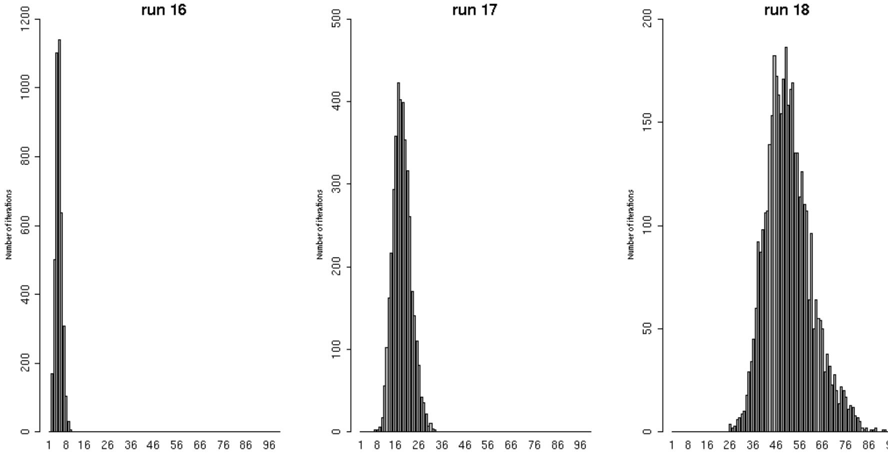


Figure: Number of iterations of the runs 16,17 and 18 associated with a number of selected variables from 1 to 100. For each run, there were 4000 post burn-in iterations.