

# Fully Bayesian Logistic Regression with Hyper-Lasso Priors for High-dimensional Feature Selection

Longhai Li\* and Weixin Yao†

12 May 2014

## Abstract

High-dimensional feature selection arises in many areas of modern sciences. For example, in genomic research we want to find the genes that can be used to separate tissues of different classes (eg. cancer and normal) from tens of thousands of genes that are active (expressed) in certain tissue cells. To this end, we wish to fit regression and classification models with a large number of features (also called variables, predictors), which is still a tremendous challenge to date. In the past few years, penalized likelihood methods for fitting regression models based on hyper-lasso penalization have been explored considerably in the literature. However, fully Bayesian methods that use Markov chain Monte Carlo (MCMC) for fitting regression and classification models with hyper-lasso priors are still lack of investigation. In this paper, we introduce a new class of methods for fitting Bayesian logistic regression models with hyper-lasso priors using Hamiltonian Monte Carlo in restricted Gibbs sampling framework. We call our methods BLRHL for short. We use simulation studies to test BLRHL by comparing to LASSO, and to investigate the problems of choosing heaviness and scale in BLRHL. The main findings are that the choice of heaviness of prior plays a critical role in BLRHL, and that BLRHL is relatively robust to the choice of prior scale. We further demonstrate and investigate BLRHL in an application to a real microarray data set related to prostate cancer, which confirms the previous findings. An R add-on package called BLRHL will be available from <http://math.usask.ca/~longhai/software/BLRHL>.

**Key phrases:** high-dimensional, feature selection, heavy-tailed prior, hyper-lasso priors, MCMC, Hamiltonian Monte Carlo, Gibbs sampling, fully Bayesian.

---

\*Correspondance author, Associate Professor, Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, S7N5E6, CANADA. e-mail: [longhai@math.usask.ca](mailto:longhai@math.usask.ca).

†Associate Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506, U.S.A. e-mail: [wxyao@ksu.edu](mailto:wxyao@ksu.edu).

# 1 Introduction

Today, high-throughput biotechnologies (such as microarray and sequence methods) can easily measure expression levels of thousands of genes. An interesting problem in genomic research is to identify the genes whose expression levels are relevant to a certain complex disease classification (eg. cancer or diabetes), called class label or response hereafter. Once such genes are found and verified by biological experiments, they can be used for prognosis or diagnosis of the disease. For this purpose, researchers collect some “training” tissues, for which their true class labels are known. Typically the number of training tissues is very small, such as *tens*, but the number of candidate genes is large, such as *thousands*. Identifying the most relevant genes for a disease from thousands of candidates is still a tremendous challenge to statisticians to date. An analogy is that we are looking for a few “needles” (useful features) from a huge “haystack” (irrelevant features).

The most commonly used method in today’s practice may be univariate methods that measure the strength of relationship between each gene and the class label, such as  $t$  or  $F$  tests, or model-based inference methods with independence assumption for genes within classes, such as DLDA (Dudoit et al., 2002), and PAM (Tibshirani et al., 2002). However, univariate methods ignore the correlations between genes, which are prevalent in gene expression data due to gene co-regulation, see Ma et al. (2007), Clarke et al. (2008) and Tolosi and Lengauer (2011) for real examples. The consequence is that many redundant differentiated genes are included, meanwhile, useful but weakly differentiated genes may be omitted. Methods by fitting classification models, which attempt to capture the conditional distribution of class label (response) given features, can take correlations among features into account. However, when the number of observations is not much larger than the number of features, maximizing likelihood of a classification model will overfit the data, with noise rather than signal captured. Therefore, when the number of features is greater than the number of observations, we need to shrink the coefficients towards 0 for avoiding overfitting.

One approach is to maximize a likelihood function with a penalty term, which can often interpreted as maximizing a posterior (MAP) based on a certain prior. The most widely method

is the LASSO (Tibshirani, 1996), which uses  $L_1$  penalty (or Laplace prior in Bayesian interpretation). However, Laplace cannot distinguish the “needles” and “hay” very well due to its light tails. In the last 15 years, many researchers have worked on proposing priors with tails heavier than Laplace distribution, called hyper-lasso priors (Griffin and Brown, 2011) or global-local priors (Polson and Scott, 2010), for high-dimensional regression problems. A common goal of these priors is to shrink each coefficient differently so that large signal can be less penalized while small signals are shrunk towards 0. These priors are often given in the form of scale mixture of normal or Laplace distributions. Such alternative penalties or priors to Laplace include, but not limited to: SCAD (Fan and Li, 2001), horseshoe and normal/inverted-beta priors (Gelman, 2006; Carvalho et al., 2009, 2010; Polson and Scott, 2012c; van der Pas et al., 2014), NEG priors (Griffin and Brown, 2011), Normal/Gamma (Caron and Doucet, 2008; Griffin and Brown, 2010), adaptive LASSO (Zou, 2006), generalized double-pareto (Armagan et al., 2010) and Dirichlet-Laplace and Dirichlet-Gaussian (Bhattacharya et al., 2012). For unified reviews of these penalties, one can refer to Kyung et al. (2010); Polson and Scott (2010, 2012a,b); in addition, Bhattacharya et al. (2012) compare a broad class of shrinkage priors in terms of prior and posterior concentration, and show that the commonly used Laplace is suboptimal in high-dimensional problems.

In Bayesian community, the most popular approach is to explore a posterior based on “spike-and-slab” priors — mixture distributions of a point mass at 0 and a continuous distribution, see Mitchell and Beauchamp (1988); Brown et al. (1998); Sha et al. (2004); Ishwaran and Rao (2005a,b); Hans et al. (2007); Bottolo and Richardson (2010); Fan and Zhang (2010); Guan and Stephens (2011); Rockova and George (2013), among many others. Such priors may express the best our prior belief. However, the inference results using this approach are highly sensitive to the choice of width of the continuous part. If the width is large enough, any features, no matter how large signals they carry, will be classified as “irrelevant”, with data information totally ignored. The sensitivity has been discussed in the context of regression problems theoretically in Denison et al. (2002, p.23), and empirically in Lamnisos et al. (2011); indeed, the problem has long been known as Jeffreys-Lindley’s paradox regarding Bayesian test of null hypothesis about a point (Lindley, 1957; Shafer,

1982; Robert, 2013).

The fully Bayesian approach that uses Markov chain Monte Carlo (MCMC) to explore a posterior based on hyper-lasso priors is an important alternative. However, to our knowledge, Bayesian approach is still lack of exploration in the literature to date. The reason is probably the difficulty in computation. First of all, one needs to compute the likelihood function of a large number of coefficients. In addition, posterior distributions based on hyper-lasso priors are highly multimodal, and within each mode some coefficients may be highly correlated, as will be discussed later. Therefore, if we explore the posterior with ordinary MCMC simulations, such as with one-by-one Gibbs sampling, we may see that using priors of heavy-tails (such as Cauchy) does not produce better results than using priors with lighter tails, such as Gaussian (see eg. Li and Neal, 2008). Therefore, to our knowledge, fully Bayesian methods appearing in the literature still use Laplace or  $t$  with large degree freedom as priors for the sake of computational ease attributed to unimodality or smoothness between modes; but the sacrifice is that their applications are restricted to problems of small dimensions. For example, Laplace prior is used in Bayesian lasso methods (Park and Casella, 2008; Hans, 2009; Li et al., 2011; Cai et al., 2011), among many others; Bae and Mallick (2004) use  $t$  with 6 degree freedom as a prior for a problem of only 500 features; Yi and Xu (2008) use Laplace prior and  $t$  with degree freedom treated as hyperparameter, but test their methods with a problem of only 127 features. Recently, there have been a few papers reporting fully Bayesian methods using hyper-lasso priors, for example, Alhamzawi et al. (2012); Leng et al. (2014). However, their empirical studies are confined to problems with only tens or ones of features. On the other hand, because of the multimodality in posterior distributions, fully Bayesian methods have potential advantage that a well-designed MCMC algorithm can explore multiple modes and choose among them according to the integrals under the modes; whereas penalized likelihood method may end up with a local mode that depends much on an arbitrary choice of initial value (see eg. Griffin and Brown, 2011).

In this paper, we show how to achieve the benefit of hyper-lasso penalization — shrink small signals strongly towards 0, but leave large signals less punished than Laplace — in fully Bayesian

methods so that they can be applied to problems of higher dimension, such as thousands. In particular, we have developed a class of fully Bayesian logistic regression methods using hyper-lasso priors (shortened by **BLRHL**) with tails similar to Cauchy ( $t$  with 1 degree freedom). In BLRHL we use Hamiltonian Monte Carlo (Neal, 2010) in restricted Gibbs sampling framework to explore the multiple modes of the posterior. This paper focuses on reporting our empirical results from investigating the issues of choosing heaviness and scale of priors in fully Bayesian methods. Our empirical studies demonstrate the following three points, which appear unreported elsewhere in the literature, but are important to the applicability of fully Bayesian methods using hyper-lasso priors. First, the choice of heaviness should be appropriate, which cannot be too large or too small; our empirical results show that Cauchy is optimal. Second, the fitting results are relatively stable for the choice of scale when Cauchy prior is chosen; therefore, it is unnecessary to treat the scale as hyperparameter (which greatly increases MCMC simulation difficulty). Third, in fully Bayesian methods, the tails of hyper-lasso priors play more important role than their shapes at 0.

This article will be structured as follows. In Section 2, we first discuss some properties of hyper-lasso priors by comparing to Laplace and Gaussian using geometric illustration. In Section 3, we describe BLRHL in technical details. In Section 4 we use simulated data sets to test our method and investigate the issue of choosing heaviness and scale. In Section 5, we report the analysis results by applying our methods to a real microarray data set with  $p = 6033$  related to prostate cancer. The article is concluded in Section 6 with discussions.

## 2 Preliminaries of Hyper-Lasso Priors

We first consider the simple logistic regression model for binary class label for examine two important properties of hyper-lasso priors. Suppose we have collected data of features and responses (class labels) on  $n$  training cases. For a case indexed by  $i$ , we denote its class label by  $y_i$ , which can take integers 1 and 2, and denote  $p$  features associated with it by a row vector  $\mathbf{x}_{i,1:p}$ . The logistic regression model for the data is:

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

Table 1: 4 scale-mixture-normal distributions.  $\alpha$  is denoted by “df” in Figure 1

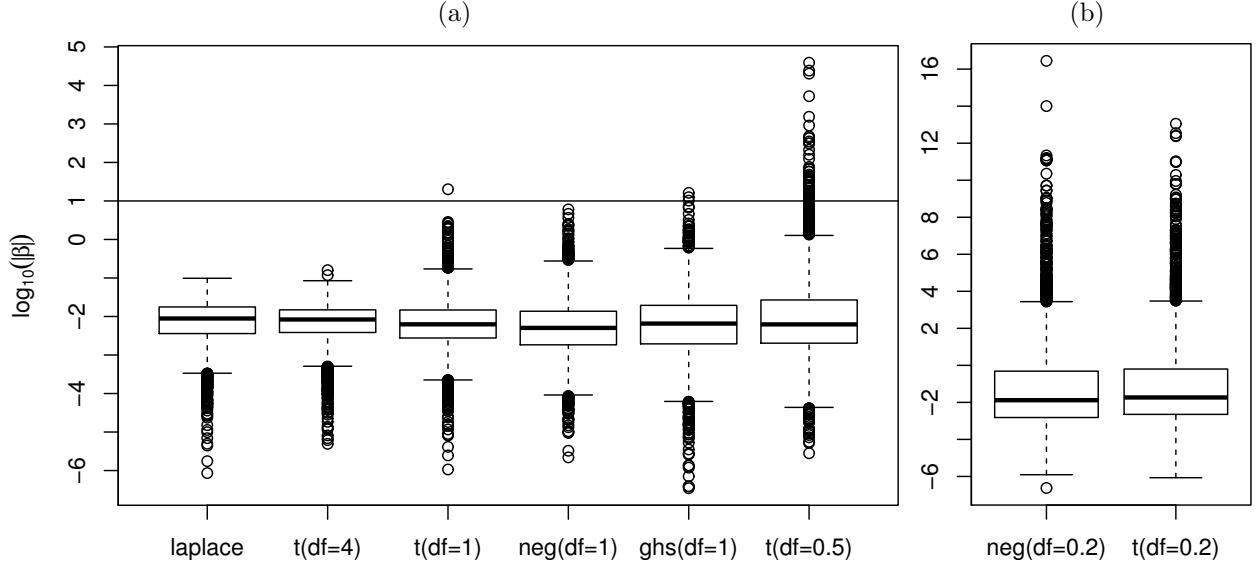
Name	Random Numbers Generator	Parameters used in Figure 1
$t$	$N(0, 1) \times \sqrt{\text{IG}(\alpha/2, \alpha/2)} \times \gamma$	four groups of $(\alpha, \log(\gamma))$ used: (4, -4.5)/(1, -5)/(0.5, -5.5)/(0.2, -6)
GHS	$N(0, 1) \times  N(0, 1)  \times \sqrt{\text{IG}(\alpha/2, \alpha/2)} \times \gamma$	$(\alpha, \log(\gamma)) = (1, -4.5)$
NEG	$N(0, 1) \times \sqrt{\exp(1)} \times \sqrt{\text{IG}(\alpha/2, \alpha/2)} \times \gamma$	$(\alpha, \log(\gamma)) : (1, -5)/(0.2, -6)$
Laplace	$N(0, 1) \times \sqrt{\exp(1)} \times \gamma$	$\log(\gamma) = -4$

for  $k = 0$  and  $1$ ,  $i = 1, \dots, n$ , where  $\beta_{1:p}$  is a column vector of regression coefficients, and  $I(\cdot)$  is indicator function, which is equal to 1 if the condition in bracket is true, 0 otherwise. We will assume that all the features are commensurable and we will select features by looking at the values of  $\beta_{1:p}$ . We will consider how to infer  $\beta_{1:p}$  from the training data.

