Josephine Asafu-Adjei[1] / Mahlet G. Tadesse[2] / Brent Coull[3] / Raji Balasubramanian[4] / Michael Lev[5] / Lee Schwamm[6] / Rebecca Betensky[7]

Bayesian Variable Selection Methods for Matched Case-Control Studies

[1] Departmetn of Biostatistics, University of North Carolina at Chapel Hill, 3104-E McGavran-Greenberg Hall, Chapel Hill, NC 27515, USA; Department of Nursing, University of North Carolina at Chapel Hill, 2005 Carrington Hall, Chapel Hill, NC 27515, USA, E-mail: jasafuad@email.unc.edu
[2] Department of Mathematics & Statistics, Georgetown University, Washington, DC, USA
[3] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[4] University of Massachusetts, Amherst, MA, USA
[5] Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[6] Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[7] Harvard University, Cambridge, MA 02138, USA

Abstract:
Matched case-control designs are currently used in many biomedical applications. To ensure high efficiency and statistical power in identifying features that best discriminate cases from controls, it is important to account for the use of matched designs. However, in the setting of high dimensional data, few variable selection methods account for matching. Bayesian approaches to variable selection have several advantages, including the fact that such approaches visit a wider range of model subsets. In this paper, we propose a variable selection method to account for case-control matching in a Bayesian context and apply it using simulation studies, a matched brain imaging study conducted at Massachusetts General Hospital, and a matched cardiovascular biomarker study conducted by the High Risk Plaque Initiative.

Keywords: Bayesian analysis, conditional logistic regression, matched case-control studies, variable selection methods

DOI: 10.1515/ijb-2016-0043

1 Introduction

In matched case-control studies, subjects from a particular diagnostic group(s) (cases) are matched with those from a comparison group(s) (controls) based on important demographic characteristics, such as age or gender. Matching on these potential confounders can result in substantial improvements in efficiency and statistical power to identify feature variables that are associated with case-control status [1]. Matched case-control studies are becoming increasingly popular in studies involving high dimensional data that aim to identify subsets of relevant features. Examples include brain imaging studies aimed at identifying brain regions associated with comorbidities or genomic studies focused on discovery of cancer biomarkers. Despite the popularity of matched high dimensional studies, it is quite common for these studies to ignore the matched design used when applying variable selection techniques (e.g., Anglim et al. [2], Westman et al. [3]). Failure to account for matching has been shown to decrease variable selection accuracy [1] and lead to biased results [4].

Currently, there are several frequentist variable selection approaches for matched high dimensional data that incorporate case-control matching. Tan et al. [5] develop a modified paired t-test statistic to identify a subset of relevant features that serves as a basis for classification via support vector machines (SVM). Although matching is accounted for with respect to variable selection, it is ignored with respect to building the SVM classifier. In addition, their approach of identifying relevant features involves univariate tests, which do not control for the effects of other features and, thus, can lead to spurious identification of relevant features. Adewale et al. [6] develop two modified versions of boosting for correlated binary response data. The first version utilizes a loss function for the generic gradient descent boosting algorithm [7] that handles correlated binary responses. The second version modifies the likelihood optimization boosting algorithm [7] via a generalized linear mixed modeling approach in order to handle correlated binary responses. However, boosting approaches may have trouble identifying interactions among different features [1] and have decreased accuracy for data

Josephine Asafu-Adjei is the corresponding author.

sets with relatively small sample sizes [8]. In matched studies, a standard analytic approach to identify features significantly associated with case-control status is conditional logistic regression (hereafter denoted as CLR) modeling. However, for high dimensional data sets, CLR can not only become computationally intensive, but can also quickly run into model convergence problems. To address these computational issues, Balasubramanian et al. [1] develop a random penalized CLR (hereafter denoted as R-PCLR) variant that merges ridge penalized CLR [9] with Random Forests [10] to identify relevant features and two-way interactions. In addition, Qian et al. [11] develop two selection methods based on the conditional and unconditional logistic likelihood functions, as well as the lasso and elastic net penalties [12, 13]. Their first method uses a two-stage approach to estimate the logistic regression parameters that are subsequently used to predict case-control status, while their second method simultaneously computes both the regression coefficient estimates and the predicted values for case-control status.

Alternately, one can approach variable selection from a Bayesian standpoint, which has several important benefits. With regards to variable selection, penalized methods identify features for inclusion by determining which of them have nonzero model coefficient estimates. Bayesian variable selection (BVS) provides more information by giving not only coefficient estimates, but also inclusion probability estimates for each feature. One common BVS approach utilizes the spike and slab prior [14–16], which involves assigning hierarchical priors to the regression coefficients by introducing indicator variables to determine whether each predictor should be considered for inclusion or removal from the model. The use of this prior is fairly widespread among BVS techniques [17–21], mainly due to its flexibility and ease of application, particularly for high dimensional data [22]. Lee et al. [23], Sha et al. [24], and Zhou et al. [25] all developed BVS approaches based on spike-and-slab priors [14–16] for binary outcomes, with applications to genetic microarray data. While [23] and Sha et al. [24] developed their approach using multinomial probit models by introducing latent variables, Zhou et al. [25] used logistic models. BVS methods have also yielded relatively high selection accuracy for linear regression [16, 26], and have been shown to efficiently handle ultra-high dimensional data sets [27]. Another key benefit of BVS is that it can naturally incorporate auxiliary information regarding different spatial, network, or other correlation or grouping structures among features. For instance, Smith and Fahrmeir [28] incorporate spatial correlation among features with direct applications to imaging data, while Stingo et al. [29] account for membership in a particular genetic pathway and the relationship among genes in that pathway.

