High-dimensional Feature Selection Using Hierarchical Bayesian Logistic Regression with Heavy-tailed Priors

Longhai Li

longhai@math.usask.ca

Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan

CRM-ISM-GERAD Colloque de Statistique
McGill University, 13 April 2012

Outline of this Talk

- High-dimensional Feature Selection in Classification Problems
- Review of Existing Methods
- Why Choose Heavy-tailed Priors and MCMC?
- Sketch of Our Implementation Details
- Simulation Studies and Application to a Real Microarray Data Set
- Conclusions, Discussions, and Acknowledgements

High-dimensional Feature Selection Problem

- Background: modern high-throughput biotechnologies can easily measure expression levels of thousands of genes, or SNPs of the whole genome.

- Problem: identifying a few features (such as genes) whose values are statistically relevant to a categorical response variable, such as an indicator of presence of a certain cancer.

- Mathematical Notations:
  - Response variable: y, taking integers $1, \ldots, C$.
  - Features (explanatory variables, or covariates): $\boldsymbol{x}_{1:p} = (x_1, \ldots, x_p)$.
  - Data: $(y_i,\ \boldsymbol{x}_{i,1:p})$ for $i = 1, \ldots, n$

- Challenges: high dimensionality of features (p), small sample size (n), often called $p >> n$ problems, and complex relationships between features.

- An analogue: looking for a couple of “needles” from a huge “haystack”.

Existing Methods in the Literature

- Univariate screenings, for example t-test,
- Diagonal discriminant rules, such as DLDA by Dudoit et.al.(2002), PAM by Tibshirani et.al.(2003), and BCBCSF by Li (2011).

  Limitation: ignore relationships among features
- Regularized discriminant rules, such as Guo, Hastie and Tibshirani (2007)

  Limitation: sensitive to non-Gaussian outliers
- Classification models based methods
  - Penalized likelihood methods, see review by Ma and Huang (2008)

    Limitation: results are unstable when non-convex penalties are used
  - Bayesian methods based on Spike-and-Slab priors, see Yang and Song (2010)

    Attractiveness: shrinking some coefficients to exactly 0

    Limitations: results are sensitive to choice of width of continuous distribution, see Lindley (1957), Lamnisos et.al. (2011), difficulties in MCMC sampling

Our Approach: Bayesian Logistic Regression with Moderately Heavy-tailed Priors and MCMC

- Logistic regression model for binary response y

$$P(y_i = k+1 | \boldsymbol{x}_{i,1:p}, \boldsymbol{\beta}_{0:p}) = \frac{I(k=0) + I(k>0)\exp\left(\beta_0 + \boldsymbol{x}_{i,1:p}\boldsymbol{\beta}_{1:p}\right)}{1+\exp\left(\beta_0 + \boldsymbol{x}_{i,1:p}\boldsymbol{\beta}_{1:p}\right)},$$

  for $k = 0$ and 1, $i = 1, \ldots, n$, where $\boldsymbol{\beta}_{1:p}$ is a column vector of regression coefficients, and $I(\cdot)$ is indicator function.

- Moderately heavy-tailed prior for β_j with small scale

$$\beta_1, \ldots, \beta_p \sim t(\text{df} = 1,\ \text{scale}{=}0.01), \text{or others}$$

- Computing: MCMC sampling using Hamiltonian Monte Carlo in Gibbs Sampling

Why Moderately Heavy-tailed Priors and MCMC?

Scale-Mixture-Normals Priors

- t distribution

$$\beta_j \sim N(0, \sigma_j^2), \quad \sigma_j^2 \sim \mathsf{IG}(\alpha/2, \alpha\gamma^2/2)$$

- Horseshoe or Inverted-Beta (Carvalho et al., 2010)

$$\beta_j \sim N(0, \sigma_j^2), \quad \sigma_j|\phi_j \sim N^+(0, \phi_j^2), \quad \phi_j^2 \sim \mathsf{IG}(\alpha/2, \alpha\gamma^2/2)$$

- Normal-Exp-Gamma (NEG, Griffin and Brown, 2012)

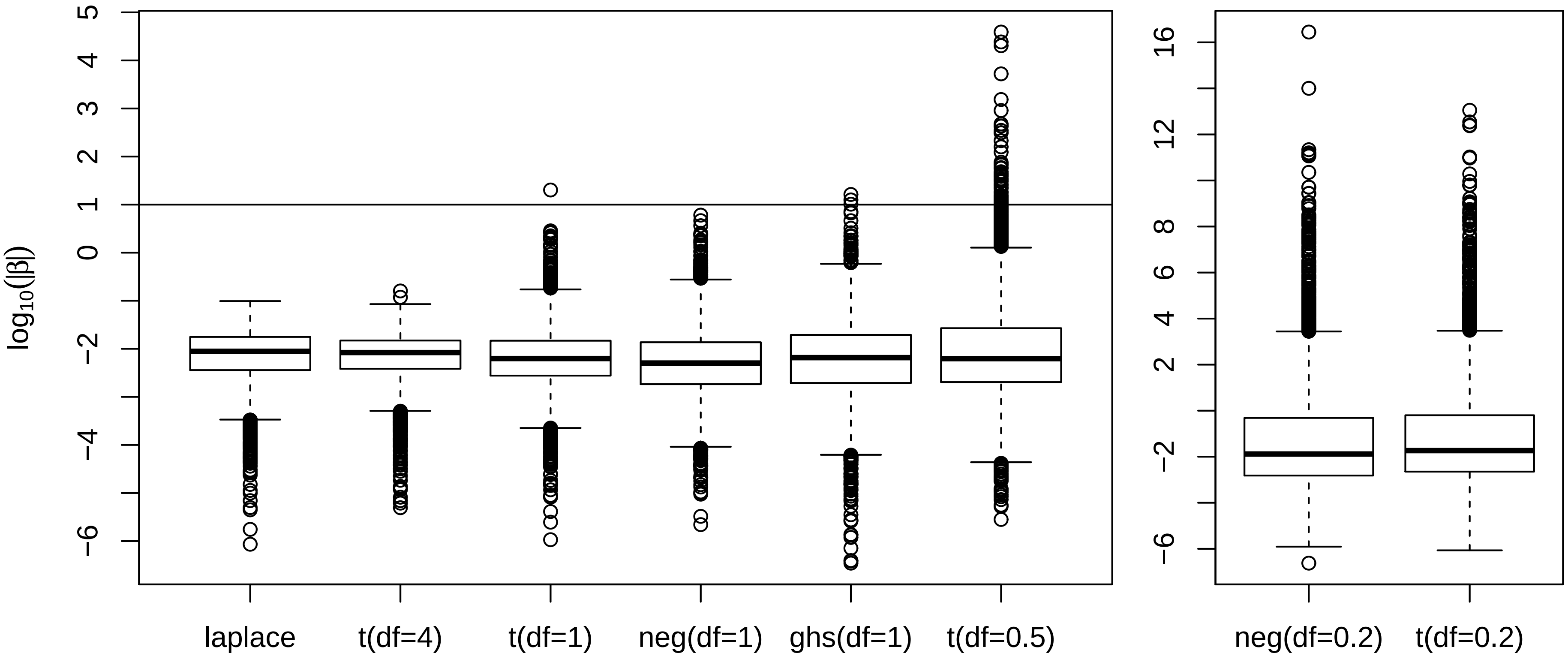
$$\beta_j \sim N(0, \sigma_j^2), \quad \sigma_j^2|\psi_j \sim \exp\left(\frac{1}{\psi_j}\right), \quad \psi_j \sim \mathsf{IG}(\alpha/2, \alpha\gamma^2/2)$$

- Laplace (used in LASSO)

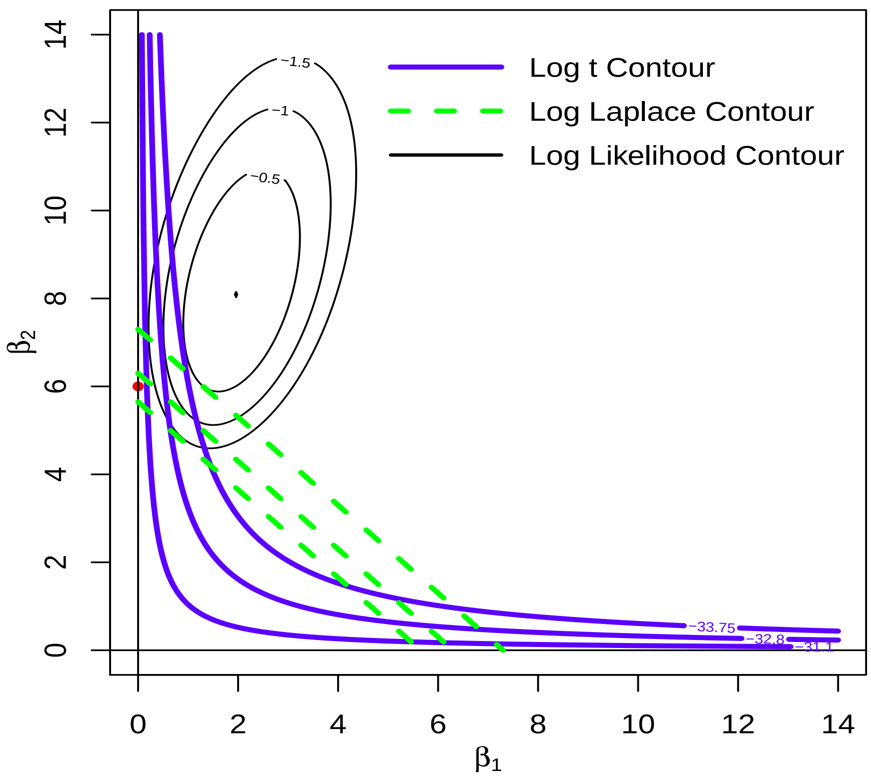
$$\beta_j \sim N(0, \sigma_j^2), \quad \sigma_j^2 \sim \exp(1/\gamma^2)$$