We consider scale-mixture-normal (SMN) distributions as priors for  $\beta_{1:p}$ . The simplest choice is  $t$  distribution with df  $\alpha$  and scale  $\gamma$ , which can be expressed with two-level conditional distributions:  $\beta_j | \sigma_j^2 \sim N(0, \sigma_j^2)$ ,  $\sigma_j^2 \sim \text{IG}(\alpha/2, \alpha\gamma^2/2)$ , where  $\text{IG}(a, b)$  stands for Inverse-Gamma distribution, the distribution of the inverse of a Gamma random variable with shape parameter  $a$  and *rate* parameter  $b$ . In terms of random numbers generator,  $t$  is the distribution of the products of two independent random variables and a scale  $\gamma$ :  $N(0, 1) \times \sqrt{\text{IG}(\alpha/2, \alpha/2)} \times \gamma$ . Similarly, Laplace is the distribution of  $N(0, 1) \times \sqrt{\exp(1)} \times \gamma$ , where  $\gamma$  is a scale, and  $\exp(1)$  is the standard exponential random variable. Note that a Laplace distribution parametrized by  $\lambda$  with PDF  $(\lambda/2)e^{-\lambda|\beta_j|}$  has our scale  $\gamma = \sqrt{2}/\lambda$ . Recently, some other penalties with SMN interpretation, including Horseshoe (Carvalho et al., 2010), and Normal-Exp-Gamma (NEG) (Griffin and Brown, 2011), are shown to be superior than LASSO in high-dimensional regression problems. In the original Horseshoe prior, positive half Cauchy prior is assigned to  $\sigma_j$ . Here we naturally generalize half Cauchy to half  $t$  for uniformity of notations for all the priors considered in this article, and call the prior GHS. In Table 1, we describe them with their random numbers generators. The detailed descriptions of GHS and NEG are given in Section 3.1 (equations (5) and (6)).

We generated 4000 of random  $\beta_j$  from different SMN priors to look at their appropriateness in expressing our belief that a couple of “needles” are present in a huge “haystack”. For this, we

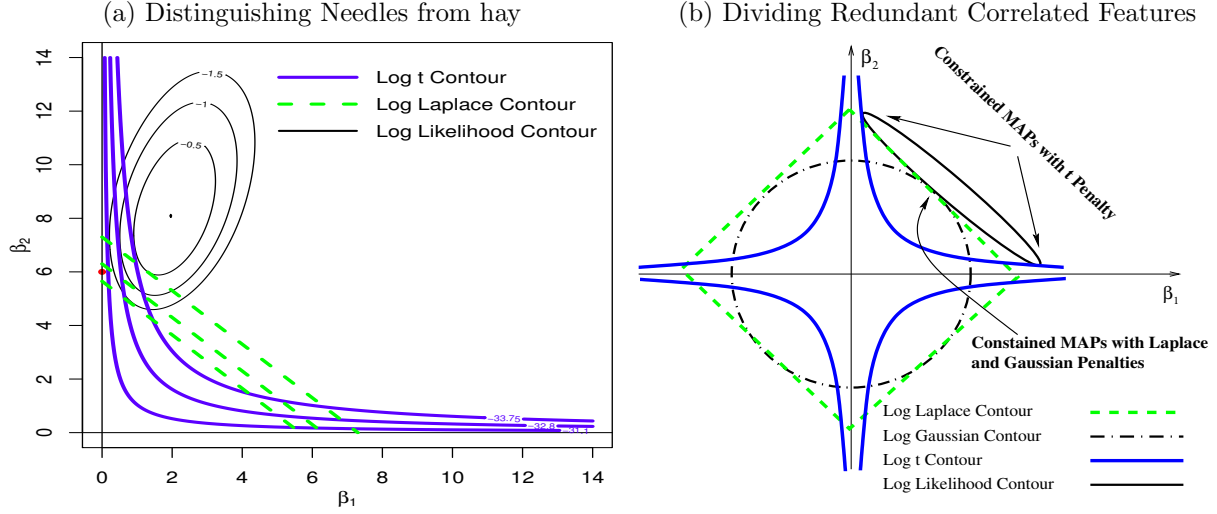
Figure 1: Boxplots of 4000  $\log_{10}(|\beta_j|)$  generated from different SMN priors. The scales can be found from the last column of Table 1.



desire that most of  $|\beta_j|$  generated from a prior are very small, but a few are very large for modelling the “needles”. To look at their tails more easily, we set the scale for each distribution such that the medians of  $\log_{10}(|\beta_j|)$  are around  $-2$ , with exact values listed in Table 1. Figure 1 shows the boxplots of  $\log_{10}(|\beta_j|)$  for 8 SMN distributions. Note that, with other larger or smaller scale, the boxplots of  $\log_{10}(|\beta_j|)$  only move up or down, without shape change. From Figure 1a, we see that  $t$ , NEG, and GHS with small dfs, such as 1, have moderately heavy tails, allowing a few  $|\beta_j|$  much larger than a great many of “hay”, expressing well our belief.  $t$  with large df (such as 4) and Laplace have very short upward tails, with coefficients very close to each others. On the other hand,  $t$  and NEG with overly heavy tails with very small dfs, such as 0.2, are not good either — the upward tails of such priors become overly flat, allowing values from  $10^{-6}$  to  $10^{16}$ , which can’t be true.

The property of moderately heavy-tailed priors in better separating “needles” from “hay” can be explained also by looking at the “path” of constrained MAPs (maximizer of posterior) — the MAP with the log likelihood constrained to a particular value (ie on a contour). A constrained MAP can be found by shrinking the contour lines of log priors toward the origin until the two lines are tangent. Figure 2a shows three such constrained MAPs for  $t$  with  $df = 0.5$  and  $\gamma = e^{-10}$ , and three for Laplace, based on a data set generated with true coefficients  $\beta_1 = 0, \beta_2 = 6, \beta_0 = 0$ .

Figure 2: Geometric illustrations of the properties of  $t$  penalty in MAP inference.



The (unconstrained) MAP can then be found from the “path” containing these constrained MAPs. Because the contour lines of  $\log t$  prior indent into the origin, the path of constrained MAPs based on  $t$  prior is flatter (with respect to x-axis) than the path based on Laplace; starting from the MLEs, the path based on  $t$  prior goes to a point at which  $\beta_1$  is very close to 0 (but not exact 0), and  $\beta_2$  is close to its true value (6), whereas, the path based on Laplace goes to the origin. Therefore we see that  $t$  prior can shrink small “hay” without much punishment to large “needles”.

From looking at constrained MAPs, we also find that hyper-lasso penalties can *automatically* divide a group of correlated features into different posterior local modes. Figure 2b shows a conceptual illustration not based on a simulated data set. When two features are highly correlated, a contour line of log likelihood is negatively correlated as shown in Figure 2b. With  $t$  penalty, the constrained MAPs are at the two ends of the contour of log likelihood, each of which uses only one of them to explain the class label without under-estimating the importance of each of them. Therefore, the coefficients of highly correlated features are divided into different modes, each using only one of them, see Figure 11d for a demonstration with real data. When the predictive abilities of the correlated features are different,  $t$  prior can also make selection among a group of highly correlated features automatically. The selection within groups is necessary in high-dimensional problems in which often a large group of correlated features exist. By contrast, with Laplace and Gaussian



penalties, the constrained MAPs are in the middle of the contour, favoring using all features with coefficients of smaller absolute values to explain the class label. When the group of correlated features is large, they may under-estimate the absolute values of all of them, and hence, miss all of them, see a detailed discussion by [Tolosi and Lengauer \(2011\)](#).

Figure 2b is also helpful for seeing that the primary computational difficulty of using hyper-lasso priors in classification and regression problems is the existing of many local modes in the posterior distribution. An optimization algorithm can easily get trapped in a minor local mode arbitrarily depending on the initial values, so the algorithm becomes unstable and some sophisticated methods for choosing the initial values are required, see [Griffin and Brown \(2011\)](#). Therefore, full Bayesian methods have potential advantage in that they can explore and choose among many of the modes according to the integrals under the modes.

### 3 Bayesian Logistic Regression with Hyper-Lasso Priors

We will now describe our method, BLRHL, including some technical details. Throughout this article, we will denote matrices with **bold-faced** letters, with row indexes displayed in the first subscript, and column indexes in the second; we denote real-valued vectors with **bold-faced** letters too, but with only a set of indexes in subscript; the indexes of matrices and vectors are denoted by  $i:j$  — integers from  $i$  to  $j$ , or a single integer for a row or column.

#### 3.1 The Model

Suppose we have collected data of features and responses (class labels) on  $n$  training cases. For a case indexed by  $i$ , we denote its class label by  $y_i$ , which can take integers  $1, \dots, C$ , and denote  $p$  features associated with it by a row vector  $\mathbf{x}_{i,1:p}$ . The hierarchical Bayesian polychotomous logistic regression model used by us is described as follows:

$$P(y_i = c | \mathbf{x}_{i,1:p}, \boldsymbol{\beta}_{0:p,1:C}) = \frac{\exp(\beta_{0,c} + \mathbf{x}_{i,1:p} \boldsymbol{\beta}_{1:p,c})}{\sum_{c=1}^C \exp(\beta_{0,c} + \mathbf{x}_{i,1:p} \boldsymbol{\beta}_{1:p,c})}, \text{ for } c = 1, \dots, C, \quad (2)$$

$$\boldsymbol{\beta}_{j,1:C} | \sigma_j^2 \sim N(0, \sigma_j^2), \text{ for } j = 0, \dots, p, \quad (3)$$

$$\sigma_{1:p}^2 \sim \text{IG}(\alpha/2, w\alpha/2), \quad (4)$$

where  $\beta_{0,p,1:C}$  are regression coefficients, and other variables are called hyperparameters, which are introduced to define the prior for  $\beta_{1,p,1:C}$ , and for convenience in MCMC sampling.

In this hierarchy,  $\sigma_j^2$  indicates the importance of  $j$ th feature — the feature with larger  $\sigma_j^2$  is more useful for predicting  $y$ , provided that all features are commensurable (which can be enforced by standardization). Note that we fix  $\sigma_0^2$ , not controlled by  $w$ , because we believe that the variability of intercepts is quite different from the variability of  $\beta_{j,1:C}$  for features. With  $\sigma_{1,p}^2$  marginalized with respect to IG prior (4), equations (3) and (4) assign  $\beta_{j,1:C}$  ( $j > 0$ ) with  $C$ -dimensional  $t$  prior with df  $\alpha$ , and  $I_C \times \gamma$  as its covariance, whose PDF can be found from [Kotz and Nadarajah \(2004\)](#). For  $C = 2$ , this bivariate  $t$  prior is equivalent to assigning independent  $t$  priors for  $p$  coefficients  $\delta_j = \beta_{j,2} - \beta_{j,1}$ , as explained in Section 3.2.

As alternatives of (4), other priors for  $\sigma_{1,p}^2$  have been proposed in the recent literature for regression problems with the goal of shrinking  $\sigma_{1,p}^2$  with weak signals more towards 0. [Carvalho et al. \(2010\)](#) propose Horseshoe prior for coefficients by assigning half (positive) Cauchy distribution for  $\sigma_{1,p}$ . For uniformity of notations, we describe half Cauchy distribution using half  $t$  with various degree freedoms for  $\sigma_j$ , inducing a prior for  $\sigma_j^2$ :

$$P_{ghs}(\sigma_j^2) = \frac{\Gamma((\alpha + 1)/2)}{\Gamma(\alpha/2)\sqrt{\alpha\pi}\sqrt{w}} \left( \frac{1}{1 + \sigma_j^2/(\alpha w)} \right)^{\frac{\alpha+1}{2}} \frac{1}{(\sigma_j^2)^{1/2}}, \quad \text{for } \sigma_j^2 > 0. \quad (5)$$

[Griffin and Brown \(2011\)](#) propose NEG prior for coefficients by assigning exp-gamma prior for  $\sigma_{1,p}^2$ :  $\sigma_j^2 | \psi_j \sim \exp(\frac{1}{\psi_j})$ ,  $\psi_j \sim \text{IG}(\alpha/2, \alpha w/2)$ , where  $\psi_j$  is the mean parameter of exp distribution. We can marginalize  $\psi_j$ , and obtain a closed-form PDF for  $\sigma_j^2$ :

$$P_{neg}(\sigma_j^2) = \frac{\kappa}{\lambda} \left( \frac{1}{1 + \sigma_j^2/\lambda} \right)^{\alpha/2+1}, \quad (6)$$

where  $\kappa = \alpha/2$  and  $\lambda = \alpha w/2$  for notational simplicity. We will call (5) and (6) as **GHS** and **NEG** priors for  $\sigma_j^2$ , though the names are also used for the priors for coefficients. The PDFs of GHS and NEG priors for  $\sigma_j^2$  do not converge to 0 as  $\sigma_j^2$  goes to 0 (as IG prior does), therefore possibly regression coefficients are better shrunk towards 0 without punishing large signals. This property is indeed desired. Our numerical studies that follows confirm the improvement over IG

prior, though it seems very little for logistic regression models, if small scale  $\sqrt{w}$  is used. On the other hand, the sampling methods with adaptive rejection sampling (ARS, [Gilks and Wild, 1992](#)) for the posteriors of  $\sigma_j^2$  given  $\beta_{j,1:C}$  based on GHS and NEG priors are substantially slower than the standard sampling method for the IG posterior based on IG prior, which increase the total Markov chain sampling time greatly, for example from 3 hours to 8 or 10 hours in one of our examples.