To account for matching in BVS, we can specify the likelihood based on a CLR model. Based on the key benefits of BVS and CLR modeling, we propose a new methodology that formulates BVS in a CLR framework (hereafter denoted as BVS CLR) and evaluate its performance using simulation and actual studies. For comparative purposes, we also examine the performance of CLR using the lasso penalty (denoted as lasso CLR). In our applications to actual studies, we also assess the performance of BVS CLR relative to that of R-PCLR and the methods proposed by Qian et al. [11], both of which were based on lasso penalized conditional logistic likelihood functions. Several penalties exist for variable selection, e.g., elastic net, group lasso [30, 31], and sparse group lasso [32]. Since the variable selection method we propose does not take the correlation or group structure among features into account, it is more comparable to lasso. In addition, we compare our approach with that of the lasso penalty because of its ease of implementation for CLR using available software (e.g., R) relative to other variable selection penalties, e.g., SCAD [33] or adaptive lasso [34].

In Section Section 2, we specify the CLR model for paired data and describe our selection approach. We evaluate the performance of BVS CLR relative to lasso CLR using simulation studies in Section Section 3. In Section Section 4, we assess the performance of BVS CLR relative to lasso CLR, as well as R-PCLR and the selection approach of Qian et al. [11] which account for matching, using a matched brain imaging study of hospital acquired pneumonia (HAP) among stroke patients in Massachusetts General Hospital (MGH). We then perform a similar assessment in Section Section 5 using a matched study of biomarkers for near-term cardiovascular events, where we compare the performance of BVS CLR with that of lasso CLR and R-PCLR. In Section Section 6, we conclude with a discussion.

2 Bayesian Variable Selection for Paired Case-Control Data

Consider I case-control pairs, where $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijK})$ denotes the observed feature values and Z_{ij} denotes case-control status for the j^{th} member of the i^{th} pair ($i = 1, \ldots, I$; $j = 1, 2$), so that $Z_{ij} = 1$ for cases and 0 for controls. CLR models the probability that the first member of pair i is a case, given $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ and $Z_{i1} + Z_{i2} = 1$, as follows:

$$p_{i1} = P(Z_{i1} = 1 | Z_{i1} + Z_{i2} = 1, \mathbf{X}_{i1}, \mathbf{X}_{i2}) = \left\{1 + \exp\left[-\sum_{k=1}^{K} \beta_k (X_{i1,k} - X_{i2,k})\right]\right\}^{-1}, \quad (1)$$

where β_k denotes the coefficient or log odds ratio for feature X_k. For features with an effect on case-control status, i.e., relevant features, β_k is nonzero. From eq. (1), the conditional log-likelihood is given by

$$l_C(\boldsymbol{\beta}) = \log(L_C(\boldsymbol{\beta})) = \log\left(\prod_{i=1}^{I} p_{i1}^{Z_{i1}}\right). \tag{2}$$

We now introduce $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)$, a binary vector where γ_k is either 1 or 0 based on whether X_k is retained in eq. (1). We assign a mixture of normal and point mass priors to the coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$, where $\beta_k|\gamma_k \sim \gamma_k N(0, \sigma^2) + (1 - \gamma_k)\delta_0$, δ_0 corresponds to $\pi(\beta_k = 0) = 1$ and $\sigma^2 < 0$. In addition, $\gamma_k|\omega \sim$ Bernoulli(ω), where $\omega \sim$ Beta(c, d) and $c, d < 0$.

Based on the prior assumptions and the conditional likelihood in eq. (2), the posterior distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{Z}) \propto L_C(\boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \cdot \pi(\boldsymbol{\gamma}), \tag{3}$$

such that $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} \pi(\beta_k|\gamma_k)$ and $\pi(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \int_0^1 \pi(\gamma_k|\omega) \cdot \pi(\omega) d\omega$.

We use Markov chain Monte Carlo (MCMC) sampling via the Metropolis-Hastings (MH) algorithm to estimate the distribution in eq. (3). Starting from random initial values for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we apply the following moves at each of S MCMC iterations:

1. **Move 1**
   - Add or remove X_r ($r \in \{1, \dots, K\}$) by choosing γ_r at random and changing its state, i.e., $\gamma_r^* = 1 - \gamma_r^{(t)}$, where γ_r^* and $\gamma_r^{(t)}$ denote the proposed and current values of γ_r. If $\gamma_r^* = 1$, generate the proposed β_r^* value for X_r from $N(0, \tau_1 \sigma_{r*}^2)$, where $\tau_1 (> 0)$ is a proposal tuning parameter and σ_{r*}^2 is the estimated variance corresponding to the MLE of β_r from a univariate CLR model on X_r. Otherwise, let $\beta_r^* = 0$.

     Based on the posterior probabilities $p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{X}, \mathbf{Z})$ and proposal densities $q(\cdot|\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $q(\cdot|\boldsymbol{\gamma})$, compute the acceptance ratio

$$A = \frac{p(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*|\boldsymbol{X}, \boldsymbol{Z})}{p(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}|\boldsymbol{X}, \boldsymbol{Z})} \cdot \left[\frac{q(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\gamma}^*)}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t)})} \cdot \frac{q(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)})}{q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^*)}\right]. \tag{4}$$

     Accept the proposed values $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ with probability $\min(A, 1)$ and retain the current values $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})$ otherwise.

2. **Move 2**
   - For each included X_k, i.e., X_k with $\gamma_k^{(t)} = 1$, generate the proposed β_k^* value from $N(\beta_k^{(t)}, \tau_2 \sigma_{k*}^2)$, where $\beta_k^{(t)}$ denotes the current value of β_k, $\tau_2 (> 0)$ is a proposal tuning parameter, and σ_{k*}^2 is the variance estimate for the univariate CLR MLE of β_k.
   - Since we do not update $\boldsymbol{\gamma}$ and only update the β values for the included X_k, the acceptance ratio in eq. (4) reduces to

$$A = \frac{p(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)}|\mathbf{X}, \mathbf{Z})}{p(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}|\mathbf{X}, \mathbf{Z})} = \frac{L_C(\boldsymbol{\beta}^*) \cdot \pi(\boldsymbol{\beta}^*|\boldsymbol{\gamma}^{(t)})}{L_C(\boldsymbol{\beta}^{(t)}) \cdot \pi(\boldsymbol{\beta}^{(t)}|\boldsymbol{\gamma}^{(t)})}.$$

     Accept the proposed values $\boldsymbol{\beta}^*$ with probability $\min(A, 1)$ and retain the current values $\boldsymbol{\beta}^{(t)}$ otherwise.