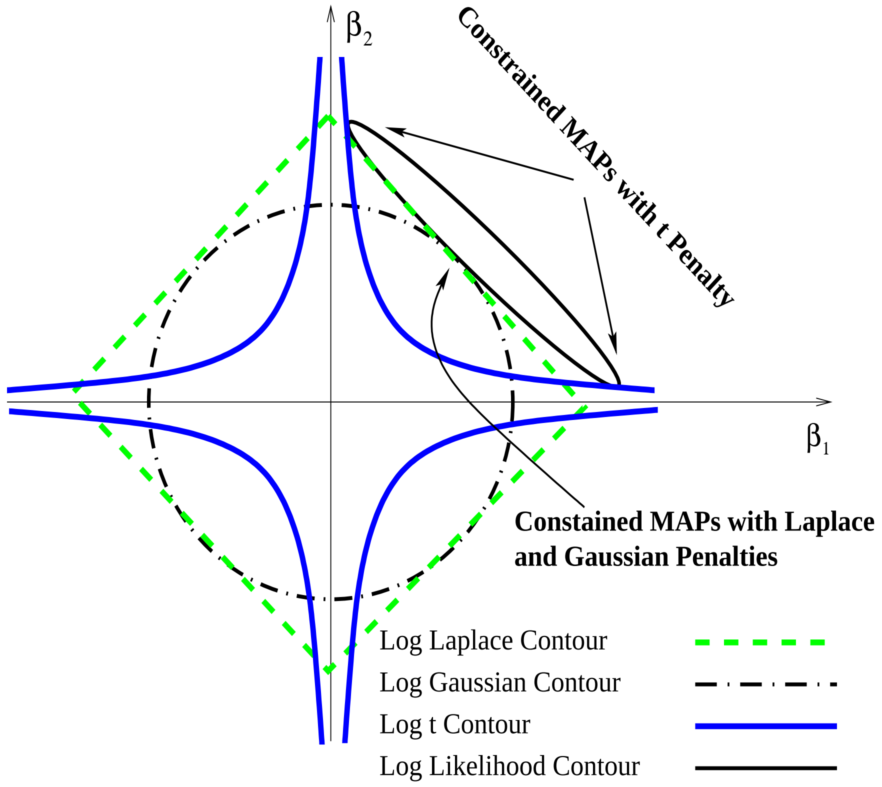
Bayesian Perspective of Choice of Prior

Boxplots of 4000 random numbers of $\log_{10}(|\beta_j|)$ from various priors



Shrinking Effect of Moderately Heavy-tailed Priors

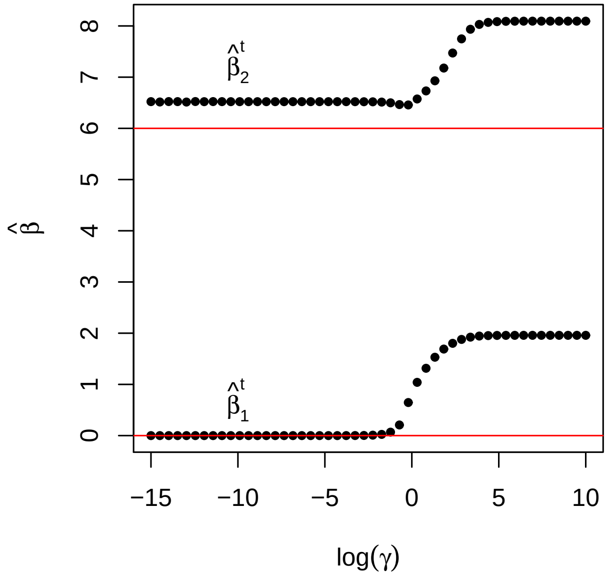


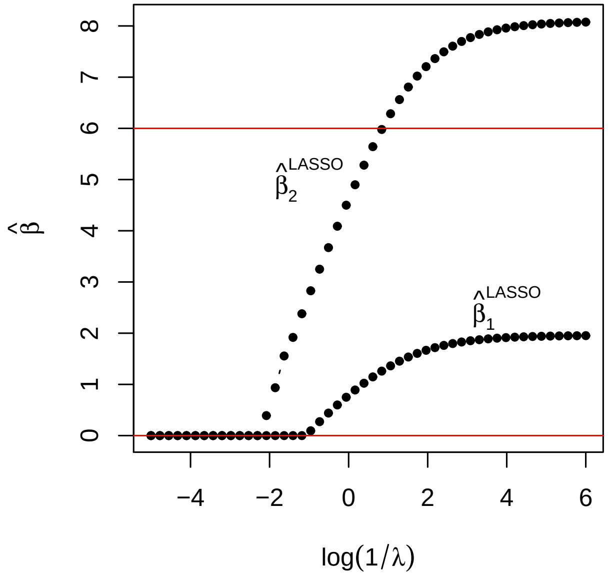


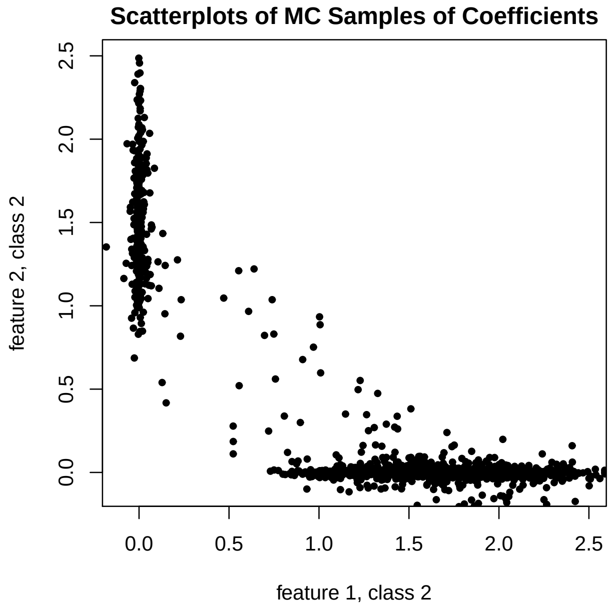
Constrained MAPs with t Penalty

Constained MAPs with Laplace and Gaussian Penalties

Examples Illustrating Shrinking Effect







Bayesian Multiclass Logistic Regression with Heavy-tailed Priors and MCMC

Bayesian Multiclass Logistic Regression Models (Original Symmetric Parameterization)

$$
\begin{aligned}
P(y_i = c|\boldsymbol{x}_{i,1:p}, \boldsymbol{\beta}_{0:p,1:C}) &= \frac{\exp\left(\beta_{0,c} + \boldsymbol{x}_{i,1:p}\boldsymbol{\beta}_{1:p,c}\right)}{\sum_{c=1}^{C} \exp\left(\beta_{0,c} + \boldsymbol{x}_{i,1:p}\boldsymbol{\beta}_{1:p,c}\right)}, \text{for } c = 1, \ldots, C, \\
\boldsymbol{\beta}_{j,1:C}|\sigma_j^2 &\sim N\ (0, \sigma_j^2),\ \text{ for } j = 0, \ldots, p, \\
\boldsymbol{\sigma}_{1:p}^2 &\sim \text{IG}\ (\alpha/2, w\alpha/2)
\end{aligned}
$$

For using Horseshoe or NEG priors for $\beta_{j,1:C}$, we need only to change the priors for σ_j^2 to other expressions, which exist in closed-form.

Bayesian Multiclass Logistic Regression Models (identifiable and symmetric parameterization)

$$
\begin{aligned}
P(y_i = k+1 | \boldsymbol{x}_{i,1:p}, \boldsymbol{\delta}_{0:p,1:K}) &= \frac{I(k=0) + I(k>0)\exp\left(\delta_{0k} + \boldsymbol{x}_{i,1:p}\boldsymbol{\delta}_{1:p,k}\right)}{1+\sum_{k=1}^{K}\exp\left(\delta_{0k} + \boldsymbol{x}_{i,1:p}\boldsymbol{\delta}_{1:p,k}\right)}, \\
\boldsymbol{\delta}_{j,1:K} | \sigma_j^2 &\sim N_K(\mathbf{0}, (I_K + J_K)\sigma_j^2), \\
\boldsymbol{\sigma}_{1:p}^2 &\sim \mathsf{IG}\ (\alpha/2, w\alpha/2),
\end{aligned}
$$

where $K = C - 1$, $\delta_{j,k} = \beta_{j,k+1} - \beta_{j,1}$.

Gibbs Sampling Procedure

The full posterior distribution of $\boldsymbol{\delta}_{0:p,1:K}$ and $\boldsymbol{\sigma}^2_{1:p}$ is:

$$P(\boldsymbol{\delta}_{0:p,1:K}, \boldsymbol{\sigma}^2_{1:p}|\boldsymbol{D}) \propto L(\boldsymbol{\delta}_{0:p,1:K}) \times P(\boldsymbol{\delta}_{0:p,1:K}|\boldsymbol{\sigma}^2_{0:p}) \times P\left(\boldsymbol{\sigma}^2_{1:p}\,|\,\alpha/2, \alpha\, w/2\right)$$

We sample the full posterior by iterating these two steps:

Step 1: Given $\boldsymbol{\sigma}^2_{1:p}$ fixed, use Hamiltonian Monte Carlo (HMC) for jointly sampling

$$P(\boldsymbol{\delta}_{0:p,1:K}|\boldsymbol{\sigma}^2_{0:p}, \boldsymbol{D}) \propto L(\boldsymbol{\delta}_{0:p,1:K}) \times P(\boldsymbol{\delta}_{0:p,1:K}|\boldsymbol{\sigma}^2_{0:p}).$$

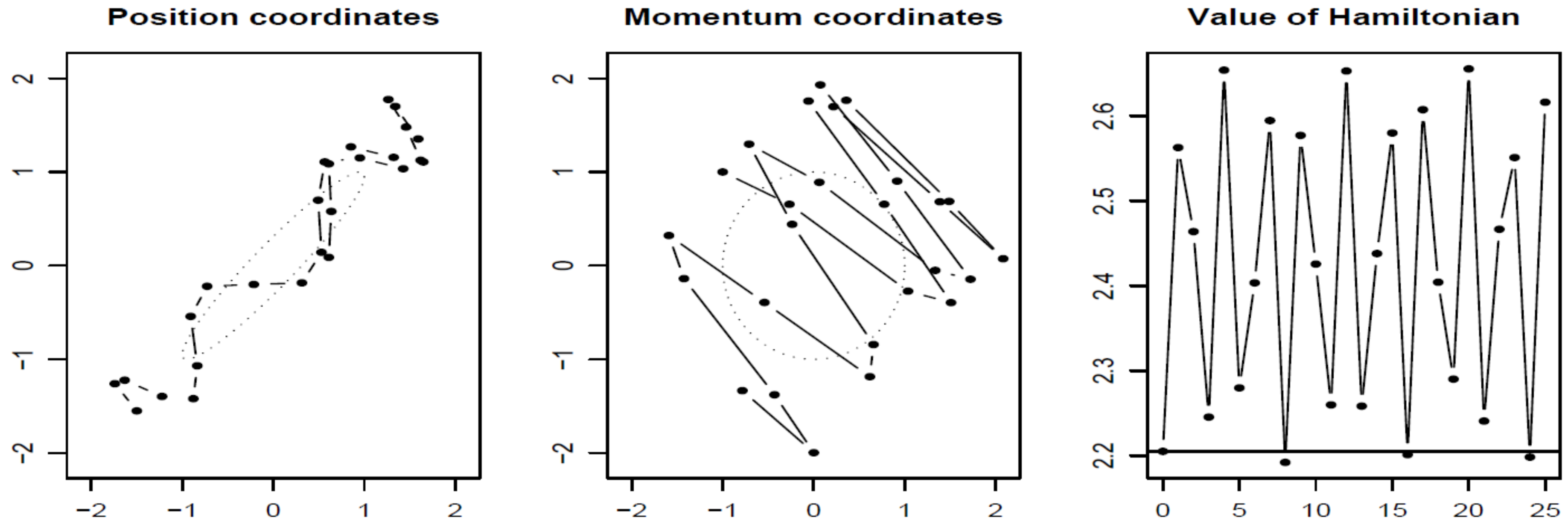
Step 2: Given value of $\boldsymbol{\delta}_{1:p,1:K}$ from Step 1, update $\boldsymbol{\sigma}^2_{1:p}$ by sampling from

$$\sigma^2_j|\boldsymbol{\delta}_{1:p,1:K} \sim \mathrm{IG}\left(\sigma^2_j \,\middle|\, \frac{\alpha + K}{2}, \frac{\alpha w + V(\boldsymbol{\delta}_{j,1:K})}{2}\right).$$

When Horseshoe and NEG priors are used for σ^2_j, we cannot sample directly for Step 2, but we can employ Adaptive Rejection Sampling. However, this step become time consuming as p is very large.

Why Use HMC for Sampling Regression Coefficients?

For highly-correlated posterior, HMC can move to a distant point with high acceptance rate, avoiding random walk of ordinary MH sampling.



The combination of HMC and the updating of σ_j^2 in Step 2 makes the whole sampler travel across multiple modes, see an explanation in slide 10.

Restricted Gibbs Sampling

When p is very large, sampling in Step 1 is very time consuming. A belief in high-dimensional classification is that most features are irrelevant and therefore most coefficients concentrate very close to 0. It is therefore useless to update them very often. In Step 1, we update only those features with σ_j greater than a small threshold ζ, say 0.05. Note that, however, no matter whether a regression coefficient of feature j is updated or not in Step 1, σ_j^2 is always updated in Step 2.

Restricted Gibbs sampling is justifiable with Markov chain theory. The sampling is still an exact MCMC sampling.

Using restricted Gibbs sampling, the time for computing linear combination $\sum_{j=1}^{p} x_{i,j}\delta_{j,c}$ in each single iteration of Step 1 is reduced greatly since $\sum_{j:\sigma_j>\zeta} x_{i,j}\delta_{j,c}$ can be reused from last iteration.

The effect of using restricted Gibbs sampling is that the coefficients of useful coefficients are updated much more frequently than those of useless.

ζ cannot be over large.

Feature Importance Indice (SDBs)

With posterior samples of $\boldsymbol{\delta}_{1:p,1:K}$, we recommend using *means* over iterations to estimate the coefficients, denoted by $\hat{\delta}_{j,1:K}$. We then compute the importance index of feature j with

$$\text{SDB}_j = \text{SD}((0, \hat{\delta}_{j,1}, \ldots, \hat{\delta}_{j,K})),$$

where $K = C - 1$.

For $C = 2$, it is just $\text{SDB}_j = |\hat{\delta}_j|/2$.

Comparisons on a Simulated Data Set with $p = 200$

Data Generation

We generated a date set of $n = 1100$ cases (of which 100 were used for fitting models, and the other 1000 were used to look at predictive performance), and $p = 200$ features from the following multivariate Gaussian model:

$$P(y_i = c) = \frac{1}{2}, \text{ for } c = 1, 2, \quad \boldsymbol{x}'_{i,1:200} \mid y_i = c \sim N_{100}\left(\boldsymbol{\mu}'_{c,1:200}\,,\, A\,A' + I_{200}\right), \quad (1)$$

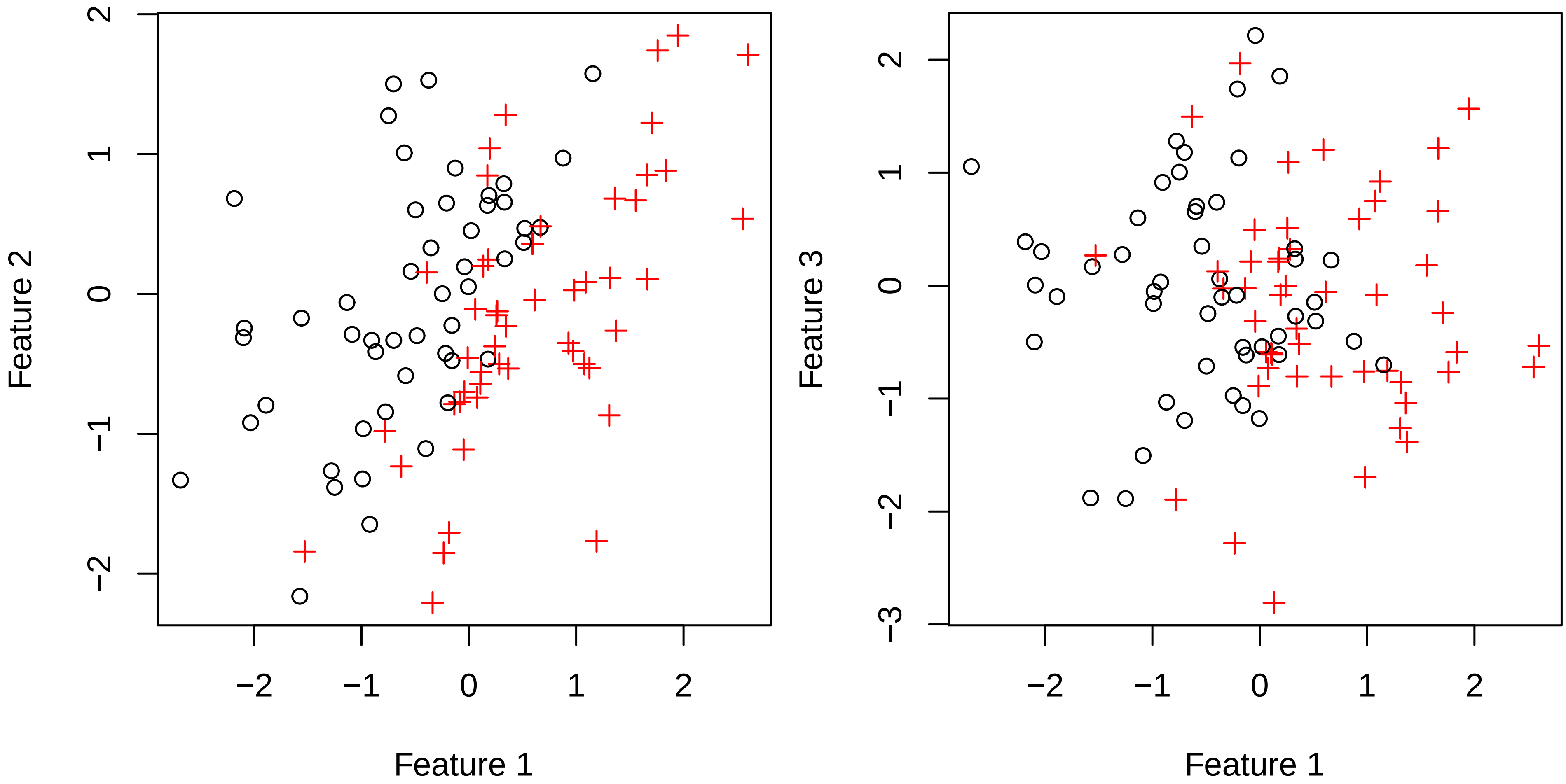
where, $\boldsymbol{\mu}_{1,1:200} = (0, \ldots, 0)$, $\boldsymbol{\mu}_{2,1:200} = (2, 0, \ldots, 0)$, $A = (a_{ij})$ with all diagonal elements equal to 1, and $a_{21} = 2$.