The value  $\sqrt{w}$  (denoted by  $\gamma$  previously) is the scale of prior distribution for regression coefficients. We recommend to fix  $\alpha$  and  $w$  at some reasonable values, for example,  $\alpha = 1$ , and  $\log(w) = -10$ . Most random numbers generated by  $t$ /GHS/NEG distributions with this setting are very small, but contain a few values between 0 and 2, which can model a wide range of problems. Our following empirical results will show that the fitting results are relatively stable to the choice of very small  $w$ . Especially, due to heavy tails, we do not need to worry that very small values may over-shrink the signals. However, when the  $\sqrt{w}$  is too large (such as greater than 0.1), the posterior is close to the likelihood function without any penalty, which is very diffuse in upper tails. We therefore should avoid using overly large scale. An alternative method for avoiding choosing  $w$  is to treat  $w$  as a hyperparameter such that it will be adjusted with actual data information. However, when  $\alpha$  is small (such as in Cauchy with  $\alpha = 1$ ), very little information about  $w$  can be captured from  $\sigma_{1:p}^2$ , *because the conditional distribution of  $w$  (given a set of fixed  $\sigma_{1:p}^2$  that are mostly very close to 0) is very diffuse*. This property for hyper-lasso priors differ greatly from Gaussian and  $t$  with large degree freedom, for which the conditional distribution concentrates much on a point estimate determined by  $\sigma_{1:p}^2$ . The consequence is that Markov chain may stick to some region that overfit the data (with very high likelihood value) for a very long time with a large value  $w$ , before it pins down  $w$ . This greatly increases the difficulty of MCMC sampling, which takes much more time, and also demands a sophisticated tuning of MCMC sampling parameters to accommodate different posterior shape determined by  $w$ . Therefore, we recommend fixing  $w$  to some reasonable values according to our prior knowledge in fully Bayesian approach.

### 3.2 BLRHL Model with Identifiable and Symmetric Priors

An important issue in *polychotomous* logistic regression models is that the coefficients  $\beta_{j,1:C}$  is non-identifiable — if we add a constant to all  $\beta_{j,1:C}$ , the conditional distribution of  $y$  given  $\mathbf{x}$  in (2) is *exactly* the same. Therefore, the data can identify only the differences of  $\beta_{j,1:C}$  to a “baseline” class, say class 1, denoted by  $\delta_{j,k} = \beta_{j,k+1} - \beta_{j,1}$ , for  $k = 1, \dots, C - 1$ . To avoid the non-identifiability problem, the coefficient for a baseline class, say  $\beta_{j,1}$ , is often fixed at 0, and  $\beta_{j,2:C}$  is assigned with a prior as in (3). However, such a prior is *asymmetric* for all classes: the prior variance of  $\beta_{j,c} - \beta_{j,c'}$  ( $c, c' \neq 1$ ) double that of  $\beta_{j,c} - \beta_{j,1}$ . This implication may not be justified for practical problems. In addition, the inferences, especially feature selection results, can vary greatly with the choice of baseline class. The common variance  $\sigma_j^2$  for all  $\beta_{j,1:C}$  makes its posterior identifiable in theory. With fixed  $\delta_{j,1:(C-1)}$ , for varying  $\beta_{j,1}$ , the normal prior [equation (3)] for  $(\beta_{j,1}, \beta_{j,1} + \delta_{j,1}, \dots, \beta_{j,1} + \delta_{j,C-1})$  is maximized at  $\beta_{j,1} = -(1/C) \sum_{k=1}^{C-1} \delta_{j,k}$ , which implies  $\sum_{c=1}^C \beta_{j,c} = 0$ . Therefore, the common variance prior, and so the posterior, identify the centralized  $\beta_{j,1:C}$ . However, the apparent non-identifiability in likelihood function does harm the efficiency of Markov chain sampling. A naive Gibbs sampler may stay for a long time where the absolute values of  $\beta_{j,1:C}$  are very large, but the differences  $\delta_{j,1:(C-1)}$  can be achieved by centralized  $\beta_{j,1:C}$  with smaller values.

We can use only the  $\delta_{j,1:(C-1)}$  as Markov chain state even when we use symmetric prior, simply by transforming  $\beta_{0:p,1:C}$  to a new set of parameters, and marginalizing the non-identifiable parameters. The transformed parameters from  $\beta_{0:p,1:C}$  are:

$$\delta_{j,k} = \beta_{j,k+1} - \beta_{0:p,1}, \text{ for } k = 1, \dots, K \equiv C - 1, j = 0, \dots, p \quad (7)$$

$$\beta_{j,1} = \beta_{j,1}, \text{ for } j = 0, \dots, p. \quad (8)$$

We will exactly transfer the symmetric prior for  $\beta_{j,1:C}$  to  $\delta_{j,1:K}$ . Conditional on  $\sigma_j^2$ , the transformed parameters  $(\beta_{j,1}, \delta_{j,1:K})$  are distributed with a joint multivariate Gaussian distribution. The conditional distribution of  $y$  given  $\mathbf{x}$  with  $\delta_{0:p,1:K}$  is:

$$P(y_i = k + 1 | \mathbf{x}_{i,1:p}, \delta_{0:p,1:K}) = \frac{I(k = 0) + I(k > 0) \exp(\delta_{0k} + \mathbf{x}_{i,1:p} \delta_{1:p,k})}{1 + \sum_{k=1}^K \exp(\delta_{0k} + \mathbf{x}_{i,1:p} \delta_{1:p,k})}, \quad (9)$$

for  $k = 0, \dots, K$ , and  $i = 1, \dots, n$ , where  $I(\cdot)$  is an indicator function, equal to 1 if the condition in bracket is true, 0 otherwise. Since  $\beta_{j,1}$  is not used in (9), therefore does not appear in the likelihood function of new parameters, we can integrate it out from the transferred prior directly conditional on  $\sigma_j^2$ , obtaining the marginal prior for  $\boldsymbol{\delta}_{j,1:K}$  given  $\sigma_j^2$ :

$$\boldsymbol{\delta}_{j,1:K} | \sigma_j^2 \sim N_K(\mathbf{0}, (I_K + J_K)\sigma_j^2), \quad (10)$$

where  $I_K$  is a  $K \times K$  identity matrix and  $J_K$  is a  $K \times K$  matrix with all elements 1. For  $C = 2$ , (10) is just a univariate normal for  $\delta_{j,1}$  with variance  $2\sigma_j^2$ . In summary, **BLRHL** model parameterized by  $\boldsymbol{\delta}_{j,1:K}$  is now a hierarchy defined by (9), (10) [replacing (2) and (3)], and one of priors (4), (5) and (6) for  $\boldsymbol{\sigma}_{1:p}^2$ .

From (10), we see that what's different from the asymmetric prior is that the prior for  $\delta_{j,1:K}$  are *correlated* for maintaining symmetry. The correlations among  $\boldsymbol{\delta}_{j,1:K}$  cause no difficulty for Markov chain sampling, because we can easily evaluate the PDF of (10):

$$P(\boldsymbol{\delta}_{j,1:K} | \sigma_j^2) = (2\pi\sigma_j^2)^{-K/2} \exp\left(-\frac{V(\boldsymbol{\delta}_{j,1:K})}{2\sigma_j^2}\right) \times |I_K + J_K|^{-1/2}, \text{ where,} \quad (11)$$

$$V(\boldsymbol{\delta}_{j,1:K}) = \sum_{k=1}^K \delta_{jk}^2 - \left(\sum_{k=1}^K \delta_{jk}\right)^2 / C. \quad (12)$$

We see that  $V(\boldsymbol{\delta}_{j,1:K})$  is the sum of variations of  $(0, \delta_{j,1}, \dots, \delta_{j,K})$  from its mean, which is the same as the sum of variations of  $\boldsymbol{\beta}_{j,1:C}$ .  $V(\boldsymbol{\delta}_{j,1:K})$  therefore measures how useful the  $j$ th feature is for predicting the class label, if the features are normalized to have standard deviation 1. For selecting features, it is more straightforward to look at **SDB** (standard deviation of  $\boldsymbol{\beta}_{j,1:C}$ ):

$$\text{SDB}(\boldsymbol{\delta}_{j,1:K}) = \sqrt{V(\boldsymbol{\delta}_{j,1:K})/C}. \quad (13)$$

Note that, if  $C = 2$  ( $K = 1$ ),  $V(\delta_{j,1}) = \delta_{j,1}^2/2$ , and  $\text{SDB}(\delta_{j,1}) = |\delta_{j,1}/2|$ .

### 3.3 Gibbs Sampling Procedure

We use Gibbs sampling procedure to sample the full posterior distribution of BLRHL model. The full posterior distribution is written as:

$$P(\boldsymbol{\delta}_{0:p,1:K}, \boldsymbol{\sigma}_{1:p}^2 | \mathbf{D}) \propto L(\boldsymbol{\delta}_{0:p,1:K}) \times P(\boldsymbol{\delta}_{0:p,1:K} | \boldsymbol{\sigma}_{0:p}^2) \times P(\boldsymbol{\sigma}_{1:p}^2 | \alpha/2, \alpha w/2), \quad (14)$$

where,  $\mathbf{D}$  represents the data  $y_i, \mathbf{x}_{i,1:p}$  for  $i = 1, \dots, p$  and other fixed values in BLRHL models —  $\alpha, \sigma_0^2$ ;  $L$  is the likelihood function:  $L(\boldsymbol{\delta}_{0:p,1:K}) = \prod_{i=1}^n P(y_i | \mathbf{x}_{i,1:p}, \boldsymbol{\delta}_{0:p,1:K})$ ; the last two parts are the PDFs of priors specified by (10), and one of priors (4), (5), and (6). We sample the full posterior in (14) by sampling the conditional distribution of  $\boldsymbol{\sigma}_{1:p}^2$  and  $\boldsymbol{\delta}_{0:p,1:K}$  given each other alternately for a number of iterations. If IG prior (4) is used, the Gibbs sampling procedure is the alternate of the following two steps:

**Step 1:** Given  $\boldsymbol{\sigma}_{1:p}^2$  fixed, update  $\boldsymbol{\delta}_{0:p,1:K}$  jointly with a Hamiltonian Monte Carlo transformation that leaves invariant the following distribution:

$$P(\boldsymbol{\delta}_{0:p,1:K} | \boldsymbol{\sigma}_{0:p}^2, \mathbf{D}) \propto L(\boldsymbol{\delta}_{0:p,1:K}) \times P(\boldsymbol{\delta}_{0:p,1:K} | \boldsymbol{\sigma}_{0:p}^2). \quad (15)$$

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

$$\sigma_j^2 | \boldsymbol{\delta}_{1:p,1:K} \sim \text{IG} \left( \sigma_j^2 \left| \frac{\alpha + K}{2}, \frac{\alpha w + V(\boldsymbol{\delta}_{j,1:K})}{2} \right. \right), \quad (16)$$

The sampling for (16) in **Step 2** is straightforward. When Horseshoe and NEG priors are used, the sampling method for **Step 1** can be the same, but we have to use the posterior of  $\sigma_j^2$  given  $\boldsymbol{\delta}_{j,1:K}$  differently in **Step 2**. When we use GHS prior (5) for  $\sigma_j^2$ , the conditional posterior of  $\sigma_j^2$  given  $\boldsymbol{\delta}_{j,1:K}$  [in replace of equation (16)] is:

$$P_{ghs}(\sigma_j^2 | \boldsymbol{\delta}_{j,1:K}) \propto \frac{1}{(\sigma_j^2)^{K/2}} \exp \left( -\frac{V(\boldsymbol{\delta}_{j,1:K})}{2\sigma_j^2} \right) \times \left( \frac{1}{1 + \sigma_j^2/(\alpha w)} \right)^{\frac{\alpha+1}{2}} \frac{1}{(\sigma_j^2)^{1/2}}. \quad (17)$$

The induced conditional distribution for  $\xi_j = \log(\sigma_j^2)$  from the above distribution is log-concave, and can be sampled with ARS (Gilks and Wild, 1992). When we use NEG prior (6) for  $\sigma_j^2$ , the

conditional posterior of  $\sigma_j^2$  given  $\boldsymbol{\delta}_{j,1:K}$  [in replace of equation (16)] is

$$P_{neg}(\sigma_j^2 | \boldsymbol{\delta}_{j,1:K}) \propto \frac{1}{(\sigma_j^2)^{K/2}} \exp\left(-\frac{V(\boldsymbol{\delta}_{j,1:K})}{2\sigma_j^2}\right) \times \left(\frac{1}{1 + \sigma_j^2/\lambda}\right)^{\alpha/2+1}. \quad (18)$$

The induced posterior of the log transformation  $\xi_j = \log(\sigma_j^2)$  from the above distribution is log-concave, and can be sampled with ARS.

The key component in the above procedure is the use of Hamiltonian Monte Carlo (HMC) for updating high-dimensional  $\boldsymbol{\delta}_{0:p,1:K}$ . In Section A.2, we give a concise description of HMC. HMC can greatly suppress the random walk due to correlation (which is common in logistic regression posterior; see our real data examples) with a long leapfrog trajectory (Neal, 2010). The major problem of sampling from posteriors based on hyper-lasso priors is the existing of many local modes due to feature redundancy. Applying HMC in the above Gibbs sampling framework can travel across the modes fairly well for the following reason. When both  $\sigma_j^2$  for two correlated features are fairly large, the joint conditional distribution of their coefficients given  $\sigma_j^2$  in **Step 1** are highly correlated, probably close to their likelihood function as shown in Figure 2b. A fairly long HMC trajectory has much greater chance than ordinary MCMC methods to move from one end of the contour to the other end, by which the Markov chain can travel from one mode to another. For high-dimensional problems with very large  $p$ , such as thousands, this step is the bottleneck for the whole Markov chain sampling. This challenge can be relieved greatly by an important trick that we call “restricted Gibbs sampling” — only the coefficients with  $\sigma_j^2$  greater than a certain threshold are updated in **Step 1**. The details of this trick are given in Section A. A list of notations for the settings of BLRHL is given in Section B.

