Spike and Slab Variable Selection: Frequentist and Bayesian Strategies

Hemant Ishwaran and J. Sunil Rao

May 5, 2011

1. The Spike and Slab Model
2. Variable Selection and Regression
3. Experiments
4. Conclusion

The Spike and Slab Model

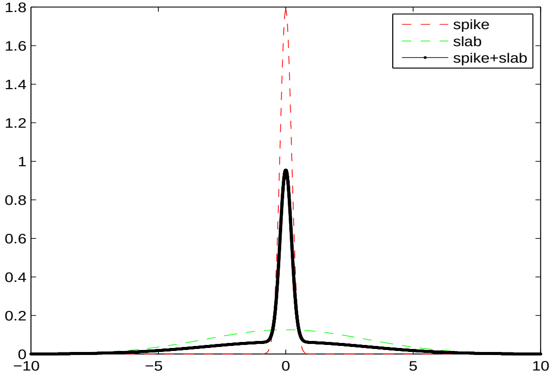
The regression problem: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

The Spike and Slab model:

$$
\begin{aligned}
(Y_i^* | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) &\sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2 n) \\
(\beta_k | \phi_k, \tau_k^2) &\sim \mathcal{N}(0, \phi_k \tau_k^2) \\
(\phi_k | v_0, w) &\sim (1-w)\delta_{v_0}(\cdot) + w\delta_1(\cdot) \\
(\tau_k^{-2} | b_1, b_2) &\sim \mathrm{Gamma}(a_1, a_2) \\
w &\sim \mathrm{Uniform}[0,1] \\
\sigma^{-2} &\sim \mathrm{Gamma}(b_1, b_2)
\end{aligned}
$$

- ⋆ $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times K}$ is the data matrix. $\mathbf{Y} = [\mathbf{Y}_1, \cdots, \mathbf{Y}_n]^\top$ is the original response. $\mathbf{Y}_i^* = \hat{\sigma}_n^{-1} n^{\frac{1}{2}} \mathbf{Y}_i$ is the normalized response with $\hat{\sigma}_n^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^o\|^2/(n-K)$ and $\hat{\boldsymbol{\beta}}_n^o = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y})$ is the OLS estimate.
- ⋆ Settings: $a_1 = 5$, $a_2 = 50$, $b_1 = b_2 = 0.0001$, $v_0 = 0.005$.
- ⋆ Notice that σ^2 is rescaled by n.

The Spike and Slab Model



$$(\beta_k|\tau_k^2, w) \sim (1-w)\mathcal{N}(0, v_0\tau_k^2) + w\mathcal{N}(0, \tau_k^2)$$

Variable Selection and Regression

In this paper, variable selection and regression are done separately in two steps.

- ⋆ Step 1: The posterior mean of the spike-slab model $\hat{\beta}_n^*$ is used to identify the variables in the model, via Zcut or svsForward.
- ⋆ Step 2: The final regression coefficient is the OLS estimate *using only the identified variables* in Step 1, i.e., $\hat{\beta}_n^o[k] = (\mathbf{X}[k]^\top \mathbf{X}[k])^{-1}(\mathbf{X}[k]^\top \mathbf{Y})$ where $\mathbf{X}[k]$ denotes a $n \times k$ matrix containing the k selected variables. $\hat{\beta}_n^o[k]$ is called a restricted OLS estimate.

Two Variable Selection Methods

- ⋆ Zcut: Hard shrinkage on posterior mean.
  $\text{Zcut} := \{\beta_k : |\hat{\beta}^*_{k,n}| \geq z_{\alpha/2}\}$
  where $z_{\alpha/2} = \text{norminv}(1 - \frac{\alpha}{2})$, $\alpha = 0.10$.

- ⋆ svsForward: Forward selection.
  First reorder the variables using $|\hat{\beta}^*_{k,n}|$.
  FOR $k = 1, 2, \cdots, K$
  Find the restricted OLS estimate $\hat{\beta}^o_n[k]$.
  Compute the Z-statistics $\tilde{Z}_{k,n} = \frac{n^{1/2}\hat{\beta}^o_{k,n}}{\hat{\sigma}_n s_{kk}^{1/2}}$.
  if $|\tilde{Z}_{k,n}| < z_{\alpha_k/2}$, return top $k-1$ variables; Stop; end
  END

$\tilde{Z}_{k,n}$ is a normalized version of $\hat{\beta}^o_{k,n}$, with $s_{kk} = ((\mathbf{X}[k]^\top \mathbf{X}[k])^{-1})_{kk}$.