In the above model, the 1st feature has difference means in two classes, all others have the same means in two classes, but the 2nd feature is positively correlated with the 1st, and is useful too.

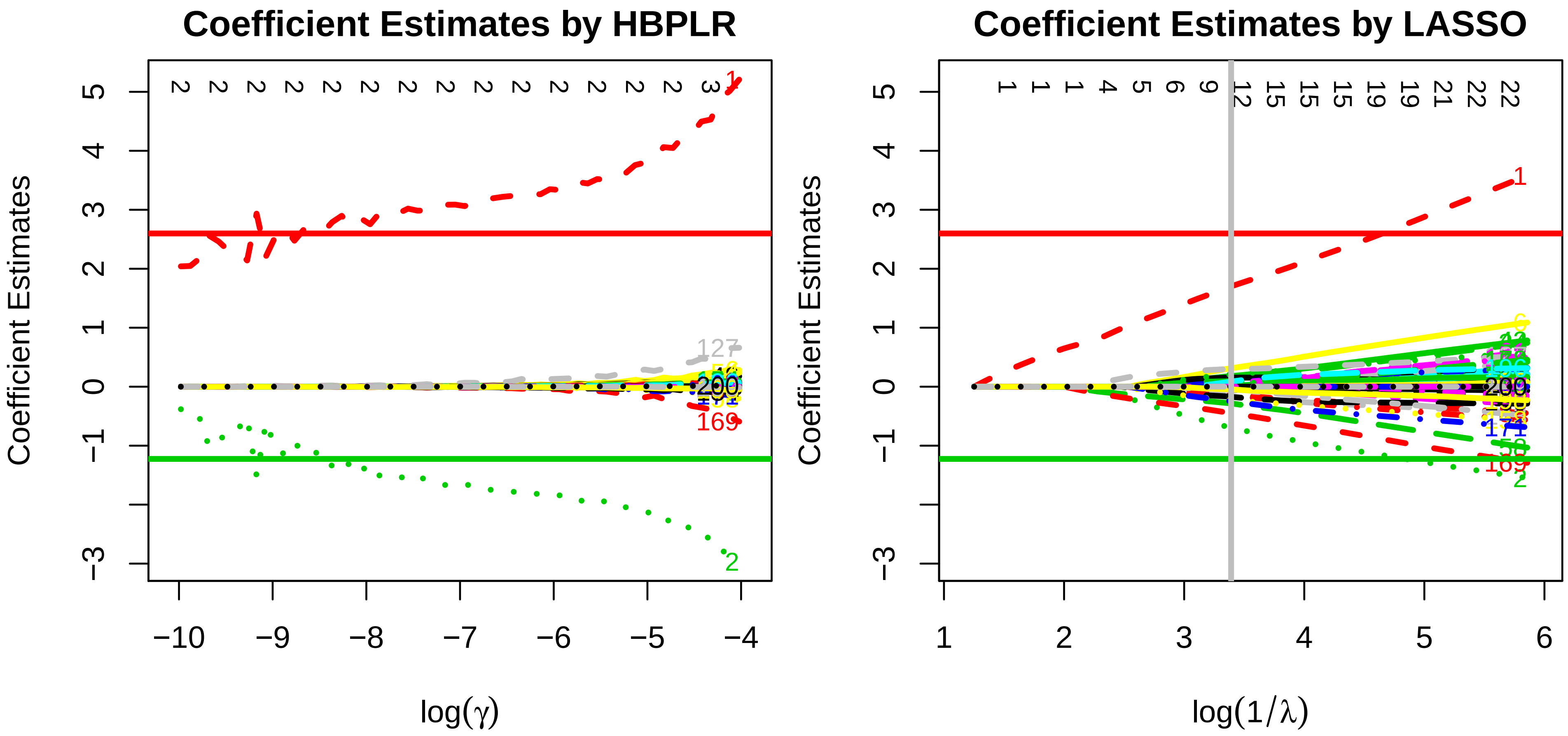
The true coefficients computed from discriminant rule is

$$\boldsymbol{\delta}_{0:200,1} = (0, 2.60, -1.22, 0, \ldots, 0).$$

Scatterplots of $x_1 - x_3$



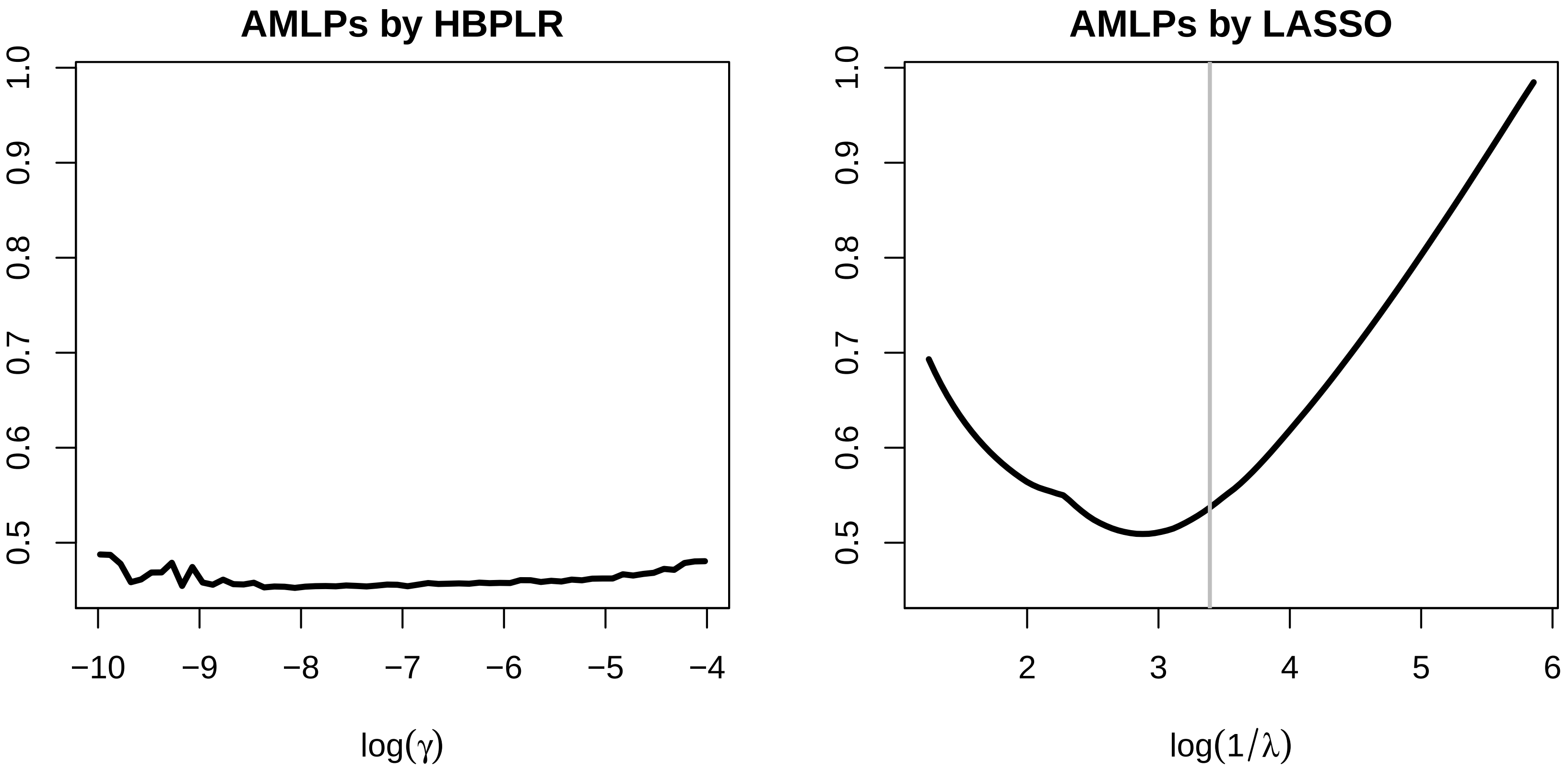
Solution Paths



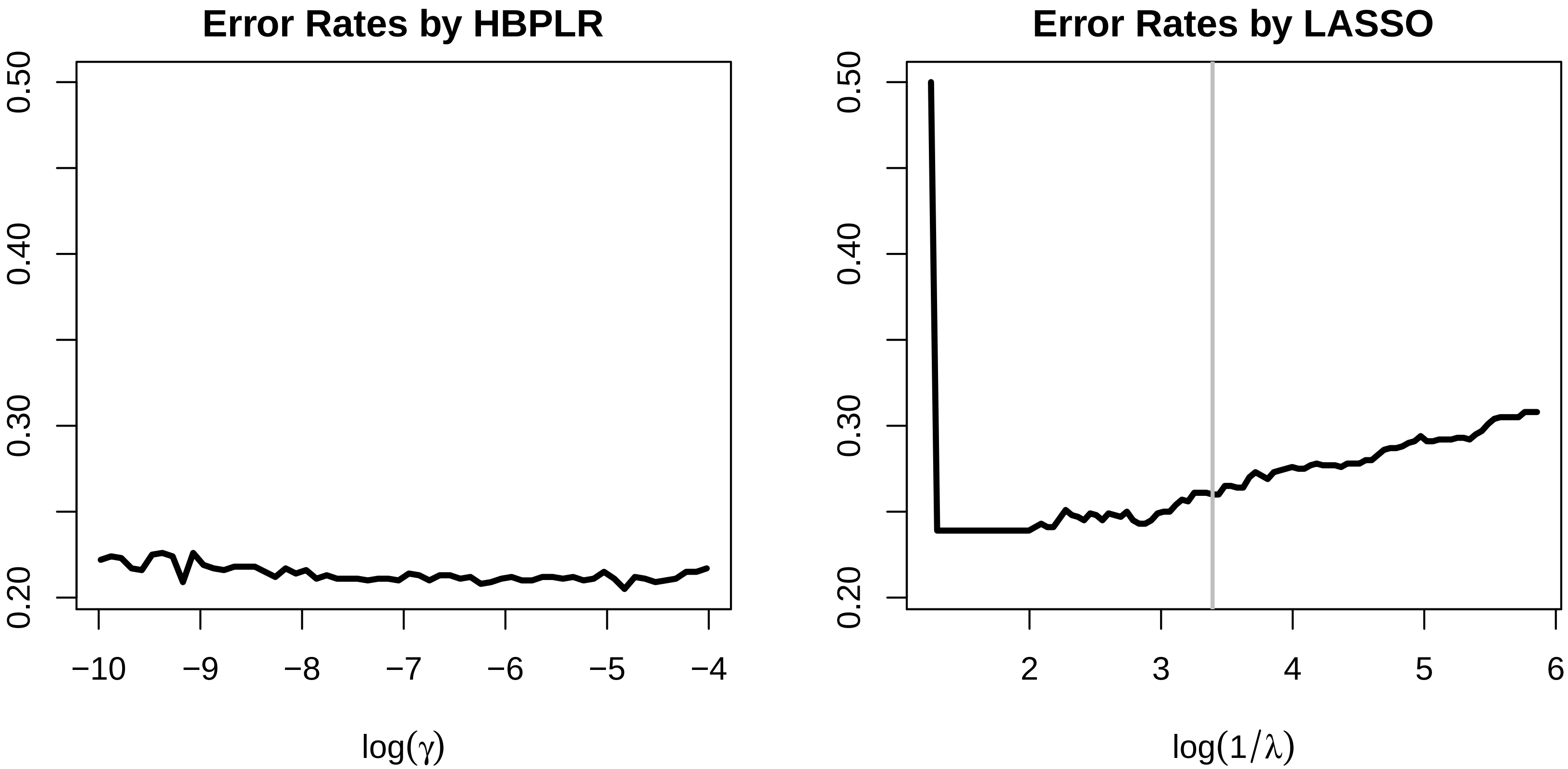
$\log(\gamma) = 0.5\log(w)$ is the log scale for t prior, and $\log(1/\lambda)$ is the scale of Laplace.