For each move type, we neither wanted to have proposal variances that are too small to avoid encountering mixing issues and high autocorrelations, nor did we want to have variances that are too large to avoid low acceptance rates. According to Gelman et al. [35] and Roberts et al. [36], asymptotically optimal acceptance rates for random walk Metropolis algorithms are approximately equal to 25%. Therefore, τ_1 and τ_2 are chosen to ensure that the acceptance rates for each move type are between 20% and 30%. Alternately, we could have used an adaptive MH algorithm, as in Lamnisos et al. [37]. However, the standard MH algorithm that we specify has good convergence properties and is easy to implement.

We then obtain the sequence $\left\{(\boldsymbol{\beta}^{[1]}, \boldsymbol{\gamma}^{[1]}), \dots, (\boldsymbol{\beta}^{[S]}, \boldsymbol{\gamma}^{[S]})\right\}$. Assuming a burn-in period of B iterations, estimates of the posterior inclusion probabilities $p(\gamma_k|\boldsymbol{X}, \boldsymbol{Z})$ and coefficients β_k are given by

$$\hat{p}(\gamma_k = 1|\mathbf{X}, \mathbf{Z}) = \frac{\sum_{v=B+1}^{S} \gamma_k^{[v]}}{S - B}, \quad \hat{\beta}_k = \hat{p}(\gamma_k = 1|\mathbf{X}, \mathbf{Z}) \cdot \left[\frac{\sum_{v=B+1}^{S} \gamma_k^{[v]} \beta_k^{[g]}}{\sum_{v=B+1}^{S} \gamma_k^{[v]}}\right]. \tag{5}$$

A similar averaging approach can be used to obtain variance estimates of $\hat{\beta}_k$.

Since we use standard spike-and-slab priors and Metropolis-Hastings moves to update the model parameters, the ergodicity of our MCMC sampler is guaranteed. The introduction of Move 2 does not compromise ergodicity; it is used to provide faster convergence by refining the parameter space within the selected model. To determine the convergence of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in our later applications of BVS CLR to actual data, we consider a multivariate version of Gelman and Rubin's potential scale reduction factor using the **coda** R package, where values substantially above 1 indicate lack of convergence [38, 39].

3 Simulation Studies

We first assess the operating performance of BVS CLR using several simulated datasets. In doing so, we focus on the level of accuracy in identifying the relevant and non-relevant features and in predicting case-control status. We also assess coefficient estimation accuracy by examining the mean squared error (MSE) of the β_k estimates. In our study, we examine the performance of BVS CLR relative to that of lasso CLR with respect to selection and prediction accuracy. Although variance estimation approaches for lasso regression exist [40, 41], we take a more straightforward approach and consider, for both BVS and lasso CLR, a rough MSE approximation using the empirical mean of the squared deviations between the estimated and actual β_k values. We run all analyses using R software version 3.1.2 [42].

3.1 Simulation Designs

3.1.1 Simulation of Paired Response and Feature Data

First, we simulate $M = 10,000$ observations for variables we term as age and gender. Gender values are generated from a Bernoulli(0.5) distribution. Age values are generated from a truncated normal distribution in the range $(0, 60)$ with mean 30 and variance 100, and then rounded to the nearest integer. To examine the performance of BVS CLR for different data types, we consider the cases of binary and normal features. To simulate $\boldsymbol{X}$ values such that the first L ($L < K$) features are related to gender and standardized age and the first Q ($Q < L < K$) features are relevant, we use the following approach:

- **Binary case** Assume $p_{m,k} = P(X_k = 1) = \left\{1 + \exp\left[\text{age.std}_m + 1.5\text{gender}_m\right]\right\}^{-1}$ for observation m ($m = 1, \ldots, M$) for $k = 1, \ldots, L$ and $p_{m,k} = 0.5$ otherwise, where age.std_m and gender_m denote standardized age and gender for observation m. Each age value is standardized by subtracting its mean value and dividing by its standard deviation. Since age is included as a linear effect in our model for simulating case-control status $p_{m,k}$, we standardize age to ensure that the regression coefficient we use corresponds to a reasonable effect size. To build correlation among $(X_{m,1}, \ldots, X_{m,K})$, we specify correlation matrix $\boldsymbol{\Sigma}$ with entries ρ_{ij} obtained from the phi coefficients computed from the MGH brain imaging data, where a phi coefficient measures the association between a pair of binary features. We consider the cases of both high correlation (ρ_{ij} obtained from the phi coefficients ranked in the top 50 in absolute value for $X_1, \ldots, X_Q$; ρ_{ij} obtained from the remaining phi coefficient values for $X_{Q+1}, \ldots, X_K$; $X_1, \ldots, X_Q$ are uncorrelated with $X_{Q+1}, \ldots, X_K$) and low correlation (for $X_1, \ldots, X_K$, ρ_{ij} obtained from the phi coefficient values not ranked in the top 50 in absolute value) among the K features. Observations $(X_{m,1}, \ldots, X_{m,K})$ are simulated using the **bindata** R package.
- **Normal case** Using the **mvtnorm** R package, we generate $X_{m,k}$ from a normal distribution with mean $\text{age.std}_m + 1.5 \cdot \text{gender}_m$ for $k = 1, \ldots, L$ and 0 otherwise and covariance matrix $\boldsymbol{\Sigma}$ with entries ρ_{ij} equal to the correlation coefficients computed from the matched biomarker study discussed in Section Section 5. As in the binary case, we consider the cases of both high correlation (ρ_{ij} obtained from the correlation values ranked in the top 50 in absolute value for $X_1, \ldots, X_Q$; ρ_{ij} obtained from the remaining correlation values for $X_{Q+1}, \ldots, X_K$; $X_1, \ldots, X_Q$ are uncorrelated with $X_{Q+1}, \ldots, X_K$) and low correlation (for $X_1, \ldots, X_K$, ρ_{ij} obtained from the correlation values not ranked in the top 50 in absolute value) among the K features.