### 3.4 Feature Importance Indices from Markov chain Samples

With posterior samples of  $\boldsymbol{\delta}_{1:p,1:K}$ , we recommend using *means* over iterations to estimate the coefficients, denoted by  $\hat{\boldsymbol{\delta}}_{j,1:K}$ . We then compute  $\text{SDB}(\hat{\boldsymbol{\delta}}_{j,1:K})$  using formula (13) to obtain an importance index for feature  $j$ . The features can then be ranked by  $\text{SDB}(\hat{\boldsymbol{\delta}}_{j,1:K})$ . As we have discussed in Section 2, the Markov chain sample pool is a mixture of subpools from different modes, each corresponding to a succinct feature subset. The mean over Markov chain is a summary of the

importance of the feature, not an estimate of the true coefficient. This method omits some useful features that appear with low frequency in Markov chain samples. In the context of high-dimensional problems, there is often a large number of such correlated features, therefore discriminating them according to their predictive ability is desired. The ranking by *means* can omit many of such correlated features with weaker predictive ability, as well as those totally useless, and therefore pin down a very small subset of highly relevant features. Note that *Median* is not recommended since it may miss useful features that appear in Markov chain iterations with frequencies even *slightly* smaller than 0.5.

In light of the multi-modality in posterior, the means may not be the best choice for interpreting the MCMC simulation results, especially in the data sets with a large number of highly correlated features. However, in this paper we wish to focus on discussing the choice of heaviness and scale of hyper-lasso priors. We will discuss other alternatives in the Section 6.

## 4 Simulation Studies

### 4.1 Investigation of Choice of Scale with Small Data Sets

We generated a data 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:

$$P(y_i = c) = \frac{1}{2}, \quad \text{for } c = 1, 2, \quad \mathbf{x}'_{i,1:200} \mid y_i = c \sim N_{100}(\boldsymbol{\mu}'_{c,1:200}, A A' + I_{200}), \quad (19)$$

where,  $\boldsymbol{\mu}_{1,1:200} = (0, \dots, 0)$ ,  $\boldsymbol{\mu}_{2,1:200} = (2, 0, \dots, 0)$ ,  $A = (a_{ij})$  with all diagonal elements equal to 1, and  $a_{21} = 2$ . To make the 200 features commensurable, we standardize them to have means 0 and sds 1. In this model, only the first feature is differential across two classes, and the 2nd is non-differential but correlated with the 1st. Therefore, only the first two features are useful for predicting response  $y$ , with **true coefficients**  $\boldsymbol{\delta}_{0:2,1} = (0, 2.60, -1.22)$ , and all other coefficients are 0. The relationship between  $y$  and  $\mathbf{x}$  is simple, but the signals are placed among the other 198 useless features. Figure 3 shows the scatterplots of the 2nd and 3rd features to the 1st with shapes



representing two classes.

Figure 3: Scatterplots of the 2nd and 3rd features against the 1st generated from (19).

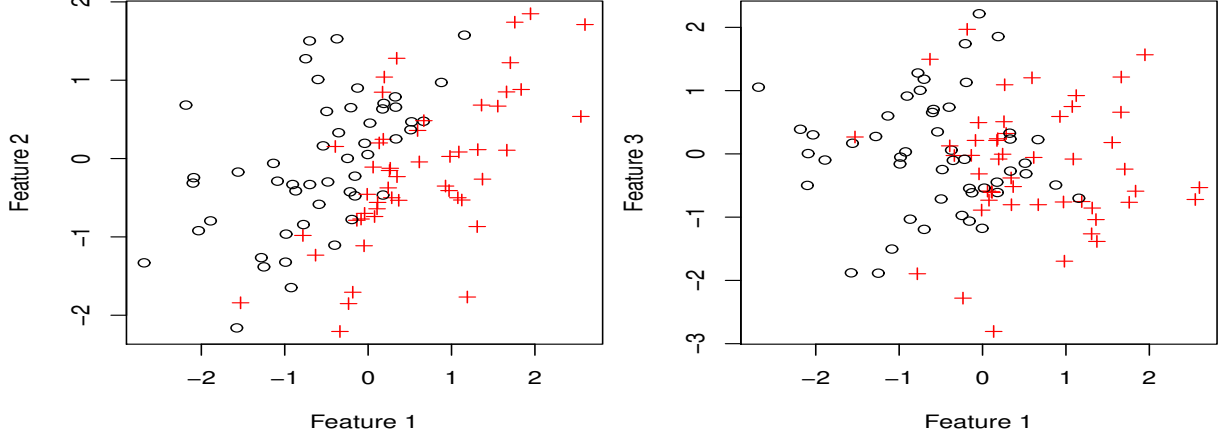
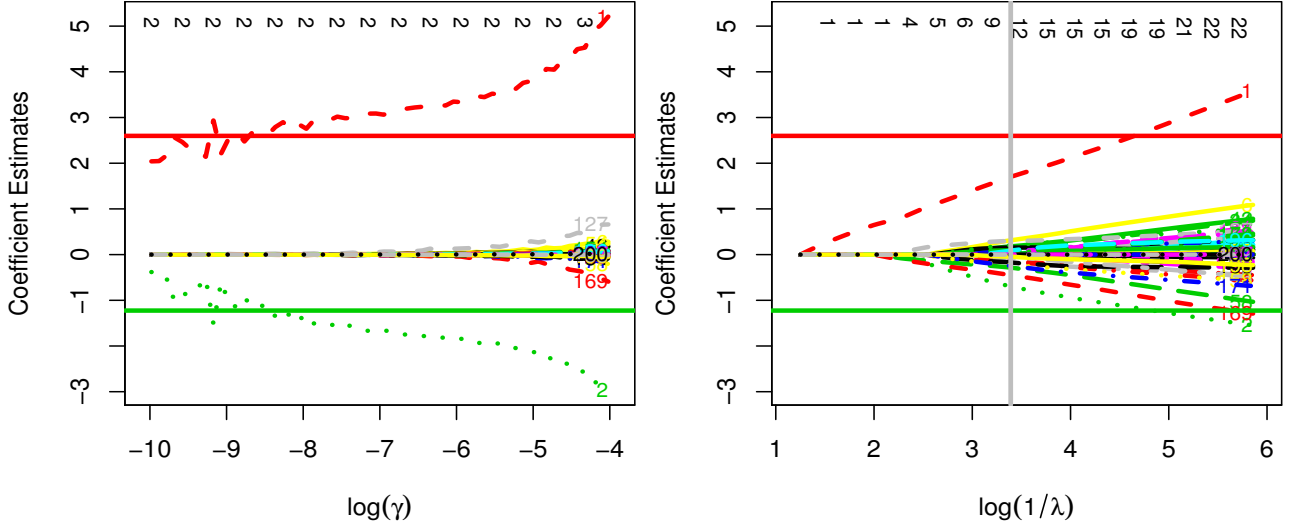


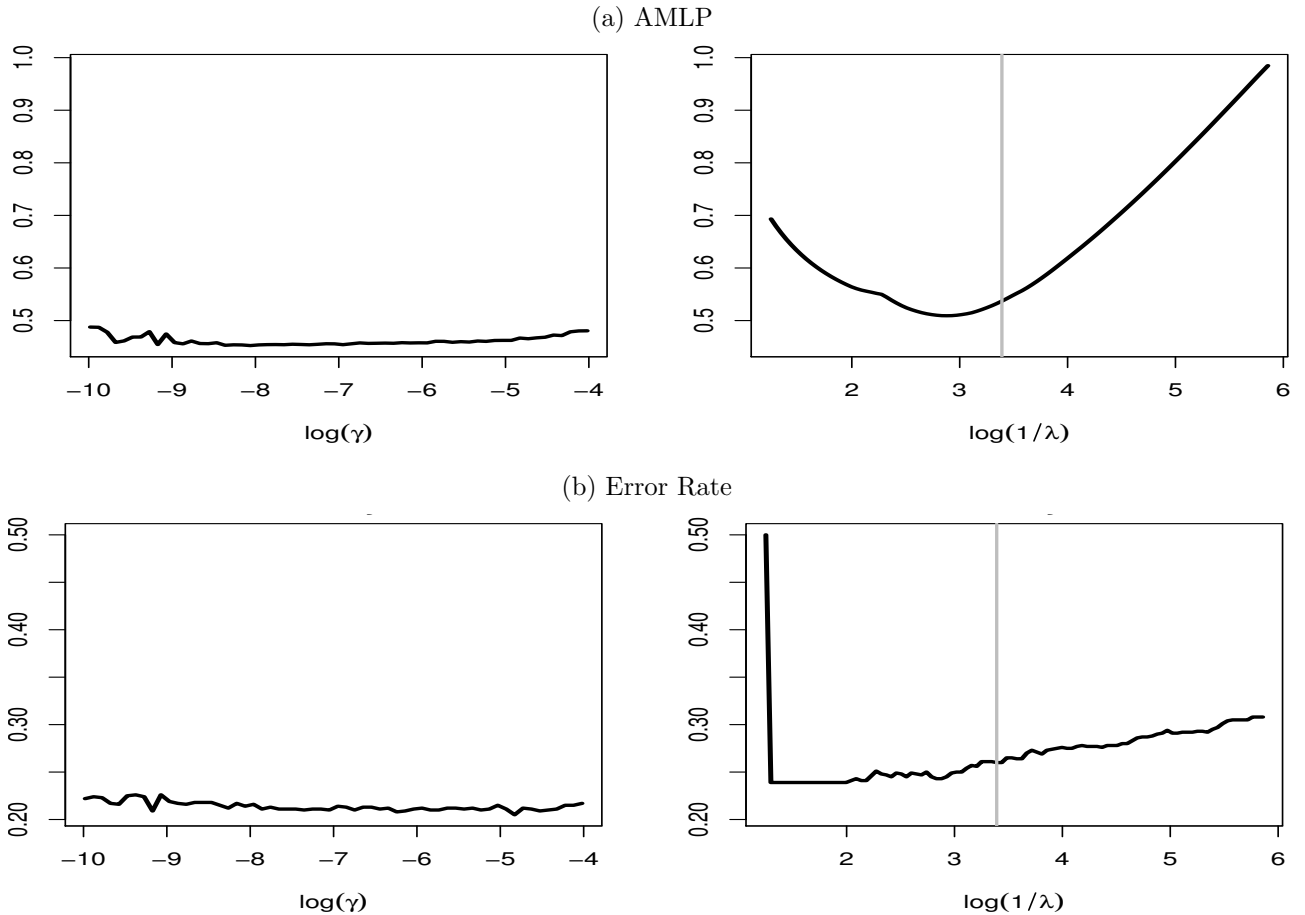
Figure 4: Solution paths of BLRHL (left) and LASSO (right). The numbers on the top show the number of coefficients with absolute values not smaller than 0.1 times the maximum value in all coefficients. The vertical line in LASSO path shows the cross-validation choice of  $\log(1/\lambda)$ . The numbers beside paths on the right are feature indice.



We ran BLRHL using **t prior** (ie, IG prior for  $\sigma_j^2$ ) with prior settings  $\alpha = 1$  and MCMC settings  $n_1 = 50K, \ell_1 = 5, n_2 = 100K, \ell_2 = 50, \epsilon = 0.3, \zeta = 0$ , and 100 different  $\log(w)$  spaced evenly from  $-24$  to  $-8$ . *The meanings of these setting parameters are listed in Section B.* The Markov chains are unnecessarily long as the dimension is very small, but we used it since the computation is fast, taking about 30 mins for each chain. For each choice of scale, we estimated the coefficients using *means* in Markov chain samples. These results allow us to draw the solution paths (Figure 4) of all the coefficients against  $\log(\gamma)$ , and compare the path given by LASSO (using R package

glmnet version 1.7). We can see that BLRHL gives much more distinctive estimates of the two non-zero coefficients from those of the other 198 hay than LASSO. LASSO cannot clearly make the second feature stand out from many other “hay”. From comparing these paths, we also see that the estimates by BLRHL are very stable for the choices of  $\gamma$  in a very wide range. There is upward bias in mean estimates when  $\gamma$  is large, because the marginal posterior distributions of two coefficients for two correlated features are skewed to large absolute values (which is explained on page 26 with Figure 11). This bias, however, doesn’t affect the predictive performance and feature selection.

Figure 5: Predictive performance of BLRHL (left) and LASSO (right) for different scale choice. The x-axis is the log scale, and the y-axis is value of AMLP (top) or error rate (bottom)



We also compare the predictive performance of BLRHL and LASSO in terms of AMLP — *the average minus log predictive probabilities at the true labels*, and error rates. The AMLP and error rate paths are shown in Figure 5. We see that BLRHL predicts better than LASSO. Most importantly, the predictive performance of BLRHL is very stable for the choice of  $\gamma$ . By contrast,

LASSO is very sensitive to the choice of  $\gamma$ : if a small scale is used, LASSO omits more useless features, but also over-shrinks the really large coefficients, especially weaker  $x_2$ , therefore gives poor prediction; conversely, if a large scale is used, LASSO takes many “hay” into the model, which also downgrades the predictive performance. We see that the cross-validated choice based on the small training data set did not pick up the optimal  $\log(1/\lambda)$ , around 3; and even when this optimal choice is made, LASSO nearly omits  $x_2$ . BLRHL does not have this difficulty, and can identify the 2 “needles” from the 198 “hay”.