Comparisons of AMLPs

AMLP = average of minus log probabilities $= (1/N)\sum_{i=1}^{N} -\log(\hat{P}(y_i^*|\boldsymbol{x}_i^*))$



Comparisons of Error Rates



Simulation Studies with 50 Data Sets of $p = 2000$

Data Generation

The number of classes C is set to 3, and class labels are equally likely drawn from $1, 2,$ and 3. Values of 10 features for each case were generated as follows:

$$\begin{aligned} x_1|y=c &= \mu_{c,1} + z_1 + 0.5\epsilon_1, \\ x_2|y=c &= \mu_{c,2} + 2z_1 + z_2 + 0.5\epsilon_2, \\ x_j|y=c &= \mu_{c,j} + z_3 + 0.5\epsilon_j, \text{ for } j = 3, \ldots, 10, \end{aligned}$$

where,

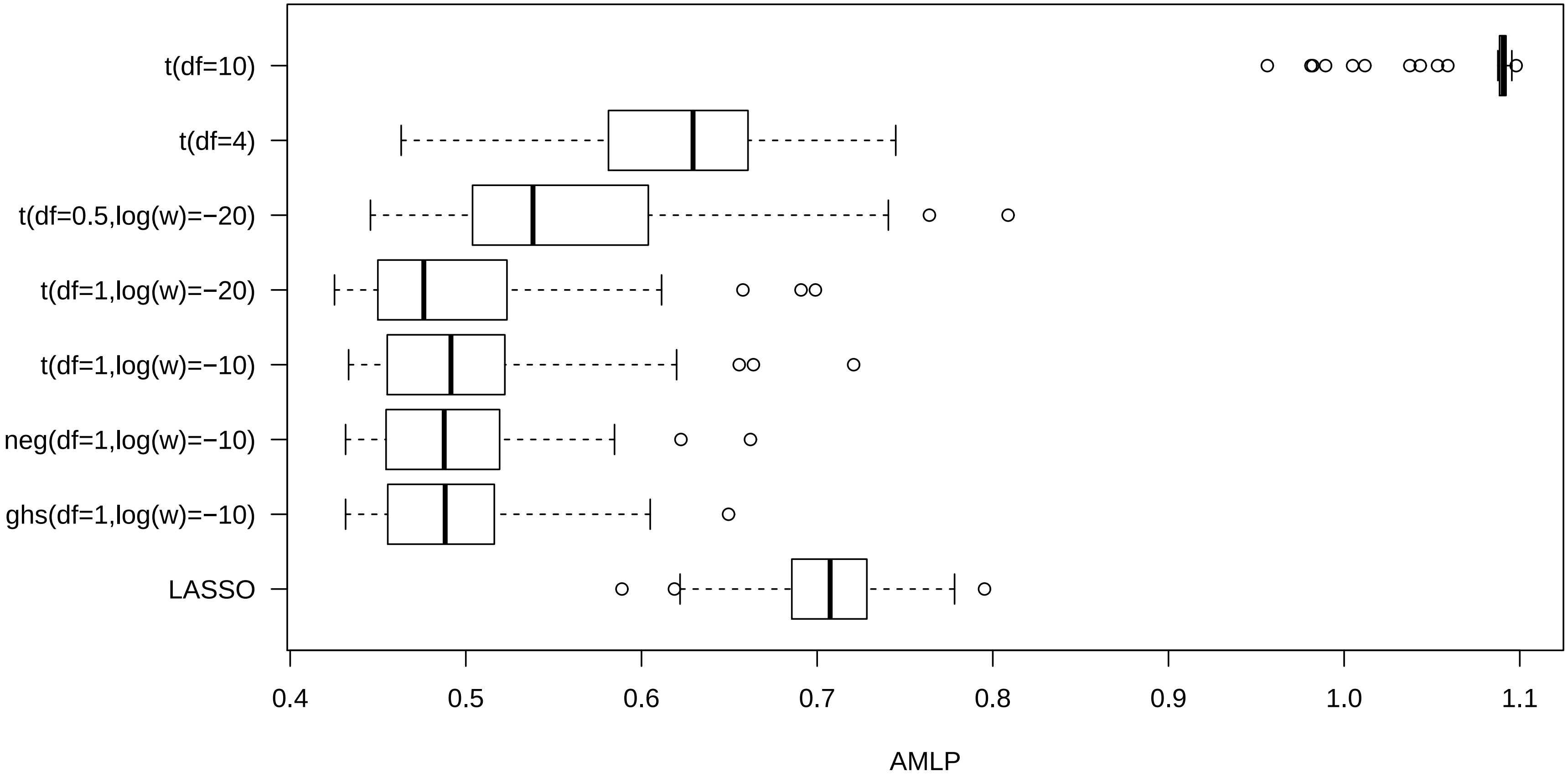
$$(\mu_{c,j})_{3\times 10} = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 2 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 2 & \ldots & 2 \end{pmatrix}, z_i, \text{and } \epsilon_j \sim N(0,1)$$

In this model, x_1 has different means in class 2 from classes 1 and 3, x_2 is positively correlated with x_1 with the same means in three classes, and x_{3-10} have different means in class 2 from classes 1 and 3, but are redundant. Another 1990 noise features with values drawn from $N(0,1)$ were then added to the data set.

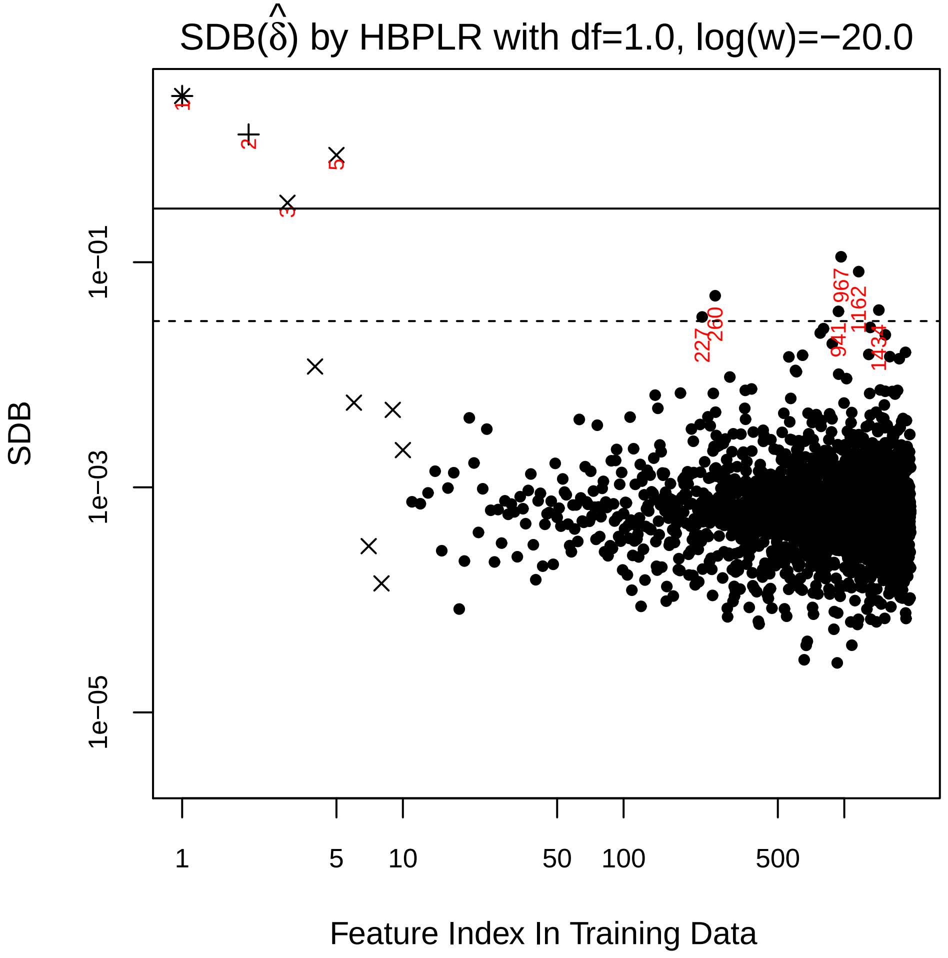
MCMC Settings and Computation Times

- Using t prior with $= 1$: 3 hours needed for running 500K iterations of Gibbs sampling with 50 leapfrog trajectories and with $\zeta = 0.05$ in restricted Gibbs sampling, for each data set
- Using Horseshoe and NEG priors with $= 1$: 8 hours needed for running Markov chains of the same settings as above
- All of others: much longer, some took over 30 hours

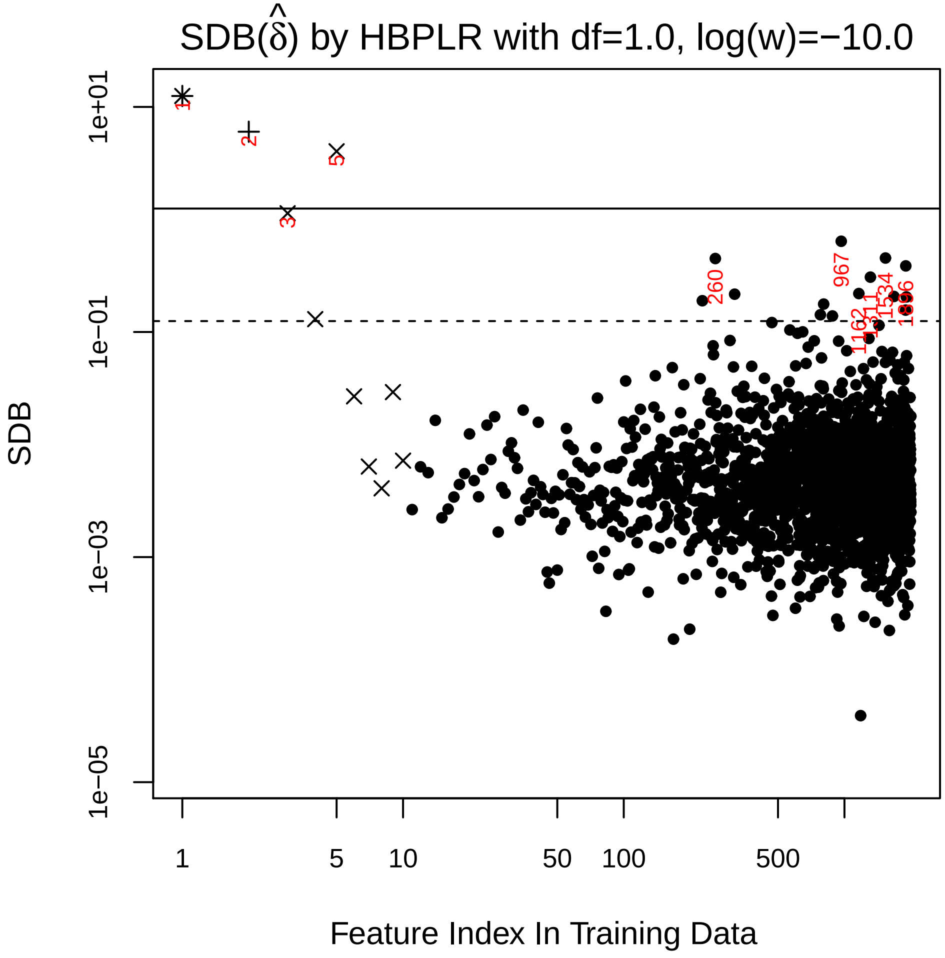
Comparisons of AMLPs



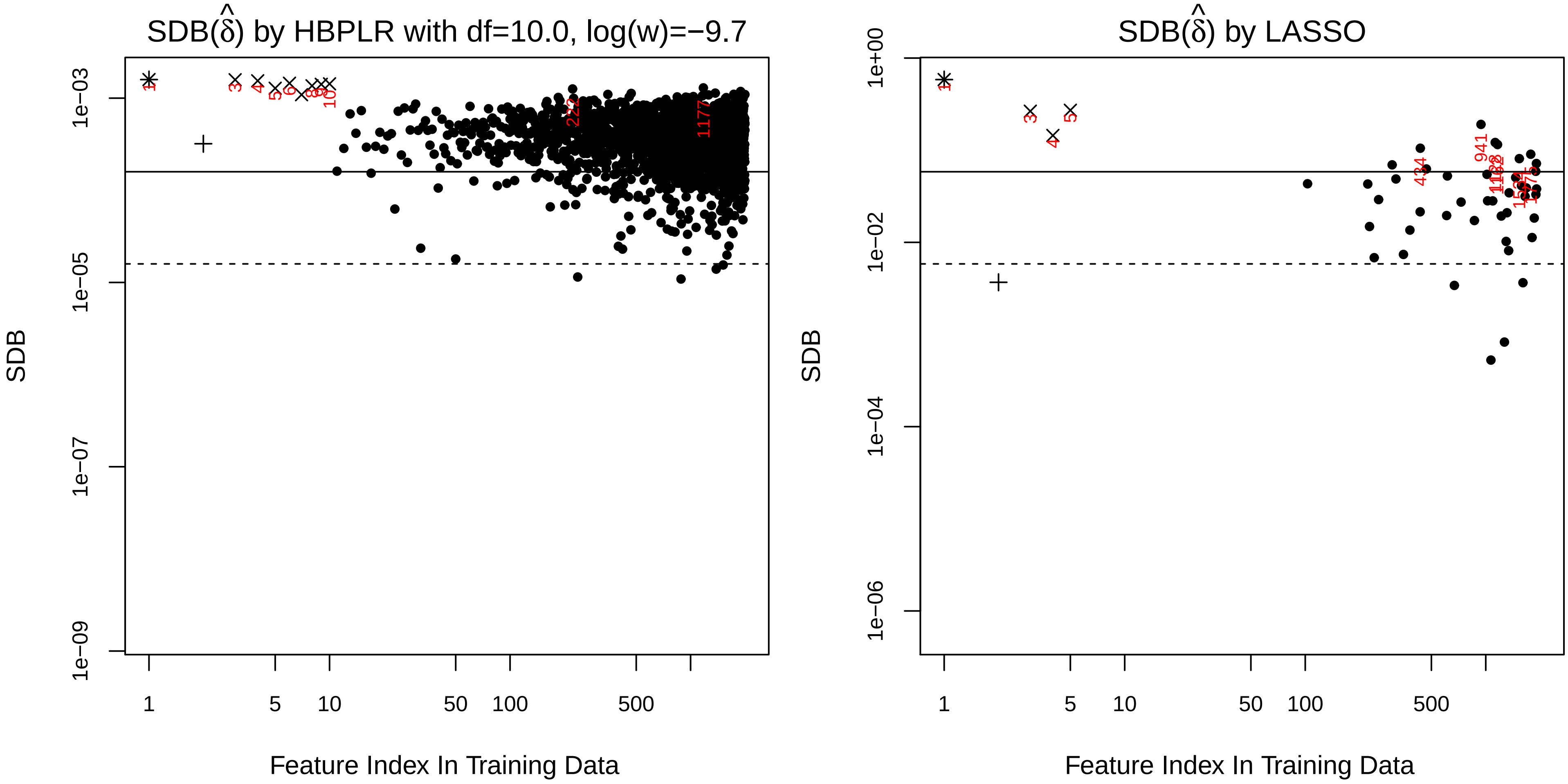
Comparisons of Feature Importance Indice