Case-control status Z_m is generated from a Bernoulli(φ_m) distribution with $\varphi_m = P(Z_m = 1|\mathbf{X}_m) = \left\{1 + \exp\left[-\sum_{k=1}^{K} \beta_k X_k - 0.3 \cdot \text{cen.age}_m - 0.5 \cdot \text{gender}_m\right]\right\}^{-1}$, where cen.age_m denotes the value of age for observation m that has been centered to have mean 0. Here, $\beta_1, \ldots, \beta_Q$ fall in the range $[1, 2]$ in magnitude in the binary case, and $[1, 1.3]$ in the normal case. For each data type, values for $\beta_1, \ldots, \beta_Q$, which capture effect sizes for the relevant features on the log odds scale, were chosen to ensure obtaining realistic odds ratios. In the normal case, BVS CLR did not converge when $\beta_1, \ldots, \beta_Q$ fell in the range $[1, 2]$. Our predictors in this case were not

standardized and had a fairly wide range of values, which leads to large values for the linear predictors and, thus, a log-likelihood that diverges. Therefore, we decreased the range of values in the normal case, relative to the binary case. All remaining $\boldsymbol{\beta}$ elements are set to zero in both cases. From this population, we randomly select $I = 50$ or $I = 200$ observations with $Z_m = 1$ as cases and match them with observations with $Z_m = 0$ (controls) based on age and gender, and let these $(\mathbf{X}_{ij}, Z_{ij})$ observations constitute the training set. Omitting this set, we randomly select $N = 2,000$ $(\mathbf{X}_n, Z_n)$ $(n = 1, \ldots, N)$ observations from the remaining population to constitute the test set. We also consider the following scenarios:

1. Scenario 1: $K = 20$ features, where $Q = 2, 5, 10$ are relevant
2. Scenario 2: $K = 100$ features, where $Q = 10, 25, 50$ are relevant
3. Scenario 3: $K = 600$ features, where $Q = 60, 150, 300$ are relevant

For each data type (binary/normal), correlation level, and scenario, we simulate 100 datasets for scenarios 1 and 2, and 20 datasets for scenario 3 due to the amount of computation time involved. Each dataset consists of a training set and test set as previously discussed. We now describe how BVS and lasso CLR are applied to each dataset.

3.1.2 Application of BVS CLR

We first apply BVS CLR to each training set, where we set the prior variance σ^2 for $\beta_k|\gamma_k$ to 1. Since the estimated standard errors for the univariate CLR coefficient estimates were all in the 10^{-2} range in this study, a value of 1 for σ^2 is fairly noninformative in this case. To ensure that a sufficient number of features are considered for inclusion, we assign a Beta$(5, 5)$ prior to ω in all scenarios. Although we omit their results in our discussion, we note that preliminary analyses have demonstrated the robustness of BVS CLR to specification of both σ^2 (values considered ranged from 0.1 to 10) and the Beta(c, d) prior.

We run BVS CLR for $S = 50,000$ iterations ($B = 20,000$ burn-in) in scenario 1, $S = 80,000$ iterations ($B = 30,000$ burn-in) in scenario 2, and $S = 100,000$ iterations ($B = 50,000$ burn-in) in scenario 3. Along with the estimates $\hat{p}(\gamma_k = 1|\mathbf{X}, \mathbf{Z})$ and $\hat{\beta}_k$ in (5), we compute the variance estimates of $\hat{\beta}_k$. For each post burn-in iteration v $(v = B + 1, \ldots, S)$, we compute the estimated case probability $p_n^{[v]} = \left\{1 + exp\left[-\sum_{k=1}^{K} \beta_k^{[v]} X_{n,k}\right]\right\}^{-1}$ for the n^{th} test set observation. Using $p_n^{[v]}$, we compute the Bayesian model averaged (BMA) case probability $\hat{p}_{n,BMA}$ using the following approaches: (1) average $p_n^{[B+1]}, \ldots, p_n^{[S]}$, (2) average $Z_n^{[B+1]}, \ldots, Z_n^{[S]}$, where $Z_n^v = 1$ if $p_n^{[v]} \geq 0.5$ and 0 otherwise, and (3) generate Z_n^v from a Bernoulli$(p_n^{[v]})$ distribution and average $Z_n^{[B+1]}, \ldots, Z_n^{[S]}$. We discuss the results obtained using the first approach, although all three approaches give similar results.

3.1.3 Application of Lasso CLR

We also examine the performance of lasso CLR, which is based on the penalized conditional log-likelihood

$$l_C(\boldsymbol{\beta}) - \lambda \sum_{k=1}^{K} |\beta_k|. \tag{6}$$

In eq. (6), the feature values are standardized and $\lambda \geq 0$ is a tuning parameter that is typically estimated using V-fold cross validation. Use of the lasso penalty in eq. (6) shrinks all β_k estimates to zero, yielding a subset of features found to be relevant in eq. (1) due to having nonzero β_k estimates. Using the **survival** and **penalized** R packages, we apply 10-fold cross-validation to estimate λ and run lasso CLR on each training set, yielding coefficient estimates $\hat{\beta}_{k,las}$ and case probability estimates $\hat{p}_{n,las} = \left\{1 + exp\left[-\sum_{k=1}^{K} \hat{\beta}_{k,las} X_{n,k}\right]\right\}^{-1}$ for the n^{th} test set observation.

A schematic diagram summarizing the simulation details, and BVS and lasso CLR applications, is provided in Figure 1.