## 4.2 Investigation of Choice of Heaviness with Large Data Sets

We generated 50 data sets ( $n = 2100$  cases, 100 of which was used for training, the other 2000 was used for looking at predictive performance) as follows. The number of classes  $C$  is set to 3, and class labels are equally likely drawn from 1, 2, and 3. The first two features were generated in a similar way as the  $x_1$  and  $x_2$  of the data set used in Section 4.1, with an addition of class 3 having 0 for means. We add another group of features ( $x_3 - x_{10}$ ) that are highly correlated within group but independent of  $x_1$  and  $x_2$ , and have mean equal to 2 in class 3. More specifically, values of these 10 features for each case were generated as follows:

$$\begin{aligned} x_1|y=c &= \mu_{c,1} + z_1 + 0.5\epsilon_1, \\ x_2|y=c &= \mu_{c,2} + 2z_1 + z_2 + 0.5\epsilon_2, \\ x_j|y=c &= \mu_{c,j} + z_3 + 0.5\epsilon_j, \text{ for } j = 3, \dots, 10, \end{aligned} \quad \text{where, } (\mu_{c,j})_{3 \times 10} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 2 & 0 & 0 & \dots & 0 \\ 0 & 0 & 2 & \dots & 2 \end{pmatrix},$$

and,  $z_j$  and  $\epsilon_j$  are independently generated from  $N(0,1)$ . In addition to these 10 features, we also attached 1990 features simply drawn from  $N(0,1)$ , which we will call absolute “hay”. In this model,  $x_1$  is differential, with different mean in class 2 from classes 1 and 3.  $x_2$  is non-differential, but correlated with  $x_1$ , therefore is useful, as shown by Figure 3.  $x_3 - x_{10}$  are all differential, with different means in class 3 from classes 1 and 2. However,  $x_3 - x_{10}$ , which have the same class means and are related to a common factor  $z_3$ , are highly correlated and redundant for predicting  $y$ ; we will call this group of features as “correlated features”.

Here we use  $p = 2000$  for experimental time consideration. Our methods can be applied to

problems of higher dimensions, such as more than 6000 features in the case study presented in Section 5. However, generally, we think that in practice, we should apply BLRHL to a much smaller subsets of features pre-selected with univariate screening to discard those obviously useless features (such as those with very small variation at all), rather than on the original data set with millions of features.

We ran BLRHL using  $t$  priors with 4 choices of  $\alpha$ : 0.2, 0.5, 1, 4, 10. The setting for  $\log(w)$  varies slightly for different  $\alpha$ . When  $\alpha = 1$ , we chose two values of  $\log(w)$ : -20 and -10. When  $\alpha = 4$  and 10, we chose to treat  $\log(w)$  as a hyperparameter assigned with a normal prior with variance 100 (The reason is that, for large  $\alpha$ , the results for feature selection and prediction are sensitive to the choices of scale and we therefore show the results with  $\log(w)$  chosen automatically during MCMC simulation). The values of  $\log(w)$  for  $\alpha = 0.2/0.5$  are -40/-20 respectively. For setting MCMC, when  $\alpha = 1$ , we chose  $n_1 = 50K, \ell_1 = 10, n_2 = 500K, \ell_2 = 50, \epsilon = 0.3, \zeta = 0.05$ . When  $\alpha$  equals to 0.2/0.5/4/10 (other than 1), we set a larger  $n_2 = 1M$  for longer chain and the same other settings as  $\alpha = 1$ . We ran BLRHL using GHS and NEG priors with settings  $\alpha = 1$  and  $\log(w) = -10$ , and the same other settings for running HPBLR using  $t$  prior with  $\alpha = 1$ . We ran LASSO with  $\lambda$  chosen by cross-validated AMLP.

Figures 6 and 7 show the SDBs by different methods of all 2000 features for a data set, which is similar to most other data sets. From Figure 6, we see that BLRHL using  $t$ /GHS/NEG priors with  $\alpha = 1$  and  $t$  prior with  $\alpha = 0.5$  does the feature selection very well by looking at relative SDBs with threshold 0.1: 1) they can distinctively separate the absolute “hay” from other useful features, with SDBs mostly only equal to  $10^{-2} - 10^{-4}$  times the maximum SDB; 2) they do not miss the 2nd feature which is useful but has weaker relevance; 3) they rank highly one feature ( $x_5$ ) from the 8 correlated features, recognize another feature  $x_3$  as useful too, and discriminate others; we believe  $x_5$  is the most useful based on the information from the training data because it is consistently identified by all the runs with  $\alpha = 1$  and 0.5; 4) the results of BLRHL using  $t$  prior with  $\alpha = 1$  are stable for the two very different choices of  $\log(w)$ : -10 and -20. We believe that this applies to GHS and NEG priors too; 5) the results of BLRHL using  $t$  prior and GHS and NEG priors with the

Figure 6: SDBs given by BLRHL using  $t$  priors and LASSO on a synthetic data set with  $p = 2000$  features. The shapes of points  $*$ ,  $+$ ,  $\times$ ,  $\bullet$  respectively stand for four groups of features:  $x_1$ ,  $x_2$ ,  $x_3 - x_{10}$  and  $x_{11} - x_{2000}$ . The red numbers below points show the indice of top 10 features. Both  $x$  and  $y$  axes are in logarithm scale. The horizontal lines indicate the values equal to 0.1 and 0.01 times the maximum SDB.

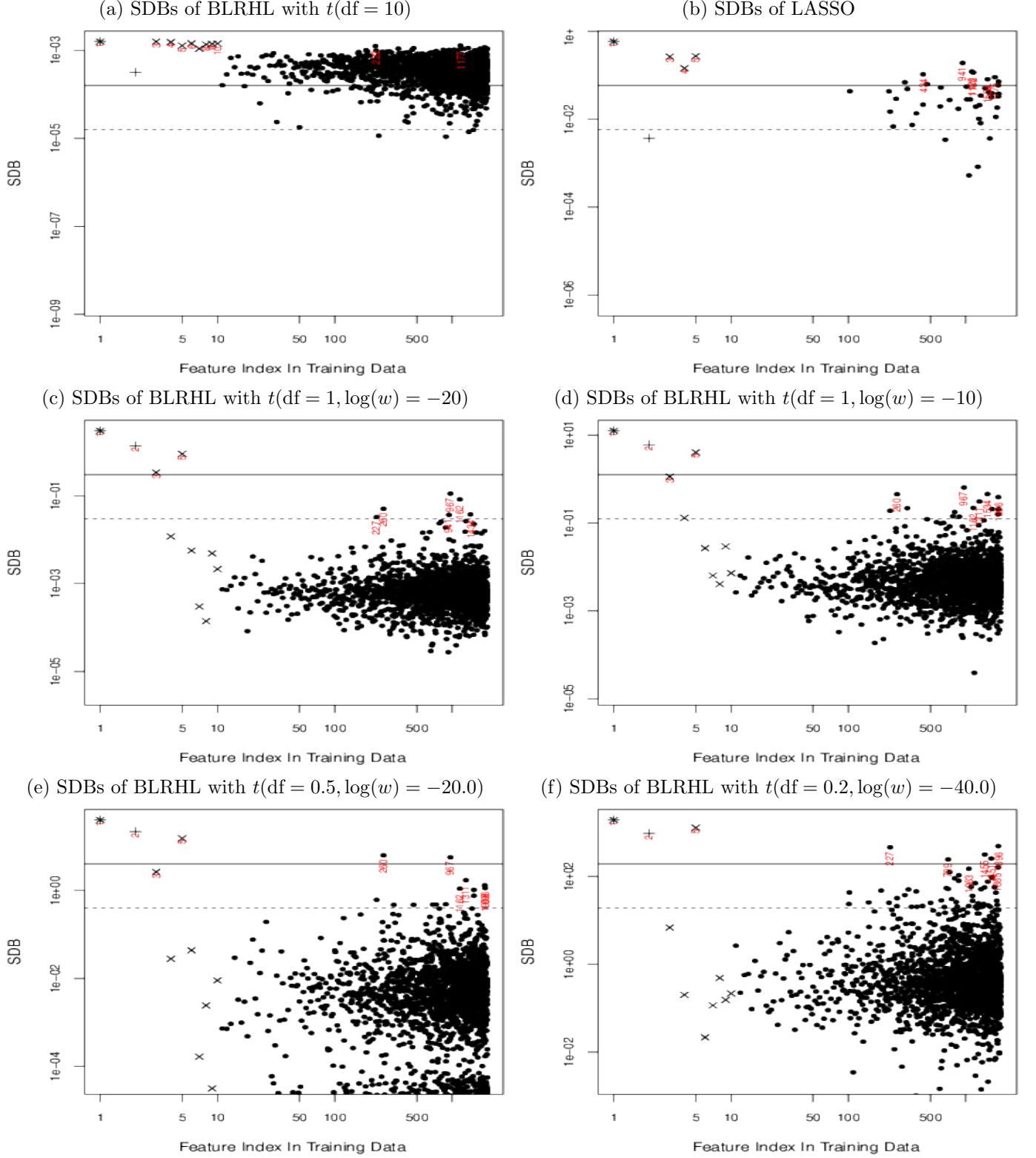
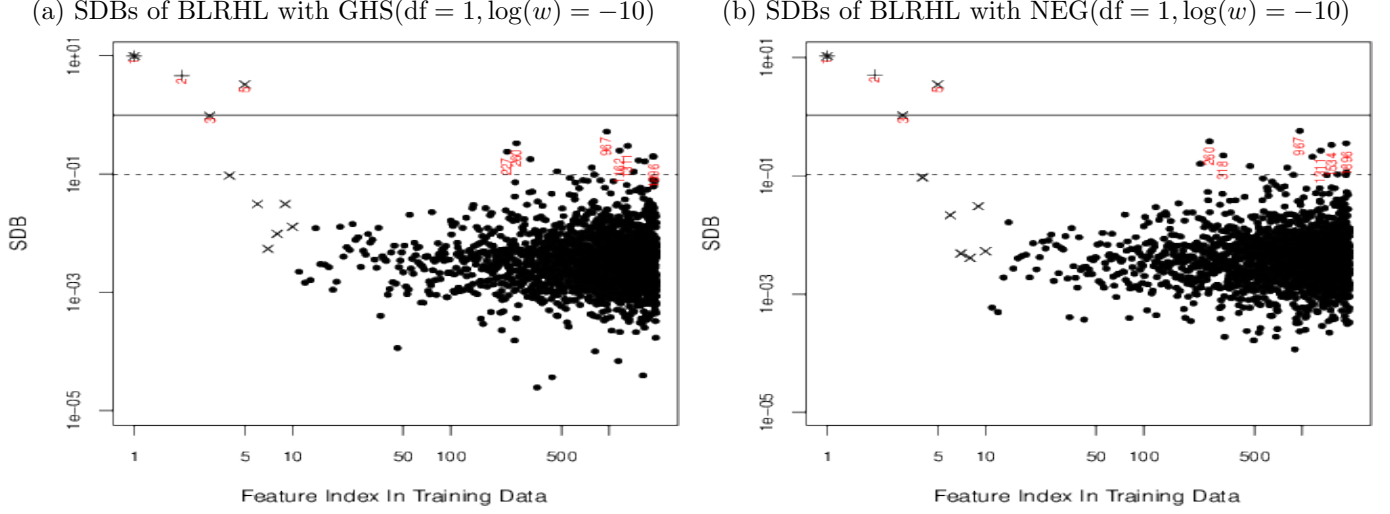


Figure 7: SDBs given by BLRHL using GHS and NEG priors on the same data set used to make plots in Figure 6.



same  $\alpha$  and  $\log(w)$  are almost the same. By contrast, 1) LASSO and BLRHL with large  $\alpha$  (such as 10) cannot separate the absolute “hay” distinctively from the few “needles”, but LASSO is better than BLRHL with large  $\alpha$ ; 2) they both miss  $x_2$  for this data set (and very often for other data sets), which we think is because they include many “hay” to overfit the data, therefore make  $x_2$  harder to identify; 3) they tend to include many of the correlated features into their unique mode, without clear discrimination for importance; when the group of correlated features is large, they could be entirely missed because the coefficients in such a mode for the correlated features have small absolute values as shown by Figure 2b. BLRHL with very small  $\alpha = 0.2$  (very heavy tails) can do feature selection well but their overly flat tails allow the “needles” to go to very large values (such as *thousands*, see Figure 6e and 6f), resulting in very poor prediction for a few cases, and infinite AMLP. Therefore priors with overly heavy tails are not good for logistic regression models. This is because the likelihood functions of logistic regression don’t punish very large coefficients once the correct subset of features are selected, and very heavy tailed priors don’t punish either.

Table 2 shows a summary of numbers of selected features in 4 different groups by retaining features with SDBs not smaller than 0.1 times the maximum SDB (ie, by thresholding relative SDBs with 0.1). Figure 8 shows the boxplots of the 50 AMLPs of the prediction on 2000 test cases given different methods. We see that BLRHL (using  $t$ /GHS/NEG priors) with  $\alpha = 1$  gives

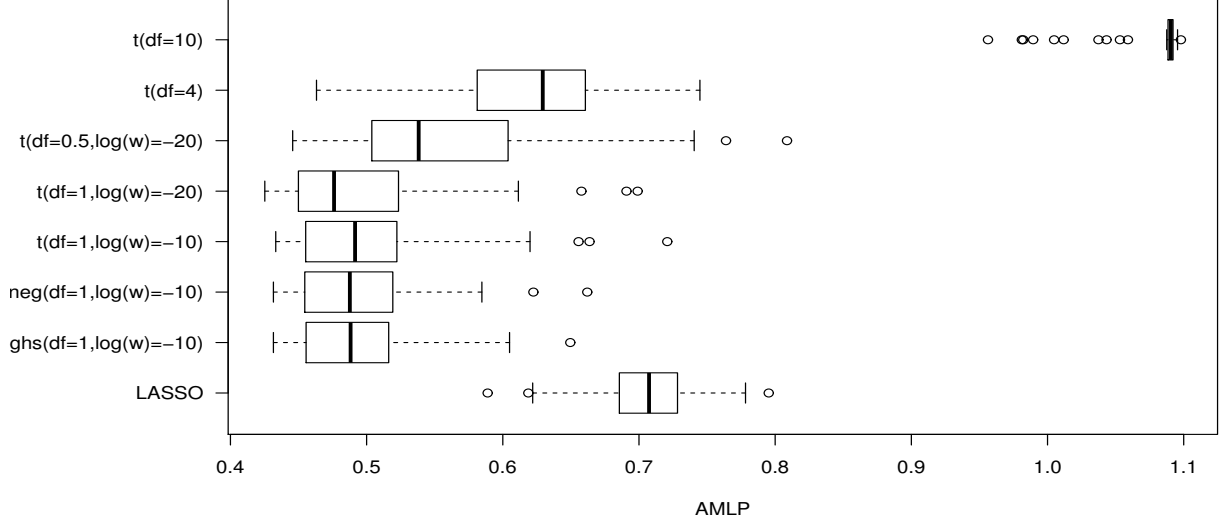
Table 2: Means of numbers of retained features by thresholding relative SDBs with 0.1 in different groups in 50 data sets. Numbers in brackets show the standard deviations of the 50 numbers.