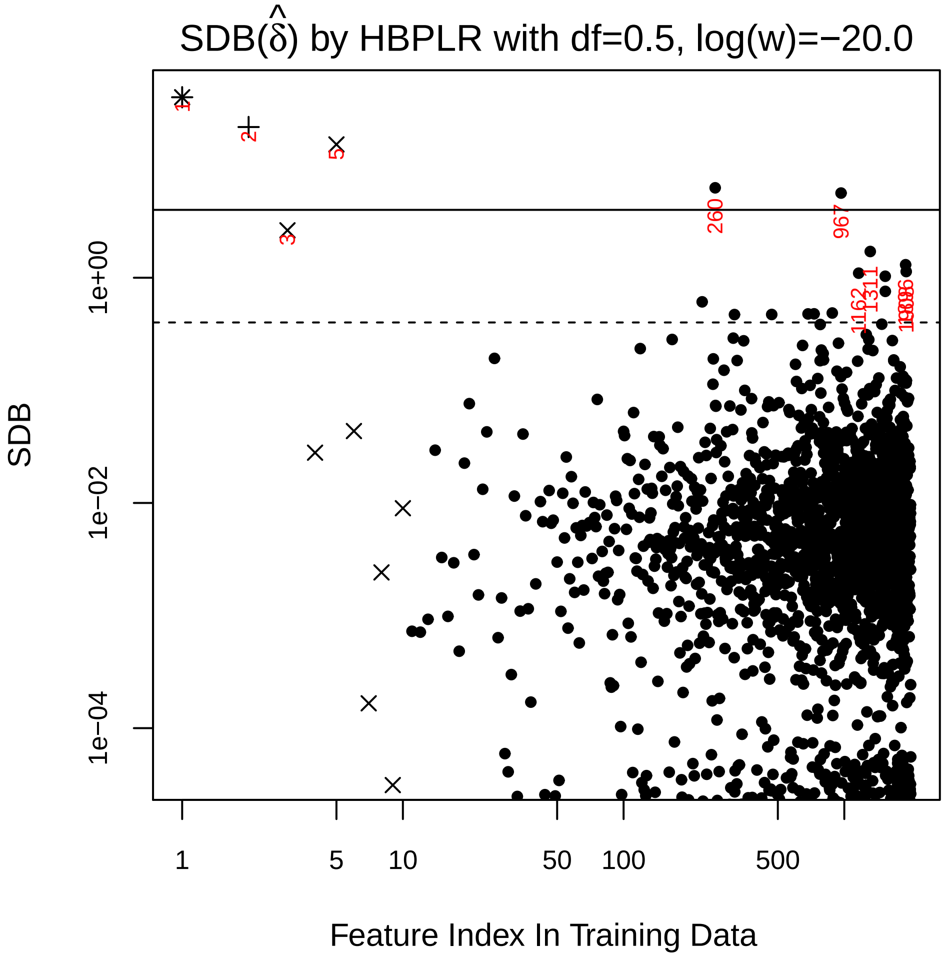


SDB($\hat{\delta}$) by HBPLR with df=1.0, log(w)=−20.0

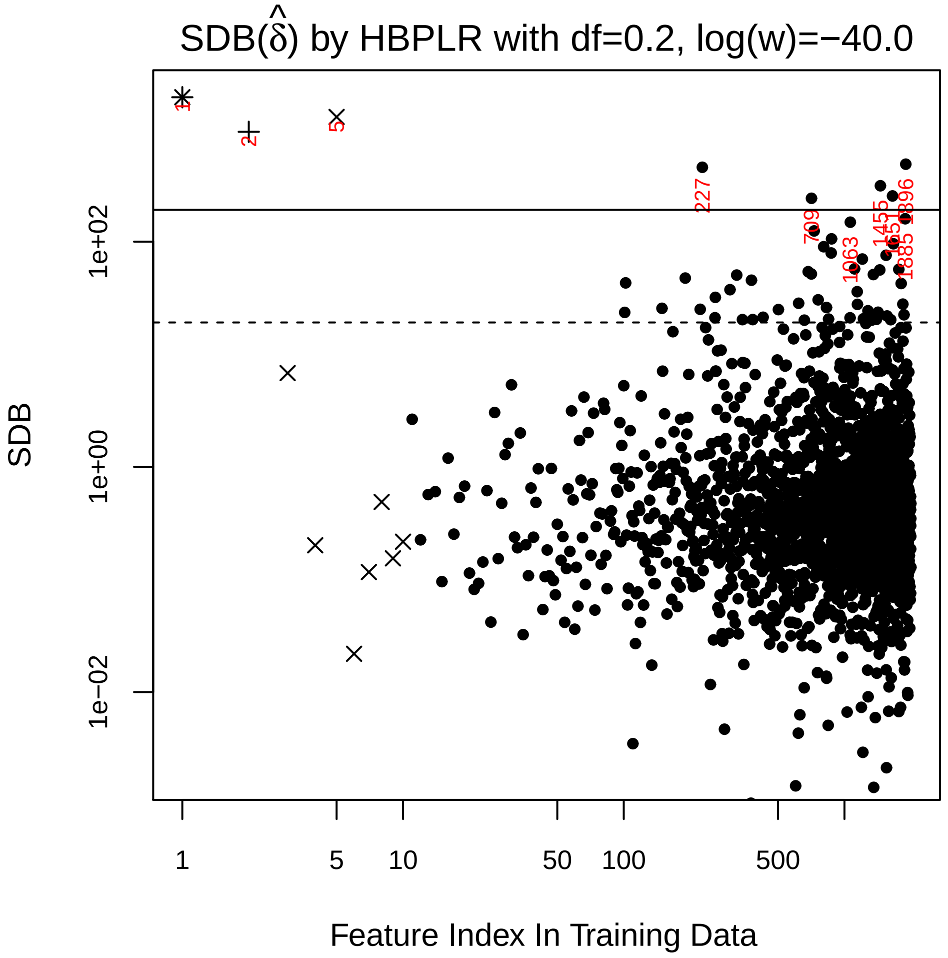


SDB($\hat{\delta}$) by HBPLR with df=1.0, log(w)=−10.0


SDB($\hat{\delta}$) by HBPLR with df=10.0, log(w)=−9.7
SDB
1e−03
1e−05
1e−07
1e−09
1
5
10
50
100
500
Feature Index In Training Data
1
3
4
5
6
8
9
10
222
1177
SDB($\hat{\delta}$) by LASSO
SDB
1e+00
1e−02
1e−04
1e−06
1
5
10
50
100
500
Feature Index In Training Data
1
3
4
5
434
941



SDB($\hat{\delta}$) by HBPLR with df=0.5, log(w)=−20.0
SDB
1e−04
1e−02
1e+00
Feature Index In Training Data
1
5
10
50
100
500
1
2
3
5
260
967
1162
1311
1896



SDB($\hat{\delta}$) by HBPLR with df=0.2, log(w)=−40.0
SDB
1e−02
1e+00
1e+02
Feature Index In Training Data
1
5
10
50
100
500
1
2
5
227
709
1063
1455
1651
1885
1896

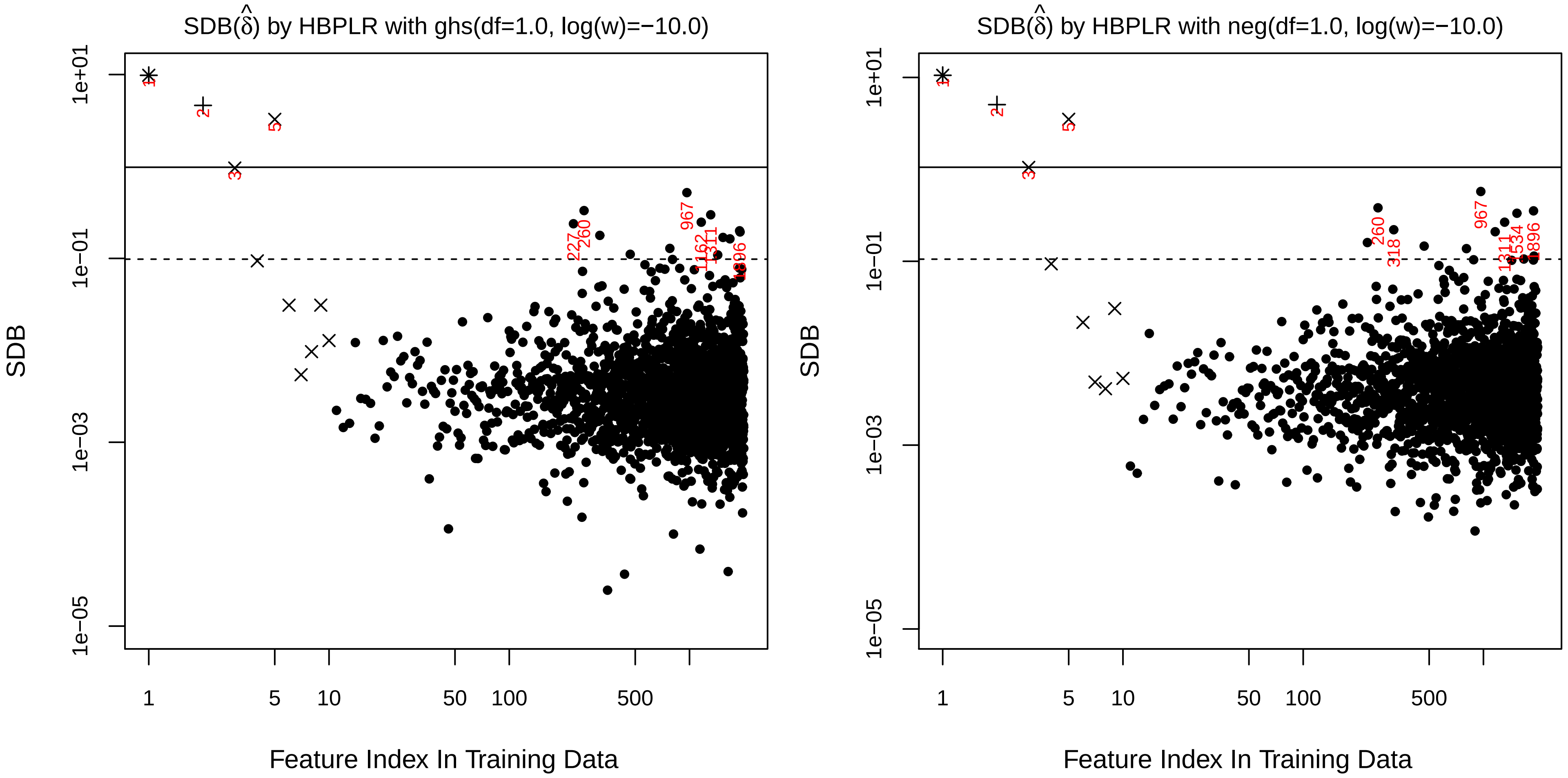

SDB($\hat{\delta}$) by HBPLR with ghs(df=1.0, log(w)=−10.0)
SDB
1e+01
1e−01
1e−03
1e−05
1
2
3
5
227
260
967
1162
1311
1896
1
5
10
50
100
500
Feature Index In Training Data
SDB($\hat{\delta}$) by HBPLR with neg(df=1.0, log(w)=−10.0)
SDB
1e+01
1e−01
1e−03
1e−05
1
2
3
5
260
318
967
1311
1534
1896
1
5
10
50
100
500
Feature Index In Training Data

Table of Feature Selection Results

Thresholding relative SDBs with 0.1, we determine whether a feature is selected. The mean numbers of selected features over 50 data sets, with bracked numbers showing sds.