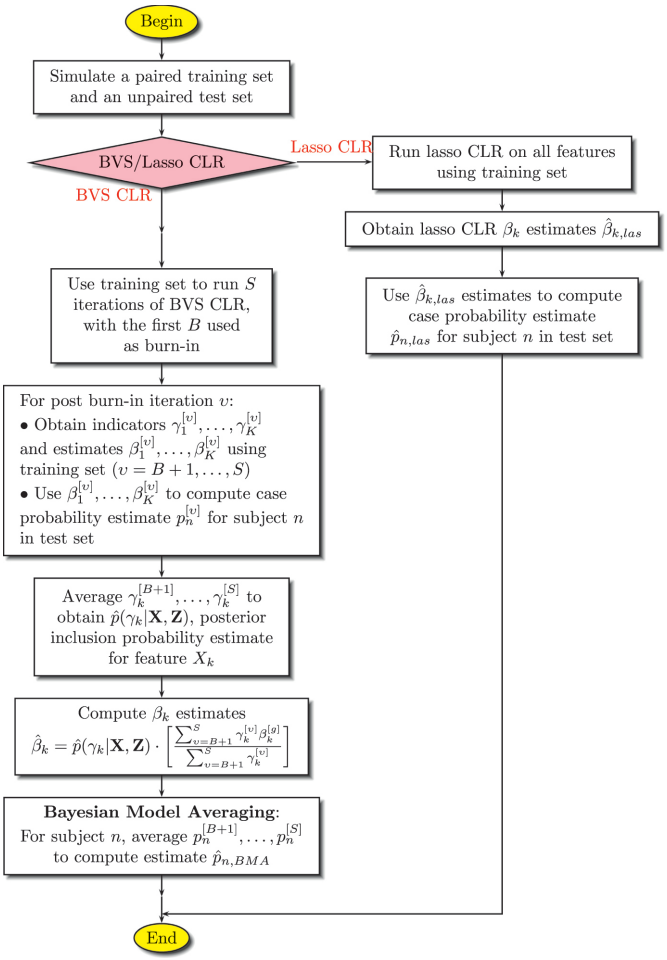


Figure 1 Schematic diagram of simulation and application details for each simulation.

3.2 Variable Selection Accuracy

To measure selection accuracy for BVS and lasso CLR, we compute the areas under the ROC curves (AUCs) using inclusion probability estimates $\hat{p}(\gamma_k = 1|\boldsymbol{X},\boldsymbol{Z})$ for BVS CLR and the magnitudes of the coefficient estimates $|\hat{\beta}_{k,las}|$ for lasso CLR. The AUC was used as a metric for selection accuracy because it gives an overall picture of the degree to which we can correctly identify relevant and non-relevant features. However, we acknowledge that, in practice, we would have to rely on the use of a defined threshold for identifying relevant and non-relevant features.

We present the median and interquartile range (IQR) values for these AUCs across simulations in the normal and binary cases for scenarios 1 and 2 in Appendix Table 5 and for scenario 3 in Appendix Table 6. For both approaches, we have that, regardless of the number of pairs I, data type, and correlation level among the relevant features, selection accuracy decreases as the total number of features K increases for a given number of relevant features Q, and as Q increases for a given K. We also have that for both approaches, selection accuracy increases with increasing I for a given Q and K, regardless of data type and correlation level. Also, as the level of correlation among the relevant features increases from low to high, selection accuracy remains comparable (magnitude of difference less than 0.05) or decreases in both the normal and binary cases. In cases where selection accuracy decreases, we have that the decrease is generally more pronounced for lasso CLR.

Regardless of the percentage of relevant features, BVS CLR generally yields higher selection accuracy than lasso CLR in the normal and binary cases when $K = 100$ features and $I = 50$ pairs. We also note an improved performance of BVS CLR for normally distributed features with high correlation when $K = 100$ features and $I = 200$ pairs. With $K = 20$ features, BVS CLR yields higher accuracy for the binary high correlation case with $I = 50$ regardless of the percentage of relevant features. When 25% and 50% of the features are relevant, this improved accuracy is also observed for the binary low correlation case and for the normal high correlation setting. With $K = 600$ features, the two methods have a comparable performance.

3.3 Prediction Accuracy

We also examine prediction accuracy for BVS and lasso CLR, which we measure using the AUCs based on the predicted case probabilities $\hat{p}_{n,BMA}$ and $\hat{p}_{n,las}$. For each approach, the median and IQR values for these AUCs across simulations are presented in Appendix Table 8 for scenarios 1 and 2 and in Appendix Table 9 for scenario 3.

In general, prediction accuracy is higher for the normal datasets than for the binary datasets, and also increases as the number of pairs I increases. Also, prediction accuracy is highest when 25% of the features are relevant for $K = 20, 100$, and remains comparable across the number of relevant features Q for $K = 600$.

BVS CLR has higher prediction accuracy than lasso CLR for the simulation setting with $I = 50$ pairs when $K = 100$ features and 50% of them are relevant, for both normal and binary cases with low or high correlation. We also observe improved prediction accuracy of BVS CLR for $K = 600$ features with 50% relevant in all cases (normal, binary, low or high correlation, I =50 or 200 pairs). For $K = 600$ with 25% relevant, BVS CLR has higher prediction accuracy in all cases when $I = 50$ pairs, and for the normal setting when $I = 200$ pairs. For $K = 600$ features with 10% relevant, BVS CLR has higher prediction accuracy for the normal low correlation case with $I = 50$ pairs. For all other settings, we find comparable prediction performance between BVS CLR and lasso CLR.

3.4 MSE for Coefficient Estimates

For each scenario, we then assess the level of coefficient estimation accuracy for BVS and lasso CLR by examining the average MSE for each feature across simulations, and then computing the median of these averaged MSEs across the relevant features and the non-relevant features. These results are reported in Appendix Table 11 and Table 13 for the relevant and non-relevant features, respectively.

Relative to lasso CLR, these median MSEs are generally lower for BVS CLR with respect to the relevant features and comparable with respect to the non-relevant features regardless of data type, correlation level among the relevant features, number of pairs, number of features (total and relevant). With fewer pairs, this decrease in MSE for BVS CLR with respect to the relevant features is more pronounced. For both approaches, we see that MSEs for the relevant features are generally larger for binary datasets, increase as the total number of features K and number of relevant features Q increase, and decrease as the number of pairs I increases. However, no apparent pattern emerges in our results as we increase the correlation level among the relevant features.

Although we do not present the results in our discussion, the level of estimation accuracy for BVS CLR was also assessed using coverage probabilities based on the 95% highest posterior density (HPD) intervals for each coefficient. In doing so, we observed high coverage probabilities across features and HPD intervals with widths that grew narrower with increasing I.