Two Baseline Variable Selection Methods

Two alternative methods based on $\hat{\beta}_n^o$ instead of $\hat{\beta}_n^*$:

- ⋆ OLS-hard: Hard shrinkage on OLS estimate.
$$\text{OLS-hard} = \{\beta_k : |\tilde{Z}_{k,n}| \geq z_{\alpha/2}\}$$
where $\tilde{Z}_{k,n}$ is computed using all variables.

- ⋆ OLSForward: Reorder the variables based on $\tilde{Z}_{k,n}$ using all variables. Then do the same sequential forward selection as in svsForward.

Zcut Vs. OLS-hard

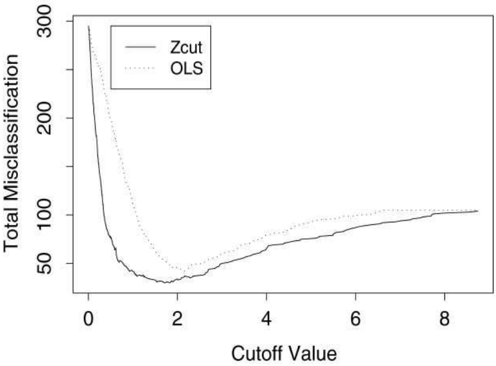


FIG. 6. *Total number of misclassified coefficients from simulation used in Figure 1. Observe how Zcut's total misclassification is less than OLS-hard's over a range of cutoff values* $z_{\alpha/2}$.

Diabetes Dataset

TABLE 1

Top 10 variables from diabetes data (ranking based on absolute posterior means $|\widehat{\beta}^*_{k,n}|$). Entries for model selection procedures are Z-statistics (12) derived from the restricted OLS for the selected model

	Variable	$\widehat{\beta}^*_{k,n}$	Zcut	OLS-hard	svsForwd	OLSForwd
1	bmi	9.54	8.29	13.70	8.15	13.70
2	ltg	9.25	7.68	0.00	7.82	0.00
3	map	5.64	5.39	7.06	4.99	7.06
4	hdl	−4.37	−4.20	0.00	−4.31	0.00
5	sex	−3.38	−4.03	−1.95	−4.02	−1.95
6	age.sex	2.43	3.58	3.19	3.47	3.19
7	bmi.map	1.61	0.00	2.56	3.28	2.56
8	glu.2	0.84	0.00	0.00	0.00	0.00
9	bmi.2	0.46	0.00	0.00	0.00	0.00
10	tc.tch	−0.44	0.00	0.00	0.00	0.00

OLS based methods missed two important variables (ltg and hdl).

Breiman simulations

TABLE 2
Breiman simulations

	$\rho = 0$ (uncorrelated X)					$\rho = 0.9$ (correlated X)				
	$\hat{k}$	Perf	TotalMiss	FDR	FNR	$\hat{k}$	Perf	TotalMiss	FDR	FNR
(A) *Moderate number of covariates with few (55%) that are zero* ($n = 200$, $K = 100$ *and 55 zero* $\beta_{k,0}$).										
Zcut	41.44	0.815	11.99	0.097	0.129	10.06	0.853	38.49	0.167	0.408
svsForwd	34.02	0.753	15.09	0.054	0.191	8.31	0.826	39.39	0.156	0.415
OLS-hard	41.99	0.791	14.06	0.128	0.145	11.08	0.707	45.31	0.496	0.446
OLSForwd	26.90	0.612	20.92	0.042	0.258	5.96	0.574	44.64	0.459	0.445
(B) *Large number of covariates with many (74%) that are zero* ($n = 800$, $K = 400$ *and 295 zero* $\beta_{k,0}$).										
Zcut	75.96	0.903	39.62	0.068	0.106	36.67	0.953	72.61	0.055	0.194
svsForwd	86.81	0.904	41.19	0.130	0.095	24.42	0.926	81.90	0.025	0.216
OLS-hard	106.74	0.883	58.54	0.279	0.097	45.41	0.706	121.37	0.676	0.255
OLSForwd	61.09	0.846	49.87	0.046	0.138	9.14	0.303	106.48	0.590	0.259

Conclusion

- ⋆ A rescaled Spike and Slab model is proposed.
- ⋆ The posterior mean of the model is used to select variables in the model via Zcut or svsForward.
- ⋆ Experiments show advantage compared with OLS based variable selection.
- ⋆ Detailed theoretical analysis is provided in the paper.