	Groups of Features			
Methods	x_1	x_2	$x_3 - x_{10}$	$x_{11} - x_{2000}$
HBPLR with t (df=10)	0.96	0.66	7.42 (1.86)	1354 (580)
HBPLR with t (df=4)	1	0.36	1.26 (0.53)	0.00 (0.00)
HBPLR with t (df=0.2,$\log(w) = -40$)	1	0.72	1.36 (0.60)	5.74 (3.12)
HBPLR with t (df=0.5,$\log(w) = -20$)	1	0.98	1.16 (0.37)	1.14 (0.97)
HBPLR with t (df=1,$\log(w) = -20$)	1	0.94	1.14 (0.35)	0.16 (0.37)
HBPLR with t (df=1,$\log(w) = -10$)	1	0.96	1.10 (0.30)	0.32 (0.55)
HBPLR with GHS (df=1,$\log(w) = -10$)	1	1.00	1.14 (0.35)	0.30 (0.51)
HBPLR with NEG (df=1,$\log(w) = -10$)	1	1.00	1.06 (0.24)	0.28 (0.50)
LASSO	1	0.34	2.72 (1.18)	6.92 (4.97)

Analysis of a Microarray Data Set with $p = 6033$

Data Description

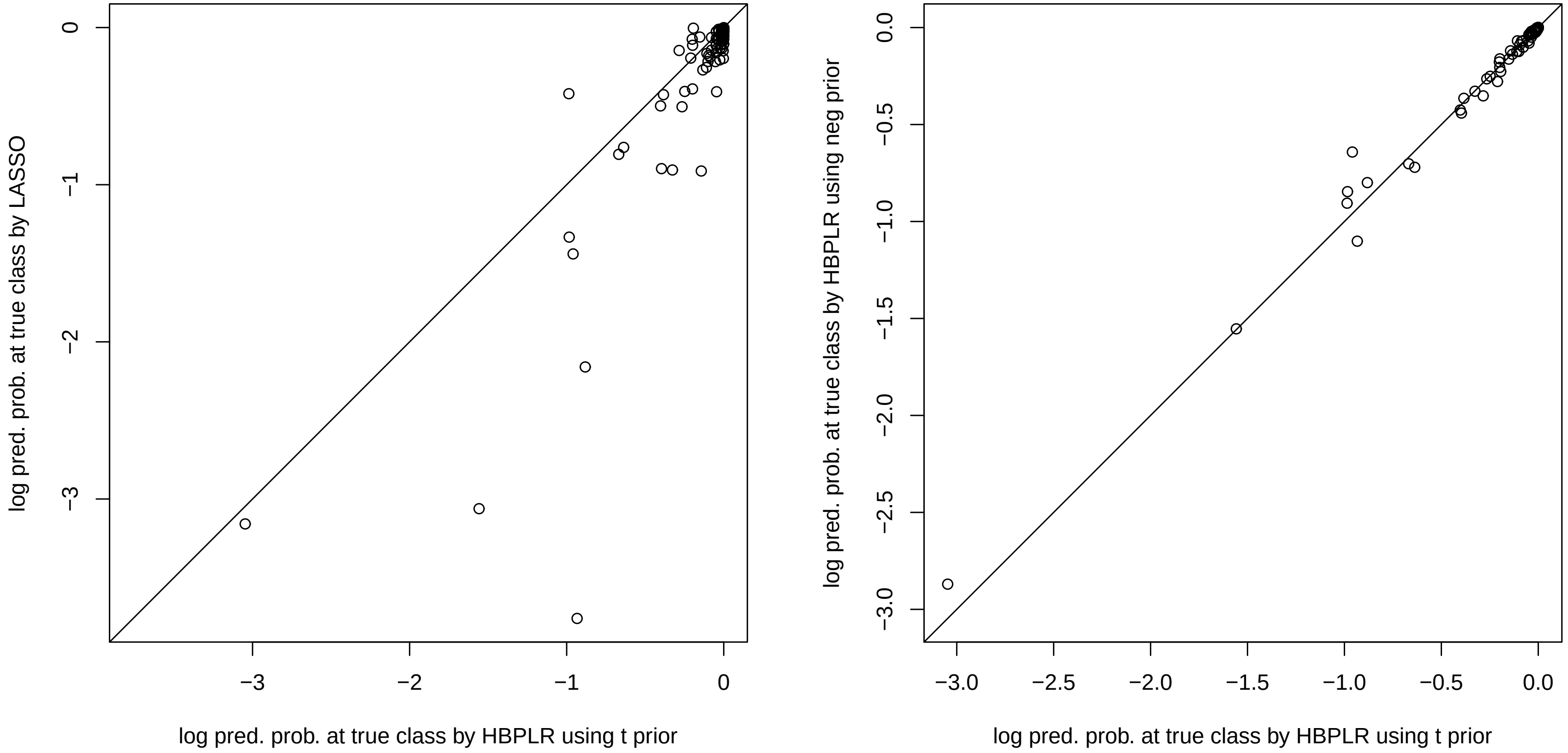
- The original data set was reported by Singh et al. (2002). We analyzed a data set downloaded from `http://stat.ethz.ch/~dettling/bagboost.html`.
- $n = 102$, 50 normal and 52 cancerous prostate tissues, $p = 6033$ genes
- For better looking at our results, we re-ordered the features by F-statistic on the whole data set, therefore the feature index is the rank based on F-statistic value.
- We standardized the data to have mean 0 and sd 1 in each split into training and test sets in leave-one-out crossvalidation.

MCMC Settings and Computation Time

- Using t Prior with df $=$ 1: 10 hours
  for running each MCMC of 1M iterations of Gibbs sampling with 50 leapfrog trajectories and setting $\zeta = 0.05$ in restricted Gibbs sampling.

- Using Horseshoe and NEG priors with df $=$ 1: 33 hours
  for running each MCMC with the same settings as above.

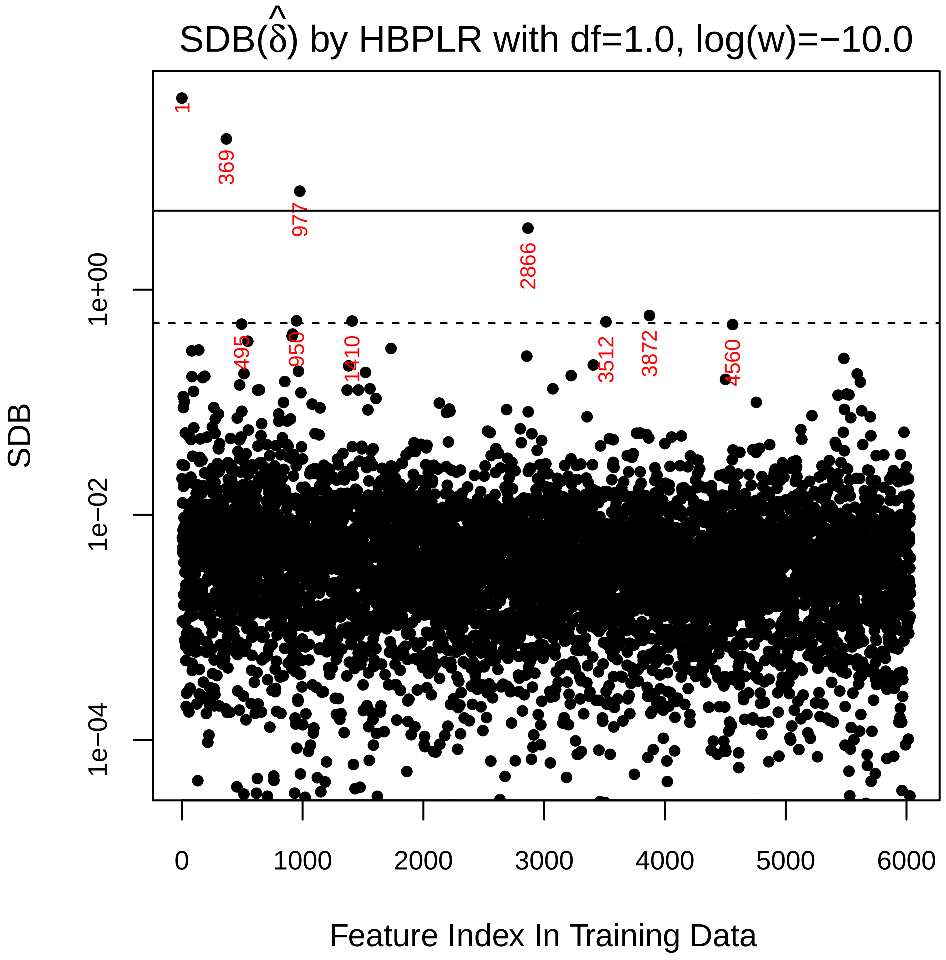
Log Predictive Probabilities at True Class Labels.

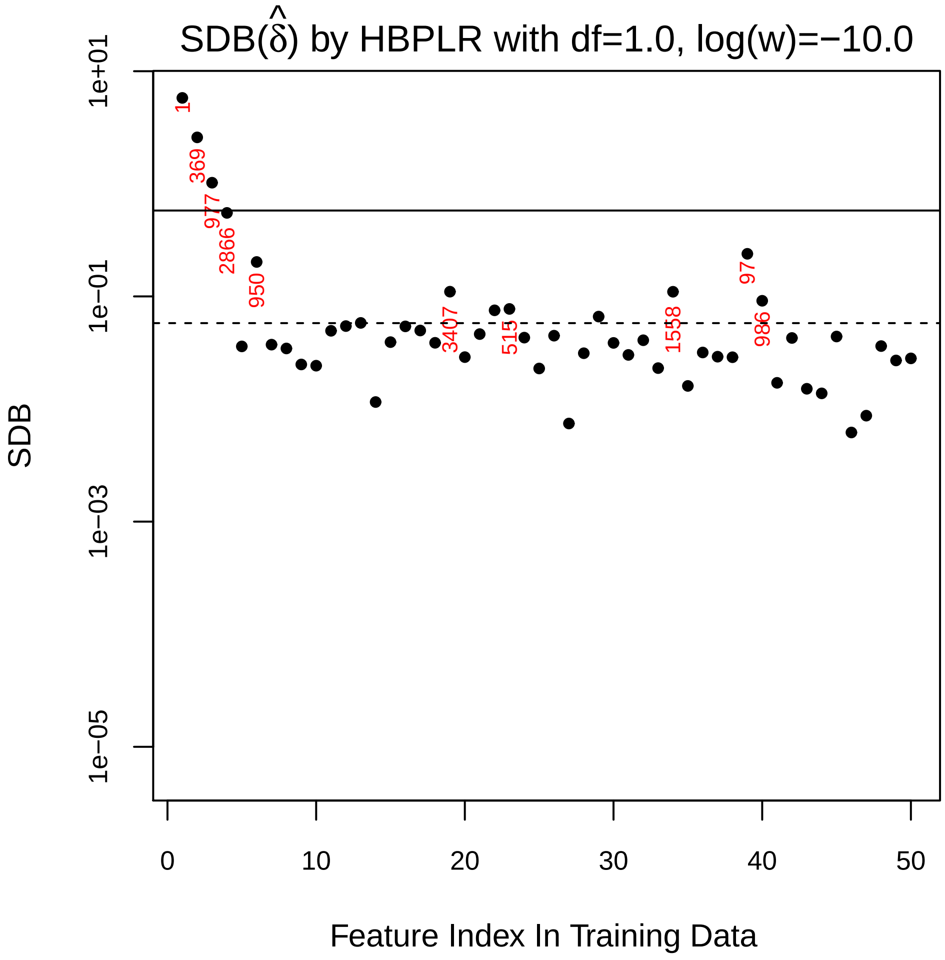


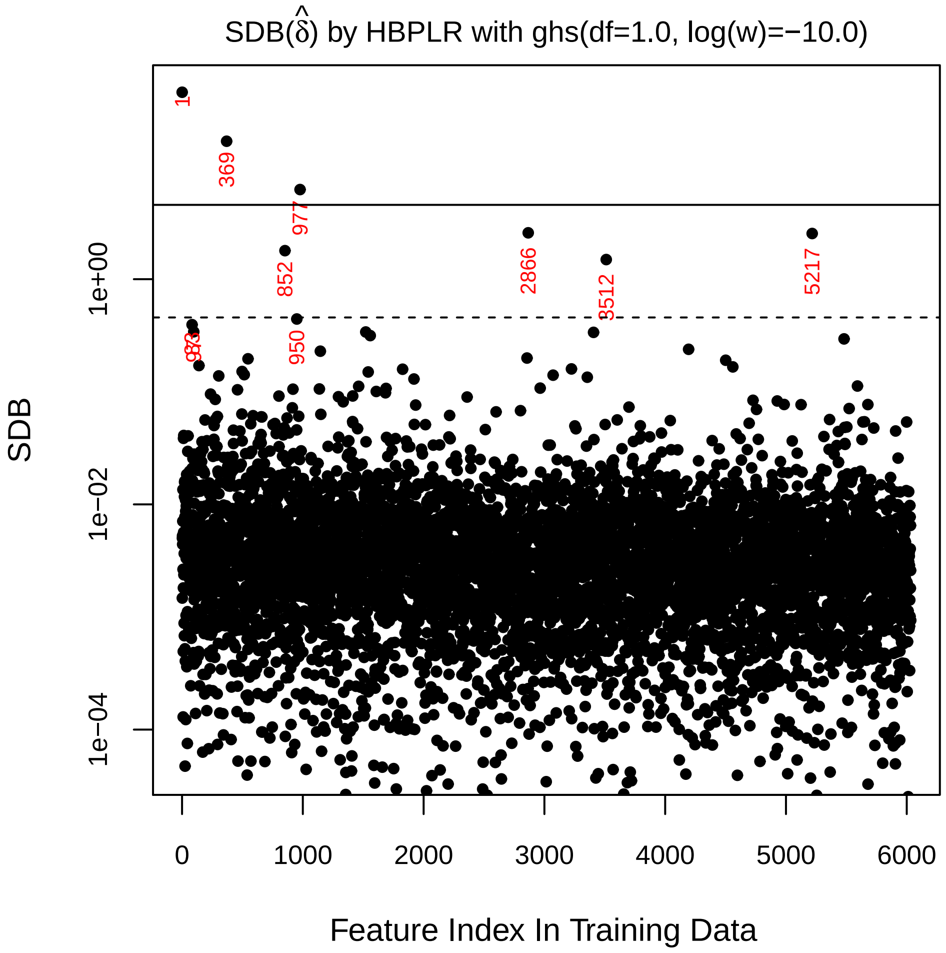
Comparisons of LOOCV Predictive Performances