3.5 Convergence of BVS CLR

To explore the convergence performance of BVS CLR, we examine, across post burn-in iterations, trace plots of the log posterior probabilities in Figure 5 and the number of selected features in Figure 7 for the normal case when $I = 50$ pairs and 25% of the features are relevant and weakly correlated. Based on these plots, we do not see any evidence of non-convergence. Although we omit their results, similar patterns are found in the trace plots for all other simulation scenarios, data types, number of pairs, and levels of correlation among the relevant features. To explore the behavior of features with the highest inclusion probabilities, we consider, as an example, the case of $K = 20$ normal features of which $Q = 5$ are relevant and weakly correlated. For this case, we examine in Figure 8 a trace plot of the inclusion status (yes/no) across post burn-in iterations in one simulation for features X_1, X_3, X_4 whose inclusion probabilities are ranked among the top three for $I = 50$ pairs.

3.6 Accuracy under Reduced Coefficient Values

To assess the performance of BVS CLR relative to lasso CLR in the case when $\beta_1, \ldots, \beta_Q$ for the relevant features are relatively small, we also consider the case where $\beta_1, \ldots, \beta_Q$ fall in the range $[0.3, 0.7]$ in magnitude, in scenarios 1 ($K = 20$ features) and 2 ($K = 100$ features). For both the binary and normal cases, the median and IQR values for the selection and prediction accuracy AUCs across simulations are reported in Appendix Table 7 and Table 10, respectively. In Appendix Table 12 and Table 14, we report the medians of the averaged MSEs across

the relevant features and across the non-relevant features, respectively. In summary, relative to when $\beta_1, \ldots, \beta_Q$ are at least 1, the improvement in performance for BVS CLR compared with lasso CLR is more pronounced with respect to selection accuracy, relatively unchanged with respect to prediction accuracy and MSE for the non-relevant features, and less pronounced with respect to MSE for the relevant features.

4 Application to MGH HAP Imaging Study

4.1 Description

We first examine a case-control brain imaging study conducted by Kemmling et al. [43], in which acute ischemic stroke patients admitted to the Stroke Service Unit at MGH stroke service were classified according to whether or not they met the criteria for having hospital acquired pneumonia (HAP), i.e., suspicion or mention of pneumonia in the patient's medical record at least 48 hours after admission requiring antibiotic treatment. 215 acute ischemic stroke patients classified as having HAP were then matched with 215 non-HAP acute ischemic stroke patients on the basis of age, gender, and NIH stroke scale (NIHSS) upon admission. In this study, each patient was measured on both clinical and neuroimaging features. Clinical features include age, gender, admission NIHSS, length of hospitalization, as well as the presence/absence of the following: dysphagia, dyslipidemia, smoking history, coronary artery disease, diabetes mellitus, atrial fibrillation, hypertension, and in-hospital mortality. In Table 1, we present the summary statistics for each of these features.

Table 1 Clinical features of matched HAP and non-HAP patients [mean $\pm$ standard error for age and length of hospitalization; median (interquartile range) for admission NIHSS; n (%) for the remaining features].

Clinical Feature	HAP (n=215)	non-HAP (n=215)	p-Value
Age (years)[a]	72.2 ± 14.9	72.3 ± 13.9	–
Male[a]	116 (54%)	116 (54%)	–
Admission NIHSS[a]	$13\ (6-19)$	$13\ (6-19)$	–
Dysphagia	151 (70.2%)	151 (70.2%)	1.00[b]
Hypertension	146 (68.0%)	139 (64.7%)	0.53[b]
Dyslipidemia	74 (34.4%)	68 (31.6%)	0.59[b]
Diabetes mellitus	51 (23.7%)	44 (20.5%)	0.49[b]
Atrial fibrillation	65 (30.2%)	58 (27.0%)	0.51[b]
Smoking history	38 (17.7%)	35 (16.3%)	0.78[b]
Coronary artery disease	59 (27.4%)	51 (23.7%)	0.39[b]
Mortality	41 (19.1%)	38 (17.7%)	0.80[b]
Length of hospitalization (days)	12.8 ± 10.2	6.1 ± 4.6	< 0.0001[c]

[a] Feature used to match HAP and non-HAP patients.
[b] McNemar's test used to compare HAP and non-HAP patients.
[c] Wilcoxon signed-rank test used to compare HAP and non-HAP patients.

To extract the neuroimaging features for each patient, subacute ischemic brain lesions were first outlined slice-by-slice in diffusion weighted magnetic resonance imaging (MRI-DWI) or computerized tomography (CT) images, each of which were chosen with an acquisition time approximately 48 hours after symptom onset. MRI-DWI/CT images and their respective binary lesion masks were affine registered to standard MNI-152 space and manually corrected for registration errors. The same imaging protocol was used for all patients.

All lesion masks were then segmented into 68 pairs (left-right hemispheres) of cortical and subcortical/brainstem white matter brain regions based on the "Johns Hopkins University white-matter" and "Harvard-Oxford cortical structural" atlases, which were created by standardized anatomic labeling of multiple subjects linearly registered to MNI-152 standard space [44, 45]. Binarized atlases defining a specific structure with at least 25% probability of anatomic localization were used. For all patients, the percentage of infarction in a specific brain region was first measured and then dichotomized using its median value as having zero or positive infarction. Specifically, the neuroimaging feature examined for each brain region was the presence/absence of positive infarction in that region. Given the relatively small sample size, brain regions with positive infarction for less than 5% of HAP patients or non-HAP patients will yield unstable log odds ratio estimates [11]. Therefore, to stabilize calculations, we only consider the 130 brain regions that have positive infarction for at least 5% of HAP patients and 5% of non-HAP patients. Another examined neuroimaging feature was infarction volume. To avoid model fitting issues due to its highly skewed distribution and the presence of outliers,

we categorized infarction volume using the tertiles of its distribution. Two indicator variables were introduced to indicate whether infarction volume was at least equal to its first tertile, and whether it was at least equal to its second tertile.