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

substantially better predictions and feature selections for most of the data sets than other choices of  $\alpha$  and LASSO. LASSO and BLRHL with very large and very small  $\alpha$  do not predict well. Note that the AMLPs of all runs with  $\alpha = 0.2$  are *infinity*, so not shown in Figure 8. BLRHL with moderate  $\alpha = 4$  does well in omitting the absolute “hay” from the “needles” with a less distinctive boundary because of the lighter tails. However, we see that it has large chance of missing the weaker feature  $x_2$ , therefore gives poorer predictions most of times, which is also due to over-shrinkage of the signals. Therefore, the choice of  $\alpha$  is critical for BLRHL to work well for high-dimensional classification, and  $\alpha = 1$  is recommended based on our investigations.

From Table 2 and Figure 8, we see that the feature selection and prediction performance measures using GHS and NEG priors are slightly better on these 50 data sets than using  $t$  priors with the same value  $\alpha$  and  $\log(w)$ . This improvement may be attributed to the nonzero values of PDFs of GHS and NEG at 0. But generally they are very similar, with differences without statistical significance. This indicates that  $t$  prior with small degree freedom and small scale performs almost as well as these two more complicated priors in logistic regression problems. On the other hand, the additional computation for using GHS and NEG compared to  $t$  is not negligible in high-dimensional problems based on our implementation. A Markov chain using  $t$  prior with  $\alpha = 1$  as described previously took only about 3 hours, whereas a Markov chain using GHS or NEG priors took about 8 hours. The time difference is merely due to the use of the ARS sampling for posteriors [equations

Figure 8: Boxplots of AMLPs on 2000 test cases by BLRHL with various priors and LASSO. “df” in the plot is  $\alpha$  of priors for BLRHL.



(17) and (18)] of  $\sigma_j^2$  given coefficients rather than standard sampling for IG posterior [equation (16)]. The time difference will be larger as  $p$  is larger. However, this time difference is possibly eliminated by using better methods than ARS for sampling (17) and (18).

## 5 Tests with Prostate Microarray Data

We applied BLRHL to a real microarray gene expression data that is related to prostate cancer, which has expression profiles of 6033 genes from 50 normal and 52 cancerous tissues. This data set was originally reported by Singh et al. (2002). We analyzed a data set downloaded from the website <http://stat.ethz.ch/~dettling/bagboost.html> for Dettling (2004), which contains more descriptions about this data set. For better looking at our results, we re-ordered the features by F-statistic on the whole data set, therefore the indice of the genes discussed below are also the ranks of features according to F-statistic. We ran BLRHL using  $t$ /GHS/NEG priors and LASSO with  $\lambda$  chosen with cross-validation, in leave-one-out cross-validation (LOOCV) fashion for comparing predictive performance. Before fitting with BLRHL, we always standardized the features with only training data (LASSO does such standardization too). We ran BLRHL with settings  $\alpha = 1, \log(w) = -10, n_1 = 100K, \ell_1 = 10, n_2 = 1M, \ell_2 = 50, \epsilon = 0.3, \zeta = 0.05$  with all 6033 genes. Each Markov chain took about 10 hours if  $t$  prior was used, and about 33 hours if GHS/NEG priors were used.



Figure 9: Gene selection results on Prostate data by different methods. The red numbers under points show the original indice of top 10 ranked genes. The horizontal lines indicate the values equal to 0.1 and 0.01 times the maximum SDB or P-value. (a) shows results using  $t$  priors, and (b) shows results from rerunning with top 50 genes selected from (a).

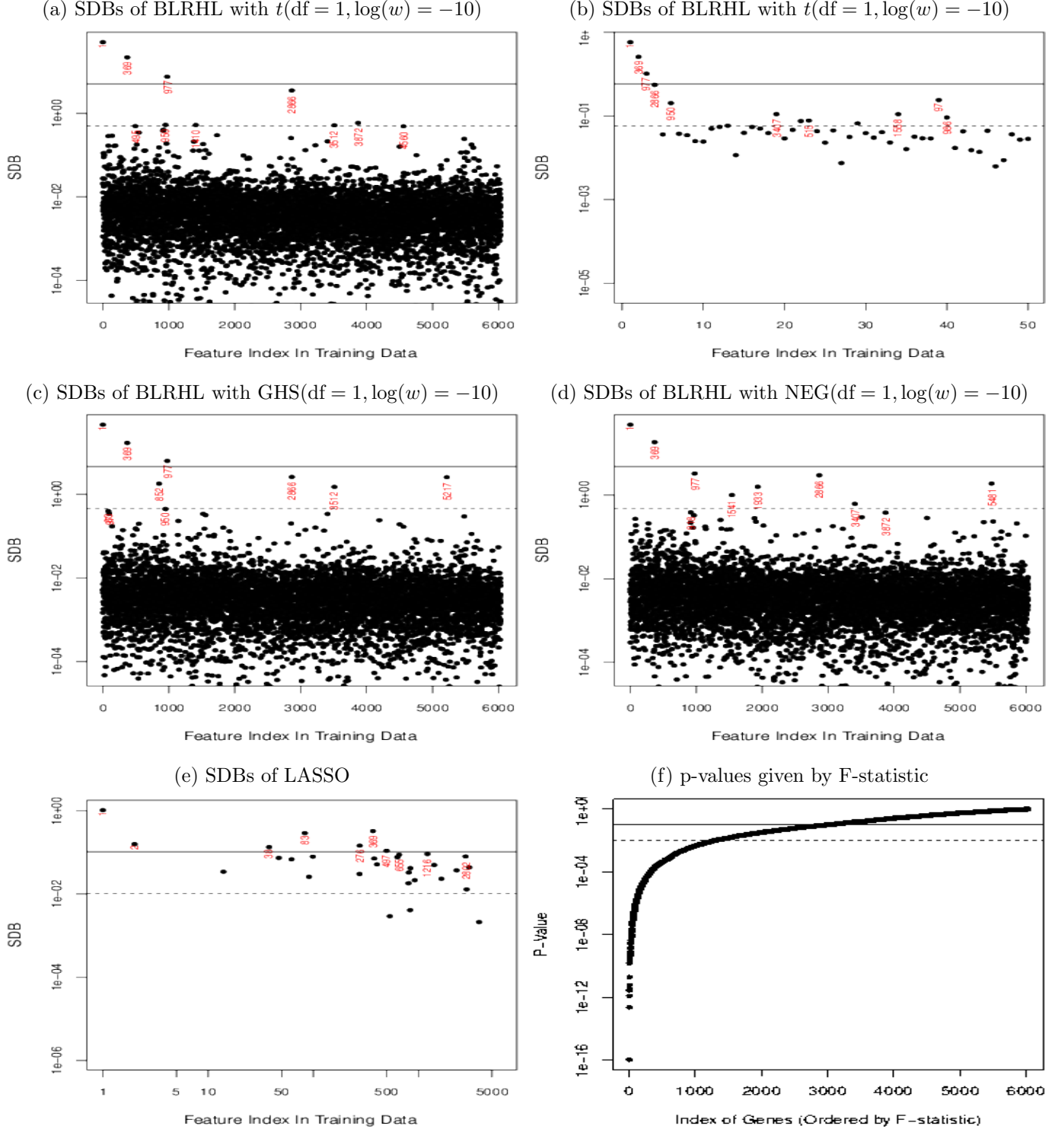


Figure 9 shows the SDBs of BLRHL and LASSO for the fold with the 2nd case left out, and P-values given by F-statistic on the whole data set. Figure 9b shows the results of rerunning BLRHL using  $t$  prior on only top 50 genes selected using the run with all 6033 genes. From these plots, we see that BLRHL using  $t$ /GHS/NEG priors perform very similarly and stably, and all distinctively select fewer than 10 genes by thresholding relative SDBs with 0.01. Comparing to F-statistic, we see that only the most differentiated gene 1 is ranked as top 1 by BLRHL, and thousands of genes with very small P-values are omitted by HPBLR methods; at the same time, other top ranked genes by BLRHL indeed have low ranks (such as 369, 977, 2866) by F-statistic. Comparing to LASSO, we see that 1) many SDBs (other than those 0) by LASSO are narrow around the value equal to 0.1 times the maximum SDB; 2) the maximum SDB value, around 1, by LASSO is much smaller than the maximum SDB value, around 10, by BLRHL, because LASSO over-shrinks the coefficients; 3) LASSO omits gene 977, which is ranked the 3rd by BLRHL, given 0 coefficient by LASSO, and is probably useful after a check with cross-validation, as shown at the end of this section.

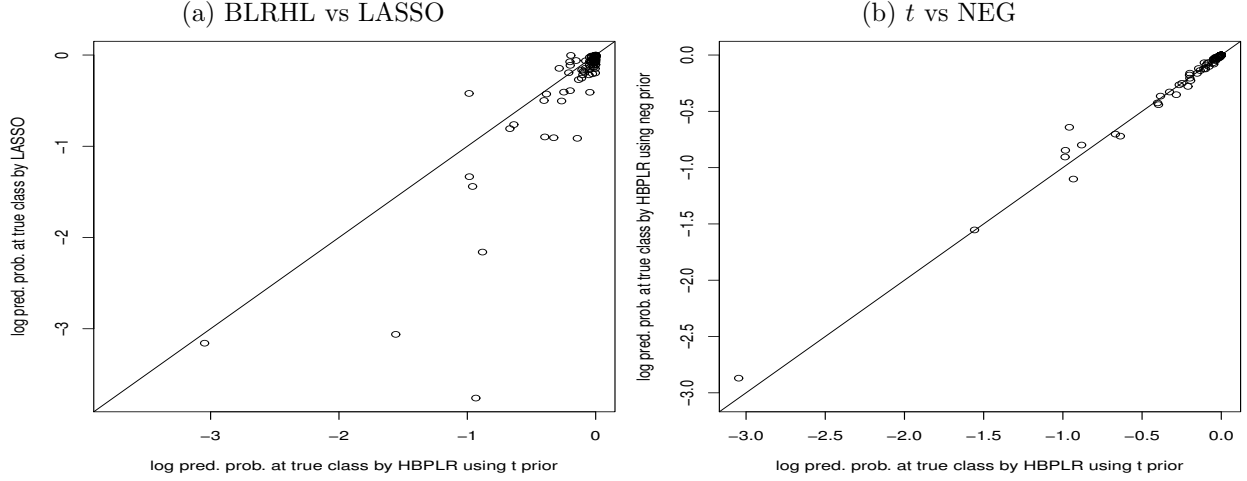
LOOCV predictive performances measured by AMLP and error rate are shown in Table 3, which also includes the results of many other methods reported by Dettling (2004). Figure 10 shows the scatterplots of log predictive probabilities at true class labels by BLRHL using  $t$  and NEG priors and LASSO. BLRHL is substantially better than many other methods, especially LASSO. The difference in AMLPs between BLRHL (using  $t$  prior) and LASSO is statistically significant, with p-value 0.00046 by paired one-sided  $t$  test. The performances of BLRHL using  $t$ /GHS/NEG priors are very similar without statistical significant difference, as shown by Figure 10b. Finally, we point out that the predictive performances of BLRHL can be improved further if we rerun them with fewer top genes selected with the runs with all genes due to improved MCMC sampling.

Table 3: Comparisons of LOOCV predictive performances of BLRHL and others. BLRt, BLRghs and BLRneg are BLRHL using  $t$ , GHS and NEG priors respectively.