Methods	HBt	HBghs	HBneg	LASSO	Bagboost	PAM	DLDA	SVM	RanFor	kNN
# genes	6033	6033	6033	6033	200	200	200	200	200	200
AMLP	.156	.158	.152	.274	-	-	-	-	-	-
ER (%)	6.86	7.84	7.84	10.8	7.53	16.5	14.2	7.88	9.00	10.59

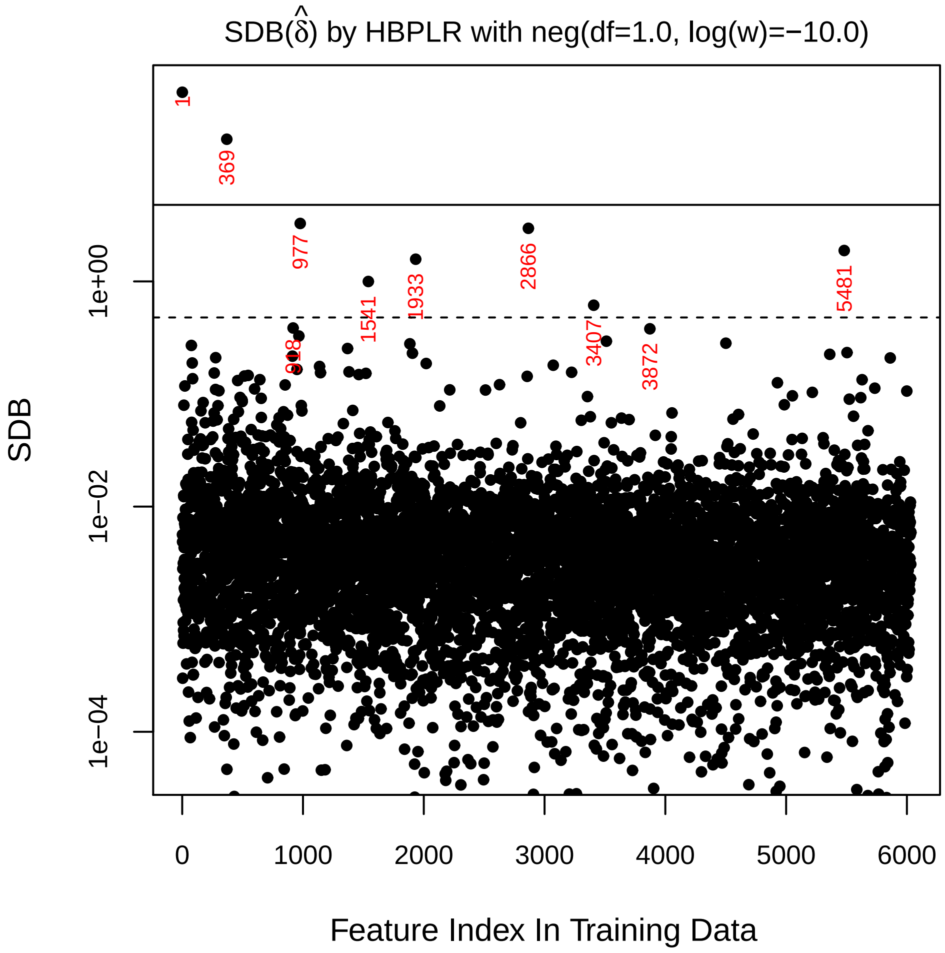
Feature Selection Results in 1 Fold



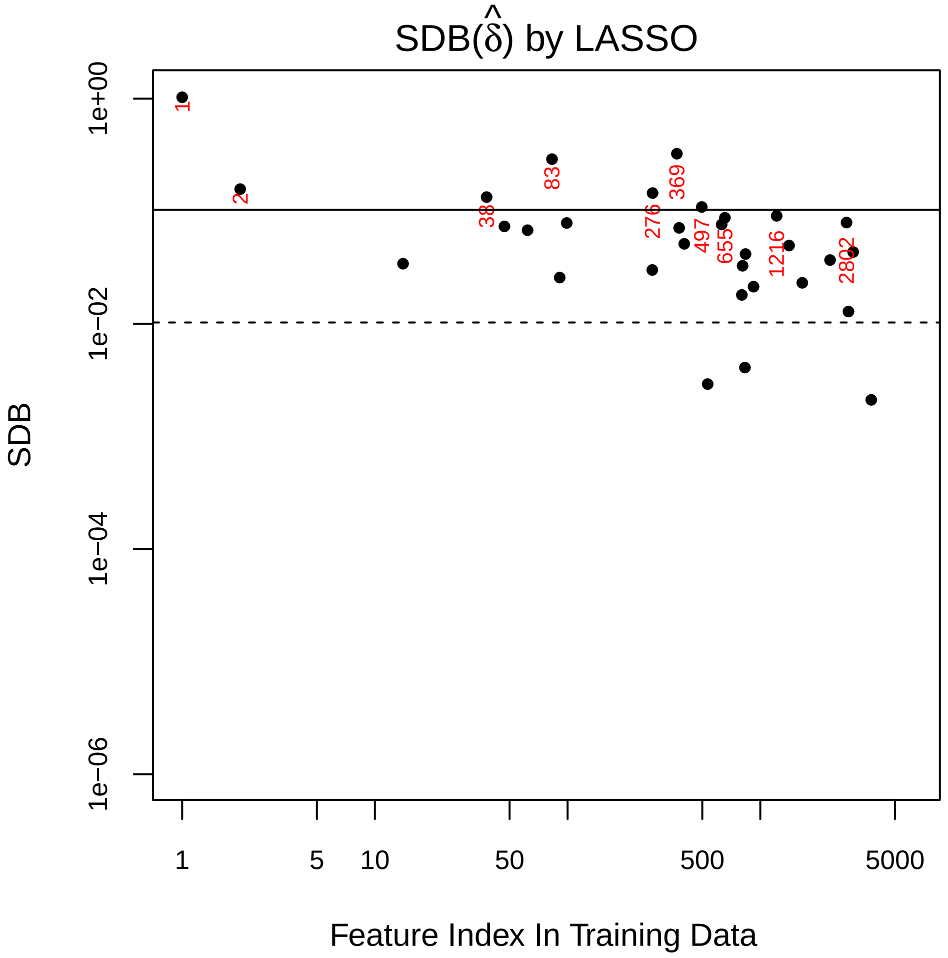




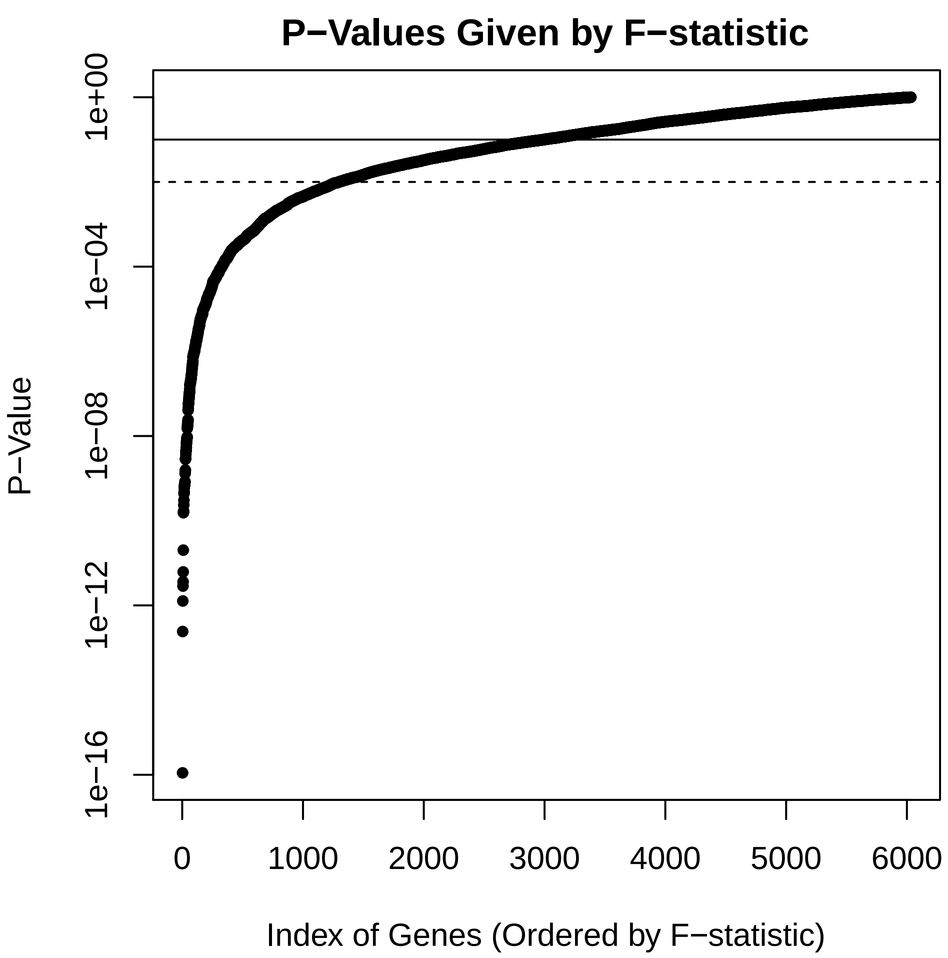
SDB($\hat{\delta}$) by HBPLR with ghs(df=1.0, log(w)=−10.0)



SDB($\hat{\delta}$) by HBPLR with neg(df=1.0, log(w)=−10.0)