In applying our proposed methodology, we aim to identify the subset of clinical and neuroimaging features that are most associated with, or most relevant to, having HAP. Although there is not necessarily a link between HAP and white matter infarctions, per se, associations between HAP and infarctions in specific brain regions have been found in prior studies. Kemmling et al. [43] showed that infarction in the right hemispheric peri-insular cortical regions was associated with the risk of acquiring HAP in acute ischemic stroke patients. They explain this finding by discussing the fact that previous studies have demonstrated the association of the right insular region with autonomically-induced immunosuppression and susceptibility to infection [46–48], as well as the association of right hemispheric peri-insular infarction to autonomic dysfunction and pathologic sympathetic activity [49].

In this application, we evaluate the performance of BVS CLR, and also assess its performance relative to lasso CLR, R-PCLR, and the selection approach of Qian et al. [11].

4.2 Method

We run BVS and lasso CLR in 50 parallel chains, i.e., both methods are applied to the data set 50 times, where a different random seed is used in each instance. To each chain, we apply (1) BVS CLR to obtain inclusion probability estimates $\hat{p}(\gamma_k = 1|\mathbf{X}, \mathbf{Z})$, and (2) lasso CLR to obtain coefficient estimates $\hat{\beta}_{k,las}$, where 10-fold cross-validation is used to estimate the tuning parameter λ in eq. (6). In applying BVS CLR, we use $S = 80,000$ iterations ($B = 40,000$ burn-in) for each chain. As in our simulation study, we set the variance σ^2 in our normal prior for $\beta_k|\gamma_k$ to 1 and assign a Beta$(5, 5)$ prior to ω. We retained these values due to the robustness of BVS CLR to the specification of σ^2 and the Beta(c, d) prior that we observed in preliminary analyses for our simulation study.

For lasso CLR, we identified features with nonzero coefficient estimates as relevant. To determine the threshold value for identifying relevant features in BVS CLR, we used the posterior inclusion probability estimates $\hat{p}(\gamma_k = 1|\mathbf{X}, \mathbf{Z})$ and applied the Bayesian FDR approach proposed by Muller et al. [50] and used in prior studies [51, 52]. This is a variation of the Benjamini and Hochberg [53] procedure where the threshold is based on increments in the ordered posterior probabilities rather than ordered p-values. We used an FDR level of 0.27 in order to identify approximately as many as features as in lasso CLR.

4.3 Results

In Table 2 and Table 3, we present the following, averaged across the 50 parallel chains: (1) BVS CLR inclusion probabilities for the features identified using the Bayesian FDR approach and their corresponding ranks, along with their coefficient estimates and standard deviations (SDs) and, (2) the lasso CLR coefficient estimates and their SDs for the features identified as relevant, along with their ranks based on the magnitude of their coefficient estimates.

There is 64% overlap among the features identified as relevant under BVS and lasso CLR. There is also a considerable degree of selection overlap using the approach of Qian et al. [11], which identifies as relevant 71% of the features selected by BVS CLR and 85% of those selected by lasso CLR. On the other hand, only 21% of the features selected by BVS CLR and 36% of the features selected by lasso CLR are also identified as relevant using R-PCLR. Relative to lasso CLR, the coefficient estimates under BVS CLR are noticeably larger in magnitude. Although the SDs for the coefficient estimates are smaller under lasso CLR, this may be due to the fact that the lasso CLR coefficient estimates are generally close to zero in magnitude. In addition, the multivariate scale reduction factors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ were 1.15 and 1.12, respectively. Since these values are not substantially greater than 1, these results do not indicate a lack of convergence.

Table 2 (MGH Imaging Study) BVS inclusion rates (and corresponding ranks), coefficient estimates and SDs for BVS CLR.

Code	Region/Volume	Rank	Inclusion Prob.	Coef. Est.	Coef. SD
v67[a,b]	Infarction volume $\geq 67^{th}$ percentile	1	0.86	0.90	0.13
j10[a]	Cerebral peduncle R	2	0.85	0.98	0.10

j34[a]	Fornix (cres) / Stria terminalis R	3	0.82	0.97	0.13
h74[a]	Temporal Fusiform Cortex – anterior division R	4	0.82	−1.12	0.14
h46[b]	Lateral Occipital Cortex – inferior division R	6	0.74	0.77	0.12
j23[b]	Posterior thalamic radiation L	8	0.67	−0.57	0.13
h32[a]	Inferior Temporal Gyrus – temporooccipital part R	10	0.67	0.63	0.11
h51[a]	Juxtapositional Lobule Cortex L	11	0.66	−0.64	0.12
h5[a]	Superior Frontal Gyrus L	13	0.64	−0.54	0.11
j19[a]	Superior corona radiata L	5	0.76	0.70	0.15
h72[a]	Lingual Gyrus R	7	0.73	0.70	0.10
cad[a]	Coronary artery disease	9	0.67	0.45	0.06
j30	Cingulum (cingulate gyrus) R	12	0.65	−0.55	0.09
h33	Postcentral Gyrus L	14	0.64	0.52	0.13

[a] Selected using Qian et al. [11] selection approach
[b] Selected using R-PCLR.

Table 3 (MGH Imaging study) Cross-validated coefficient estimates and SDs (and corresponding ranks for magnitude of coefficient estimates) for lasso CLR.

Code	Region/Volume	Rank	Coef. Est.	Coef. SD
v67[a,b]	Infarction volume $\geq 67^{th}$ percentile	4	0.18	0.02
j10[a]	Cerebral peduncle R	6	0.12	0.02
j34[a]	Fornix (cres) / Stria terminalis R	2	0.27	0.01
h74[a]	Temporal Fusiform Cortex – anterior division R	10	−0.01	0.02
h46[b]	Lateral Occipital Cortex – inferior division R	11	<0.01	0.01
j23[b]	Posterior thalamic radiation L	14	<−0.01	<0.01
h32[a]	Inferior Temporal Gyrus – temporooccipital part R	1	0.42	0.01
h51[a]	Juxtapositional Lobule Cortex L	3	−0.26	0.02
h5[a]	Superior Frontal Gyrus L	8	−0.07	<0.01
h8[a,b]	Middle Frontal Gyrus R	5	0.13	0.01
j26[a,b]	Sagittal stratum R	7	0.08	<0.01
v33[a]	Infarction volume $\geq 33^{rd}$ percentile	9	0.01	0.02
h45[a]	Lateral Occipital Cortex – inferior division L	12	<-0.01	<0.01
h25[a]	Middle Temporal Gyrus – temporooccipital part L	13	<0.01	<0.01

[a] Selected using Qian et al. [11] selection approach.
[b] Selected using R-PCLR.