Methods	BLRt	BLRghs	BLRneg	LASSO	Bagboost	PAM	DLDA	SVM	RanFor	kNN
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

We have looked at the expression values of the top genes (such as 1, 369, 977, and 2866) and

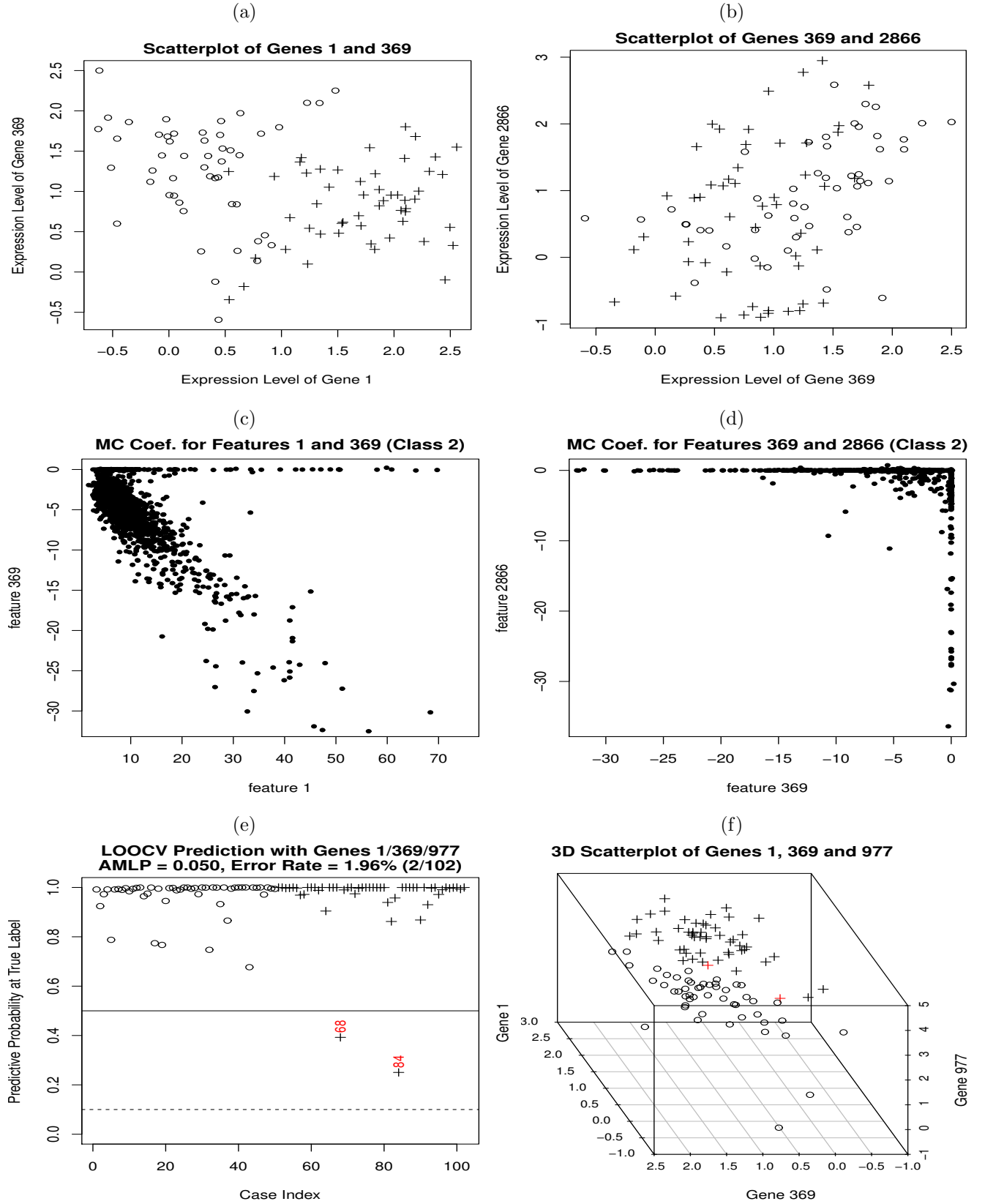
Figure 10: Comparisons of log predictive probabilities at true class labels.



the MCMC samples of their coefficients. Figure 11 shows some results with the Markov chain by running BLRHL using  $t$  priors on the data set with the 2nd case left out and only top 50 genes that were selected from a previous run using all genes (the SDB values of these two runs are shown in Figure 9a and 9b respectively). First, from these plots and others not shown, we see that genes 369, 977, and 2866 are weakly differentiated across two classes, but they are useful because they are correlated with the most differentiated gene 1. The joint posterior of coefficients for genes 1 and 369 is highly correlated, and skewed to large absolute values, as shown by Figure 11c. The reason is that in logistic regression, only the ratio between coefficients can be fairly determined by the classification boundary, but the “slope” of hyperplane cannot, therefore both log likelihood function and hyper-lasso priors allow large proportional absolute values. These observations indicate that it is necessary to use HMC to overcome the random walk of ordinary MCMC methods. Second, Figure 11b shows that genes 369 and 2866 are also correlated but both are weakly differentiated, therefore they are redundant in explaining the response. Markov chain simulation (results shown by Figure 11d) separates genes 369 and 2866 into different modes, and uses only 1 in each (or none). However, the absolute coefficient values of each gene are clearly large in each mode, indicating that both are useful but redundant. Generally, we see that the posterior of BLRHL using hyper-lasso priors is very multimodal.

Finally, we compare the predictive ability of the top 3 genes found by BLRHL with other small

Figure 11: MCMC coefficient samples, expression scatterplots, and predictive performance of some top genes selected by BLRHL. The two red numbers or + in (e) and (f) label the two cases misclassified in LOOCV.



gene subset, by looking at LOOCV prediction results given by running BLRHL (using  $t$  priors with  $\alpha = 1$ ) on the data set containing only genes of a fixed subset. The results are shown in Table 4. The subset of genes 1, 369 and 977 is substantially better than other subsets in separating the two classes. The 3D scatterplots (Figure 11f) of the top 3 genes (1, 369, and 977) suggests that this gene subset can separate the two classes very well. The LOOCV predictive probabilities at the true class labels given by BLRHL using only this feature subset are very good, shown by Figure 11e. We think that the subset of genes 1, 369, and 977 is worthy of further biological investigations. We also see that gene 977 is useful, providing additional information than genes 1 and 369 for separating the two classes. Note that, this gene is missed by LASSO.

Table 4: LOOCV predictive performances of various gene subsets.

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

## 6 Conclusions and Discussions

In this article, we introduce a class of fully Bayesian logistic regression methods with continuous hyper-lasso priors, called BLRHL, for high-dimensional feature selection. We have tested and investigated BLRHL with simulated data sets and a real microarray data set. We have demonstrated that fully Bayesian methods with hyper-lasso priors are feasible for high-dimensional problems, and can achieve the good hyper-lasso benefits — shrink a large number of small signals to a value very close to 0 but retain a few clearly large signals. The prediction and feature selection performances of BLRHL using hyper-lasso priors with tails similar to Cauchy ( $t$  with 1 degree freedom) are clearly superior than LASSO, BLRHL using priors with heavier or lighter tails, and many other methods appearing in the literature. We have also used simulated data to investigate the problems of choosing heaviness and scale in BLRHL, which have demonstrated three important and useful findings as stated in the introduction section.

In light of the fact that the posterior distributions based on hyper-lasso priors are highly multi-modal, summarizing the feature importance by averaging the coefficients over all modes may not

the best choice. Particularly, in the problems with a large group of highly correlated features, many features in the group will be selected by means of coefficients. A more sophisticated method for interpreting the fitting results is to use a clustering algorithm to divide the whole Markov chain samples into subpools, look at the subpools separately, and then deliver a list of succinct feature subsets. Once we can split Markov chain iterations as this, it will then be better to use *median* to obtain an importance index as it can better shrink the coefficients of totally useless features towards 0 and correct for the skewness of the posterior. This will demand further development of methods for interpreting the MCMC simulation results. This is my priority in future work.

The main drawback of BLRHL is that they are still much slower than many others, such as penalized likelihood methods. Therefore, there is still much room for improving the computational efficiency. The difficulties lie in the high-dimensionality and the existing of multiple posterior modes. A possible solution may be to rotate the original feature space with PCA method, and then apply Gibbs sampling to the posterior distribution of the new coefficients for the transformed feature space with lower dimension. An MCMC simulation for the transformed coefficients may travel across multiple modes with fewer iterations. The crucial step of this solution is to devise a method for sampling the original coefficients conditional on the values of transformed coefficients, a technique similar to what described by [Li and Neal \(2008\)](#).

## Acknowledgements

The research of Longhai Li is supported by Natural Sciences and Engineering Research Council of Canada (NSERC), and Canadian Foundation of Innovations (CFI). Longhai Li acknowledges inspiring conversations with Radford Neal.

## References

- Alhamzawi, R., Yu, K., and Benoit, D. F. (2012), “Bayesian adaptive Lasso quantile regression,” *Statistical Modelling*, 12, 279–297.
- Armagan, A., Dunson, D., and Lee, J. (2010), “Bayesian generalized double Pareto shrinkage,” *Biometrika*.
- Bae, K. and Mallick, B. K. (2004), “Gene selection using a two-level hierarchical Bayesian model.” *Bioinformatics*, 20, 3423–3430.

- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2012), “Bayesian shrinkage,” *arXiv preprint arXiv:1212.6088*.
- Bottolo, L. and Richardson, S. (2010), “Evolutionary stochastic search for Bayesian model exploration,” *Bayesian Analysis*, 5, 583–618.
- Brown, P. J., Vannucci, M., and Fearn, T. (1998), “Multivariate Bayesian variable selection and prediction,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 627–641.
- Cai, X., Huang, A., and Xu, S. (2011), “Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping,” *BMC Bioinformatics*, 12, 211, PMID: 21615941.
- Caron, F. and Doucet, A. (2008), “Sparse Bayesian nonparametric regression,” in *Proceedings of the 25th international conference on Machine learning*, ACM, p. 8895.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009), “Handling sparsity via the horseshoe,” *Journal of Machine Learning Research*, 5.
- (2010), “The horseshoe estimator for sparse signals,” *Biometrika*, 97, 465–465.
- Clarke, R., Renshaw, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., and Wang, Y. (2008), “The properties of high-dimensional data spaces: implications for exploring gene and protein expression data,” *Nat. Rev. Cancer*, 8, 37–49.
- Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian methods for nonlinear classification and regression*, Wiley Series in Probability and Statistics.
- Dettling, M. (2004), “BagBoosting for tumor classification with gene expression data,” *Bioinformatics*, 20, 3583–3593.
- Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), “Comparison of discrimination methods for the classification of tumors using gene expression data,” *JASA*, 97, 77–87.
- Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *JASA*, 96, 1348–1360.
- Fan, L. I. and Zhang, N. R. (2010), “Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics,” *JASA Theory and Methods*, 105, 1202–1214.
- Gelman, A. (2006), “Prior distributions for variance parameters in hierarchical models,” *Bayesian analysis*, 1, 515–533.

- Gilks, W. R. and Wild, P. (1992), “Adaptive rejection sampling for Gibbs sampling,” *Applied Statistics*, 41, 337–348.
- Griffin, J. E. and Brown, P. J. (2010), “Inference with Normal-Gamma prior distributions in regression problems,” *Bayesian Analysis*, 5, 171–188.
- (2011), “Bayesian Hyper-Lassos with Non-Convex Penalization,” *Australian & New Zealand Journal of Statistics*, 53, 423–442.
- Guan, Y. and Stephens, M. (2011), “Bayesian variable selection regression for genome-wide association studies and other large-scale problems,” *The Annals of Applied Statistics*, 5, 1780–1815.
- Hans, C. (2009), “Bayesian lasso regression,” *Biometrika*, 96, 835–845.
- Hans, C., Dobra, A., and West, M. (2007), “Shotgun stochastic search for “large p” regression,” *J. Am. Stat. Assoc.*, 102, 507–516.
- Ishwaran, H. and Rao, J. S. (2005a), “Spike and slab gene selection for multigroup microarray data,” *JASA*, 100, 764–780.
- (2005b), “Spike and slab variable selection: frequentist and Bayesian strategies,” *The Annals of Statistics*, 33, 730–773.
- Kotz, S. and Nadarajah, S. (2004), *Multivariate  $t$  distributions and their applications*, Cambridge Univ Pr.
- Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), “Penalized regression, standard errors, and bayesian lassos,” *Bayesian Analysis*, 5, 369–412.
- Lamniso, D., Griffin, J. E., and Steel, M. F. J. (2011), “Cross-validation prior choice in Bayesian probit regression with many covariates,” *Statistics and Computing*, forthcoming.
- Leng, C., Tran, M., and Nott, D. (2014), “Bayesian adaptive Lasso,” *Annals of the Institute of Statistical Mathematics*, 66, 221–244.
- Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011), “The Bayesian lasso for genome-wide association studies,” *Bioinformatics*, 27, 516–523.
- Li, L. (2012), “Bias-Corrected Hierarchical Bayesian Classification With a Selected Subset of High-Dimensional Features,” *Journal of the American Statistical Association*, 107, 120–134.
- Li, L. and Neal, R. M. (2008), “Compressing parameters in Bayesian high-order models with application to logistic sequence models,” *Bayesian Analysis*, 3, 793–821.



- Lindley, D. V. (1957), “A statistical paradox,” *Biometrika*, 44, 187–187.
- Ma, S., Song, X., and Huang, J. (2007), “Supervised group Lasso with applications to microarray data analysis,” *BMC Bioinformatics*, 8, 60–60.
- Mitchell, T. J. and Beauchamp, J. J. (1988), “Bayesian variable selection in linear regression,” *JASA*, 1023–1032.
- Neal, R. M. (2010), “MCMC using Hamiltonian dynamics,” in *Handbook of Markov Chain Monte Carlo* (eds S. Brooks, A. Gelman, G. Jones, XL Meng). Chapman and Hall/CRC Press.
- Park, T. and Casella, G. (2008), “The bayesian lasso,” *JASA*, 103, 681–686.
- Polson, N. G. and Scott, J. G. (2010), “Shrink globally, act locally: Sparse Bayesian regularization and prediction,” *Bayesian Statistics*, 9, 501538.
- (2012a), “Good, great, or lucky? Screening for firms with sustained superior performance using heavy-tailed priors,” *The Annals of Applied Statistics*, 6, 161–185, zentralblatt MATH identifier1235.91144, Mathematical Reviews number (MathSciNet) MR2951533.
- (2012b), “Local shrinkage rules, Levy processes and regularized regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 287311.
- (2012c), “On the half-Cauchy prior for a global scale parameter,” *Bayesian Analysis*, 7, 887902.
- Robert, C. P. (2013), “On the Jeffreys-Lindley’s paradox,” *ArXiv*, 1303.5973v, 1–13.
- Rockova, V. and George, E. I. (2013), “EMVS: The EM Approach to Bayesian Variable Selection,” *Journal of the American Statistical Association*, forthcoming.
- Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., and Buckley, C. (2004), “Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage,” *Biometrics*, 60, 812–819.
- Shafer, G. (1982), “Lindleys paradox,” *Journal of the American Statistical Association*, 77, 325–334.
- Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., DAmico, A. V., Richie, J. P., and Others (2002), “Gene expression correlates of clinical prostate cancer behavior,” *Cancer Cell*, 1, 203–209.
- Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *JRSS-B*, 58, 267–288.

- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), “Diagnosis of multiple cancer types by shrunken centroids of gene expression,” *Proceedings of the National Academy of Sciences*, 99, 6567.
- Tolosi, L. and Lengauer, T. (2011), “Classification with correlated features: unreliability of feature ranking and solution,” *Bioinformatics*, 27, 1986–1994.
- van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014), “The Horseshoe Estimator: Posterior Concentration around Nearly Black Vectors,” *arXiv:1404.0202 [math, stat]*.
- Yi, N. and Xu, S. (2008), “Bayesian LASSO for quantitative trait loci mapping,” *Genetics*, 179, 1045–1055.
- Zou, H. (2006), “The adaptive lasso and its oracle properties,” *JASA*, 101, 1418–1429.

## Appendices

### A Computational Method for BLRHL

This section is a continued discussion from Section 3.3 about our computational method.

#### A.1 Initial Values for Gibbs Sampling

The initial values for  $\delta_{0:p,1:K}$  are the coefficients of Bayes discriminant rule based on Gaussian distributions, whose mean vectors are estimated by the medians of Markov chain samples produced by the method described in Li (2012), and whose covariance matrix is estimated by equally weighted average of the sample covariance and the identity matrix.

#### A.2 Updating $\delta_{0:p,1:K}$ with Hamiltonian Monte Carlo

Suppose we want to sample from a  $d$ -dimensional distribution with PDF proportional to  $\exp(-U(\mathbf{q}))$ , or construct a transformation leaving it invariant. For our problem,  $U(\mathbf{q})$  is minus log of posterior distribution of  $\mathbf{q} = \delta_{0:p,1:K}$  (ie, minus log of (15)).

We will augment  $\mathbf{q}$  with a set of auxiliary variables  $\mathbf{p}$  independently distributed with  $N(0, 1)$ , also independent of  $\mathbf{q}$ . For this purpose we will randomly draw a  $\mathbf{p}$  independently from  $N(0, 1)$ . In physics,  $\mathbf{p}$  is interpreted as momentums of particles. Next we will transform  $(\mathbf{q}, \mathbf{p})$  in a way that leaves invariant  $\exp(H(\mathbf{q}, \mathbf{p}))$  — the joint distribution of  $(\mathbf{q}, \mathbf{p})$ , where  $H(\mathbf{q}, \mathbf{p})$  is often called **Hamiltonian**, which is given by:

$$H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + K(\mathbf{p}) = U(\mathbf{q}) + \frac{1}{2} \sum_{i=1}^d p_i^2.$$

At the end of this transformation, we will discard  $\mathbf{q}$ , obtaining a new  $\mathbf{p}$  that is still distributed with  $\exp(-U(\mathbf{p}))$ .

The method for transforming  $(\mathbf{q}, \mathbf{p})$  is inspired by the Hamiltonian dynamics, in which  $(\mathbf{q}, \mathbf{p})$  moves along a continuous time  $\tau$  according to the following differential equations:

$$\begin{aligned}\frac{dq_i(\tau)}{d\tau} &= \frac{\partial H}{\partial p_i} = \frac{\partial K}{\partial p_i} = p_i \\ \frac{dp_i(\tau)}{d\tau} &= -\frac{\partial H}{\partial q_i} = -\frac{\partial U}{\partial q_i}\end{aligned}$$

It can be shown that Hamiltonian dynamic keeps  $H$  unchanged and preserves volume (see details from [Neal \(2010\)](#)). These are the crucial properties of Hamiltonian dynamics that make it a good proposal distribution for Metropolis sampling.

In computer implementation, Hamiltonian dynamics must be approximated by discretized time, using small stepsize  $\epsilon$ . Leapfrog transformation is one of such methods, which is shown to be better than several other alternatives. *One* leapfrog transformation with stepsize  $\epsilon_i$  is described as follows:

**One Leapfrog Transformation**

$$\begin{aligned}p_i &\leftarrow p_i - (\epsilon_i/2) \frac{\partial U}{\partial q_i}(q_i), \\ q_i &\leftarrow q_i + \epsilon_i p_i, \\ p_i &\leftarrow p_i - (\epsilon_i/2) \frac{\partial U}{\partial q_i}(q_i).\end{aligned}$$

Note that, above we apply leapfrog transformations independently to each pair  $(q_i, p_i)$ , with different stepsizes. Applying a series of leapfrog transformations, we *deterministically* transform  $(q_i, p_i)$  to a new state, denoted by  $(q_i^*, p_i^*)$ , for  $i = 1, \dots, d$ . This transformation has the following properties:

- Value of  $H$  is nearly unchanged, if  $\epsilon_i$  is small enough. This is because each leapfrog transformation is a good approximation to Hamiltonian dynamics.
- Reversibility: following the same series of leapfrog transformations,  $(q_i^*, -p_i^*)$  will be transformed back to  $(q_i, -p_i)$ . We therefore add a negation ahead of these leapfrog transformations to form an exactly “reversible” transformation between  $(q_i, -p_i)$  and  $(q_i^*, p_i^*)$ .
- Volume preservation: the Jacobian of this transformation is 1.

A series of leapfrog transformations cannot leave  $H$  exactly unchanged. But we will use it only as a proposal distribution in Metropolis sampling. That is, at the end of the leapfrog transformations,  $(\mathbf{q}^*, \mathbf{p}^*)$  will be accepted or rejected randomly according to Metropolis acceptance probability. As a summary, the algorithm of Hamiltonian Monte Carlo is presented completely below:

## Hamiltonian Monte Carlo (HMC) with Leapfrog Transformations

Starting from current state  $\mathbf{q}$ , update it with the following steps:

**Step 1:** Draw elements of  $-\mathbf{p}$  independently from  $N(0, 1)$

**Step 2:** Transform  $(\mathbf{q}, -\mathbf{p})$  with the following two steps:

(a) Negate  $-\mathbf{p}$  to  $\mathbf{p}$ .

(b) Apply leapfrog transformation  $\ell$  times to transform  $(\mathbf{q}, \mathbf{p})$  to a new state  $(\mathbf{q}^*, \mathbf{p}^*)$ . A trajectory connecting the states along these  $\ell$  transformations are called *leapfrog trajectory* with length  $\ell$ .

**Step 3:** Decide whether or not to accept  $(\mathbf{q}^*, \mathbf{p}^*)$  with a probability given by:

$$\min \left( 1, \exp \left( - \left[ H(\mathbf{q}^*, \mathbf{p}^*) - H(\mathbf{q}, -\mathbf{p}) \right] \right) \right). \quad (20)$$

If the result is a rejection, set  $(\mathbf{q}^*, \mathbf{p}^*) = (\mathbf{q}, -\mathbf{p})$ .

At last, retaining  $\mathbf{q}^*$ , with  $\mathbf{p}^*$  discarded.

To implement HMC, we need to choose appropriate stepsizes  $\epsilon_i$  and  $\ell$  — length of leapfrog trajectory. The stepsizes determine how well the leapfrog transformation can approximate Hamiltonian dynamics. If  $\epsilon_i$  is too large, leapfrog transformation may diverge, resulting in very high rejection rate, and very poor performance; otherwise, it may move too slowly, even though with very low rejection rate. An ad-hoc choice is a value close to the reciprocal of square root of the 2nd-order partial derivative of  $U$  with respect to  $q_i$ , which automatically accounts for the width of posterior distribution of  $q_i$ . We therefore adjust the reciprocals by an adjustment factor  $\epsilon$  (which usually should be between 0.1 and 0.5, called *HMC stepsize adjustment*). The exact value of the adjustment factor can be chosen empirically such that the HMC rejection rate is less than but close to 0.2. There is often a critical point beyond which Hamiltonian diverges. A value slightly smaller than this critical point often works the best. According to our experiences, a value close to 0.25 often works well. A good thing for this choice is that it is independent of the choice of  $\ell$  — the length of leapfrog trajectory, because the value of Hamiltonian, actually the whole leapfrog trajectory, changes nearly cyclically as long as it doesn't diverge.

After we determine the stepsize adjustment  $\epsilon$ , we will determine the length of leapfrog trajectory  $\ell$ . The fact is that the appropriate values of  $\ell$  are different in two phases. In the *initial phase*, a small value  $\ell_1$  should be used such that the Gibbs sampling can quickly dissipates the value of  $U$ , as well as more frequently update the hyperparameter  $w$ . The exact choice of  $\ell$  for initial phase can be made empirically by looking at how fast the Gibbs sampling converges with different values of  $\ell$ . For our problem,  $\ell_1 = 10$  or 5 seemingly works fairly well. Another reason for the initial

phase is that a very long trajectory starting from the initial value has very high chance of being rejected. After running initial phase for a while, we need to choose a larger value, denoted by  $\ell_2$  to suppress random walk. This phase is called *sampling phase*. The advantage of using HMC instead of other samplers is that HMC can keep moving in the direction determined by the gradients of  $U$  without random walk. We therefore should choose fairly large  $\ell$  (at least larger than 1, when  $\ell = 1$ , HMC is Langevin Metropolis-Hasting method) such that the leapfrog transformation can reach a distant point from the starting one. However, if  $\ell$  is excessively large, the leapfrog transformation will reverse the direction and move back to the region near the starting point. The choice of  $\ell$  for this phase can be made empirically by looking at the curve of distance of  $\mathbf{q}$  from the origin along a very long trajectory, which changes cyclically. We will choose the largest  $\ell$  such that leapfrog transformation can move in a direction, ie, the distance of  $\mathbf{q}$  changes monotonically. From our experiences,  $\ell_2 = 50$  or 100 works well for many problems.

To apply HMC to sample (15), we need to compute the 1st-order partial derivatives of  $-\log(P(\boldsymbol{\delta}_{0:p,1:K}|\boldsymbol{\sigma}_{0:p}^2))$  with respect to  $\delta_{jk}$  for  $j = 0, \dots, p, k = 1, \dots, K$ , which is equal to the sum of the following two partial derivatives of  $L$  and minus log prior:

$$-\frac{\partial \log(L(\boldsymbol{\delta}_{0:p,1:K}))}{\partial \delta_{jk}} = \sum_{i=1}^n x_{ij}(P(y_i = k + 1|x_{ij}, \boldsymbol{\delta}_{0:p,1:K})) - I(y_i = k + 1), \quad (21)$$

$$-\frac{\partial \log(P(\boldsymbol{\delta}_{0:p,1:K}|\boldsymbol{\sigma}_{0:p}^2))}{\partial \delta_{jk}} = \left( \delta_{jk} - \sum_{k=1}^K \delta_{jk}/C \right) / \sigma_j^2. \quad (22)$$

An ad-hoc stepsizes  $\epsilon_{jk}$  is a value close to the reciprocal of the 2nd-order derivatives of  $U$ . We also use an estimate of the 2nd-order derivatives of  $U$ , which should be independent of current values of  $\boldsymbol{\delta}_{0:p,1:K}$  but could be dependent on  $\boldsymbol{\sigma}_{0:p}^2$ . They are the sum of the following two values:

$$\begin{aligned} -\frac{\partial^2 \log(L(\boldsymbol{\delta}_{0:p,1:K}))}{\partial^2 \delta_{jk}} &\approx \sum_{i=1}^n x_{ij}^2/4, \\ -\frac{\partial^2 \log(P(\boldsymbol{\delta}_{0:p,1:K}|\boldsymbol{\sigma}_{0:p}^2))}{\partial^2 \delta_{jk}} &= \frac{C-1}{C} \frac{1}{\sigma_j^2}. \end{aligned}$$

### A.3 Restricted Gibbs Sampling

When  $p$  is large, the dominating computing in applying HMC is for obtaining values of the linear functions  $\delta_{0,k} + \mathbf{x}_{i,1:p} \boldsymbol{\delta}_{1:p,k}$  for  $i = 1, \dots, n$  and  $k = 1, \dots, K$ , with which we can compute the log likelihood and its partial derivatives with respect to  $\boldsymbol{\delta}_{0:p,1:K}$  very easily.

A belief in high-dimensional classification is that most features are irrelevant and therefore most coefficients concentrate very close to 0 in a local mode of posterior. It is therefore useless to update them very often. A useful computational trick for reducing computation time is that, for each

iteration of Gibbs sampling, we update only those features with  $\sigma_j$  greater than a small threshold  $\zeta$  without much loss of efficiency. However, even a fairly small  $\zeta$  can cut off many coefficients from being updated. The consequence is that the computation time for each iteration of Gibbs sampling is reduced substantially, since we can reuse from the last iteration the sum of a large number of  $x_{i,j}\delta_{j,c}$  related to those small coefficients, which are to be fixed. We call this trick **restricted Gibbs sampling**. We want to point out that this method can be justified with Markov chain theory, therefore our computation using this trick is still an exact Markov chain simulation. The essential effect of this trick is updating those important features more often than a large number of coefficients for irrelevant features. However, note that when  $\zeta$  is chosen to be very large, only a few (say ones) coefficients are updated using HMC, and we therefore lose the ability of HMC in suppressing random walk. The consequence is that Markov chain may be harder to travel across the modes, as travelling from one mode to the other needs to update a very small coefficient to a large value. Further research is needed to find the optimal choice of  $\zeta$ . The implementation used in our examples chose  $\zeta = 0.05$  when  $\alpha$  is set to 1, for which about 10% of coefficients are updated in each iteration.

## B Notations of Prior and MCMC Settings in BLRHL

First, one needs to choose the prior type from **t**, **ghs**, and **neg**. For each choice of prior type, one needs to set these parameters for the prior and MCMC computation:

- $\alpha, \log(w)$ : degree freedom (df) and log square scale of **t**/**ghs**/**neg** prior.
- $n_1, \ell_1$ : number of Gibbs sampling iterations and length of trajectory in initial phase.
- $n_2, \ell_2$ : number of Gibbs sampling iterations and length of trajectory in sampling phase.
- $\zeta$ : the coefficients with  $\sigma_j$  smaller than  $\zeta$  are fixed in current HMC updating.
- $\epsilon$ : stepsize adjustment multiplied to the 2nd order partial derivatives of log posterior.
- the prior variances  $\sigma_0^2$  for the intercepts are always set to 2000.