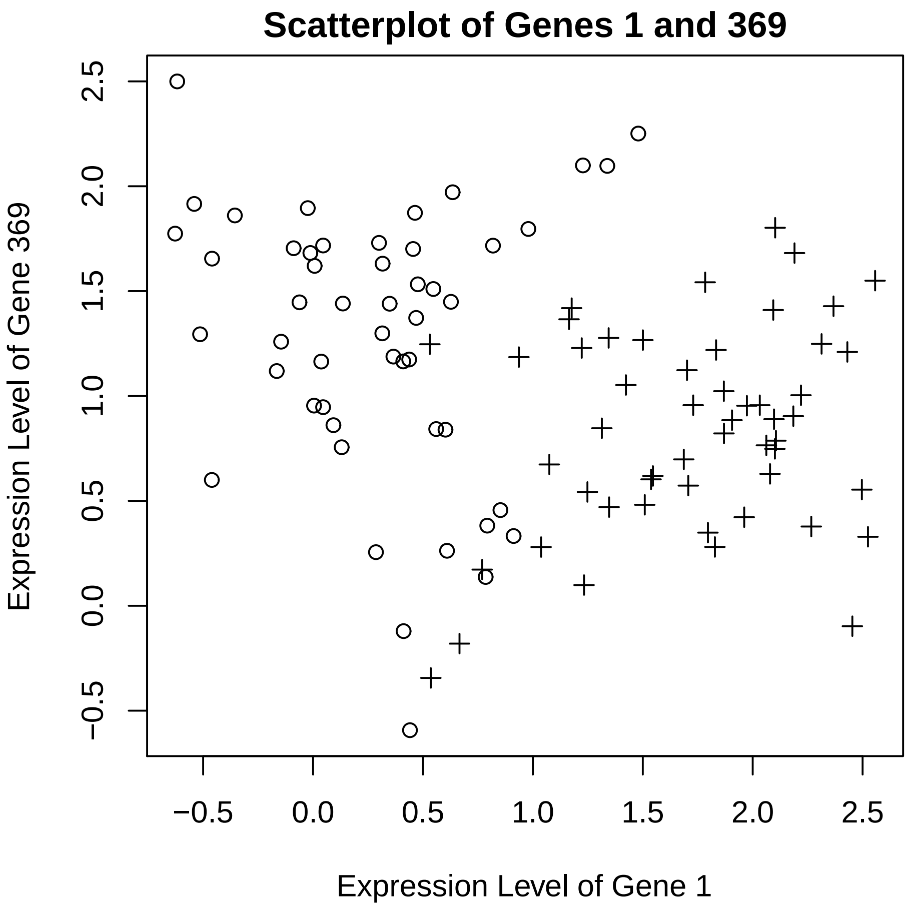


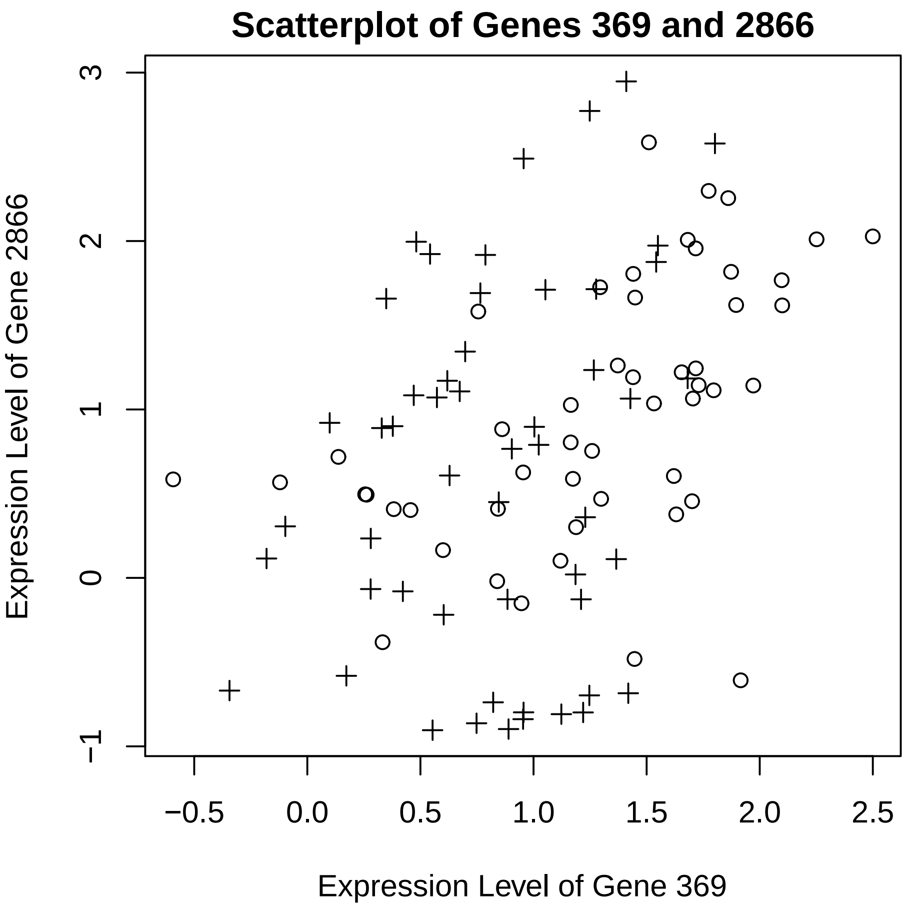
SDB($\hat{\delta}$) by LASSO

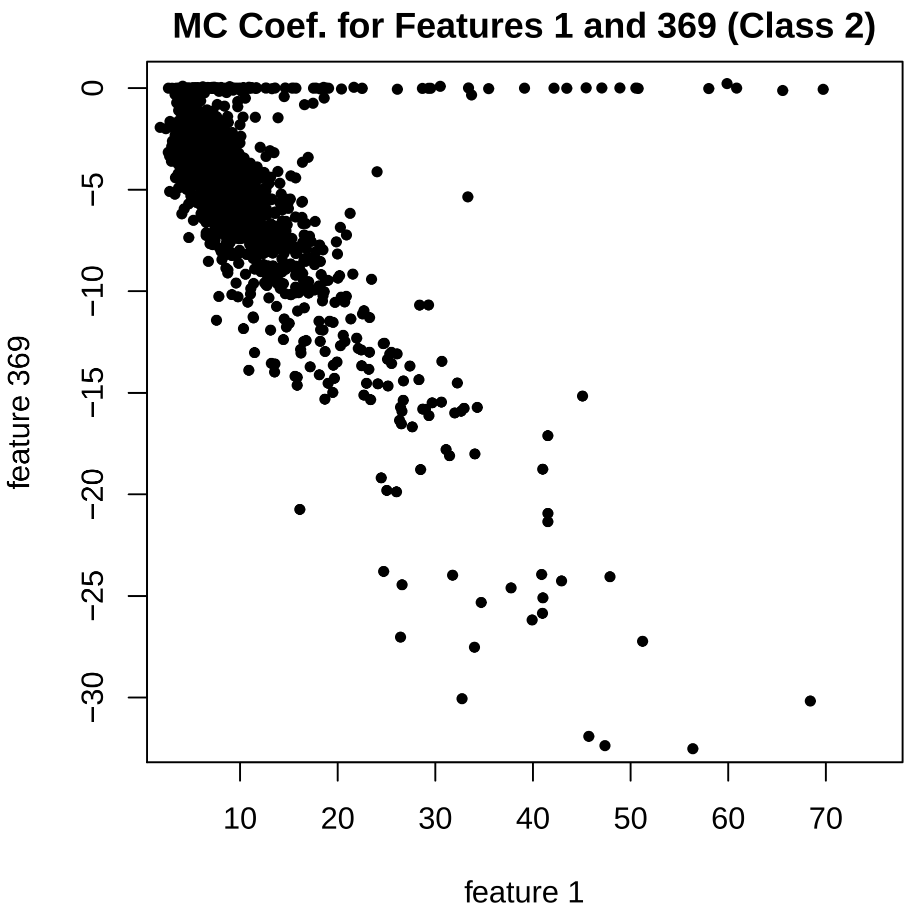


P−Values Given by F−statistic

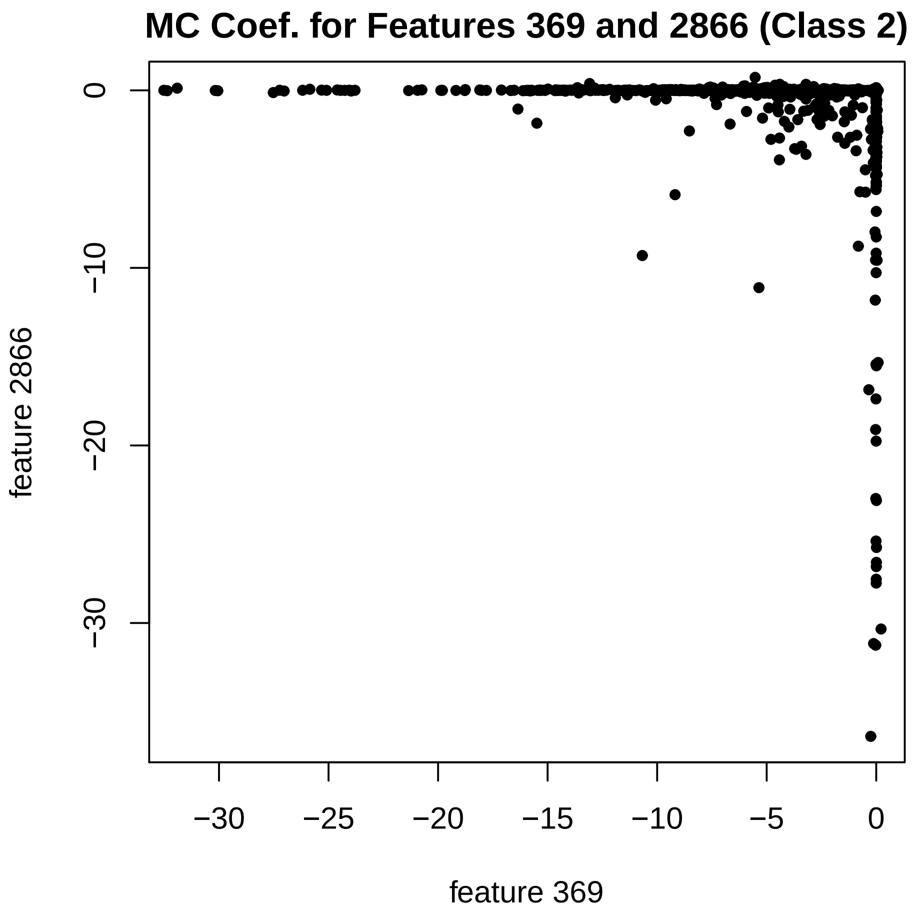
MCMC Samples and Scatterplots of Top Features







MC Coef. for Features 1 and 369 (Class 2)



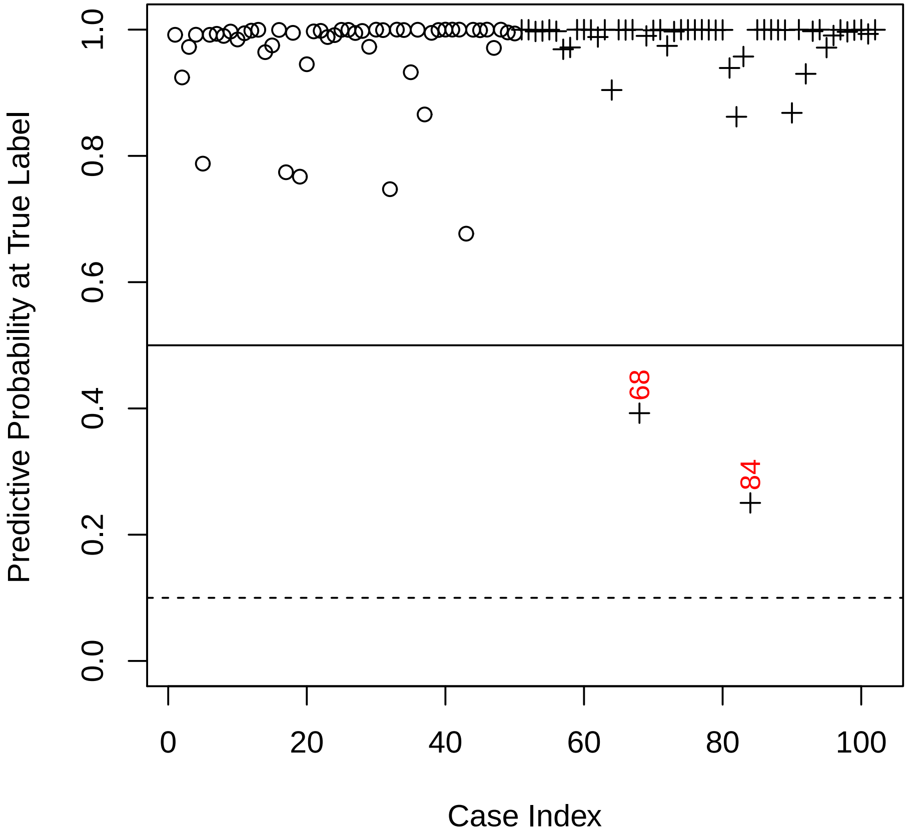
MC Coef. for Features 369 and 2866 (Class 2)

LOOCV Predictive Performances of Gene Subsets

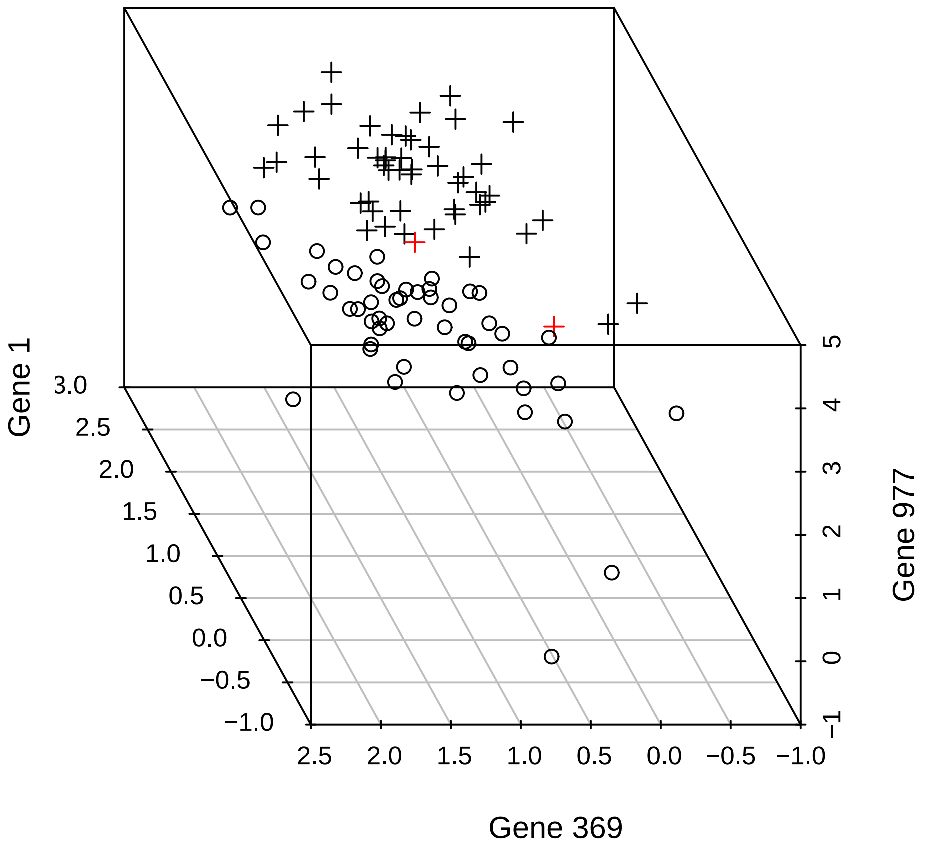
Gene subset	1, 369, 977	1, 369	1, 2, 3	1, 369, 83
Selected by	HBPLR	HBPLR and LASSO	F-Statistic	LASSO
AMLP	.050	.232	.240	.163
ER (%)	1.96	8.82	9.80	7.84

Looking at the Top 3 Genes Selected by HBPLR

LOOCV Prediction with Genes 1/369/977
AMLP = 0.050, Error Rate = 1.96% (2/102)



3D Scatterplot of Genes 1, 369 and 977



Conclusions

Bayesian logistic regression with moderately heavy-tailed, small scale and MCMC simulation works well for high-dimensional feature selection, and is feasible. It has a few good statistical properties:

- It can shrink small signals strongly towards 0 (due to small scale), but leave large signals unpunished (due to heavy tails).
- It can automatically separate a group of many redundant correlated features into different posterior modes, or eliminate many redundant and less differentiated features.
- The fitting results are stable for a wide range of small scales for heavy-tailed prior, as opposed to the instability of using "spike-and-slab" priors and the sensitivity of LASSO to the choice of scale.
- The fitting results are insensitive to initial values because MCMC can travel across many modes, as compared to penalized likelihood methods.

Discussions

- There is much room for improvement of the computational speed.
- The choice of heaviness of priors (degree freedom) is crucial for logistic regression. Our simulation studies show that df $= 1$ works better than bigger and smaller degree freedoms. This is also observed in regression problems. How to explain it theoretically?
- Ordinary t prior isn't so bad once we choose moderately small degree freedom and small scale. From our studies, the performance of t is almost the same as other more sophisticated priors. But the computation with using t prior is much faster. How much and when do we gain from using the more sophisticated priors?
- What's the best threshold in restricted Gibbs sampling? How much does it help sampling? Are there other more sophisticated methods for choosing more promising coefficients to update?
- Do we need to correct for the sknewness of the posterior of coefficients? Do we have other methods that don't get trapped in local modes?
- Using moderately heavy-tailed prior with small scale may be promising for many other high-dimensional problems. Used as priors for high-dimensional covariance matrix?

Acknowledgements

- Thanks to my collaborator Weixin Yao for his helpful contribution to the work.
- The work was inspired by my Ph.D. supervisor, Radford Neal.
- Thanks to NSERC and CFI for providing research funding for the research.
- Thank you, all of my audiences, for listening to me with your great patience.