Figure 2 and 4.4 display heatmaps of Cramèr's V matrix plot for the pairwise differences among the features selected in Table 2 and Table 3 as relevant and those not selected, where the pairwise feature differences are considered to account for matching. In these plots, black denotes perfect positive/negative correlation, white denotes no correlation, and light/dark gray denotes weak/strong correlation. In both plots, we see that the correlation among features is generally low, so that the structure of this dataset most closely corresponds to Scenario 2 ($K = 100$ features) in our simulation study when all binary features were weakly correlated.

Most of the right hemispheric brain regions identified under BVS CLR were found to be associated with HAP in Kemmling et al. [43] , who showed that infarction in the right hemispheric peri-insular cortical regions was associated with the risk of acquiring HAP in acute ischemic stroke patients. As discussed in Section Section 4.1, one explanation Kemmling et al.provided for this finding was the association of the right insular region with autonomically-induced immunosuppression and susceptibility to infection shown in Cechetto and Chen [46], Sander and Klingelhöfer [47], and Meyer et al. [48], as well as the association of right hemispheric peri-insular infarction to autonomic dysfunction and pathologic sympathetic activity shown in Colivicchi et al. [49].

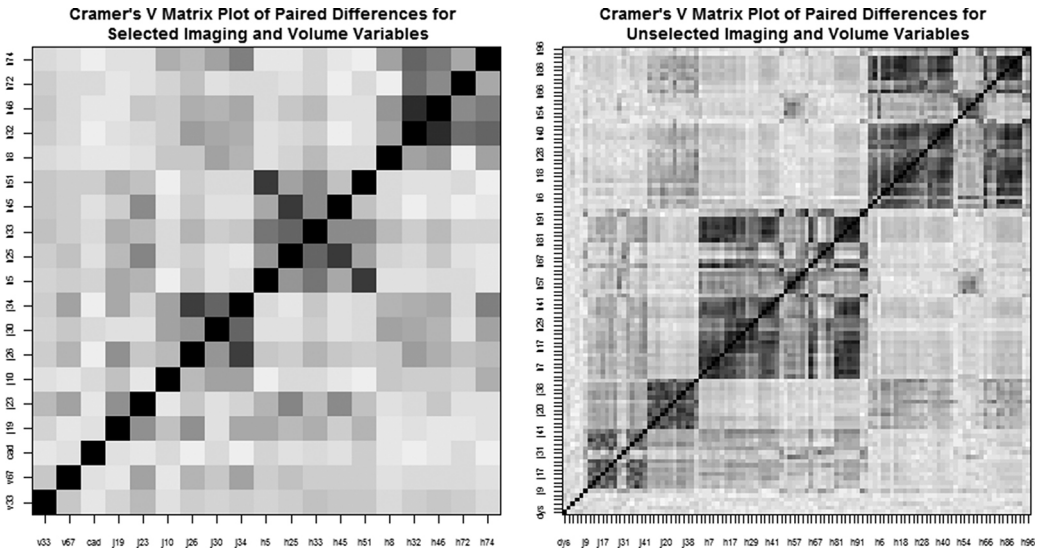


Figure 2: (MGH Imaging Study) Cramer's V matrix plots of paired differences for features selected (a) and not selected (b) in Table 2 and Table 3 (black denotes perfect positive or negative correlation, white denotes no correlation; light/dark gray denotes weak/strong correlation).

5 Application to Cardiovascular Disease Biomarker Study

5.1 Description

The cardiovascular disease biomarker study we examine was a matched case-control study conducted by the High Risk Plaque Initiative [BG Medicine Inc.(Waltham, MA) and other partners] to discover prognostic biomarkers in blood plasma for near-term cardiovascular events. Subjects from the CATHGEN study were selected for this investigation. The CATHGEN project collected peripheral blood samples from consenting research subjects undergoing cardiac catheterization at Duke University Medical Center from 2001 to 2011. 68 cases were selected from among individuals who had a major adverse cardiac event (MACE) within two years following the time of their sample collection. In a 1:1 matched study design, 68 controls were selected from individuals who were MACE-free for the two years following sample collection and were matched to cases on age, gender, race/ethnicity and severity of coronary artery disease. High-content mass spectrometry and multiplexed immunoassay-based techniques were employed to quantify 625 proteins and metabolites from each subject's serum specimen. Comprehensive metabolite profiling of the individual samples was based on a combination of four platforms employing mass spectrometry (MS) based techniques to profile lipids, fatty acids, amino acids, sugars and other metabolites. Proteomic analysis was based on a combination of targeted methods using a quantitative multiplexed immunoassay technique as well as a comprehensive protein profiling strategy based on tandem mass spectrometry. A detailed description of the mass spectrometry based platforms and proteomics analysis can be found in a previous publication [54]. In our analysis, the identities of the measured metabolites and proteins are masked due to a data confidentiality agreement with the stakeholders involved in the study.

We consider for analysis the 593 biomarkers with complete data. In this application, we evaluate BVS CLR under the assumption of normality for all biomarkers, and assess its performance in relation to lasso CLR and R-PCLR.

5.2 Method

We used the same procedure described in Section Section 4.2, except that 25 parallel chains with different random seeds are run. To account for the high dimensionality of this dataset, we use $S = 150,000$ iterations ($B = 75,000$ burn-in) for each chain. As in our previous application, we set the variance σ^2 in our normal prior for $\beta_k|\gamma_k$ to 1 and assign a Beta$(5,5)$ prior to ω. To identify relevant features, we use the Bayesian FDR approach (at an FDR level of 0.41) for BVS CLR based on the inclusion probability estimates and examine features whose lasso CLR coefficient estimates are nonzero.

5.3 Results

In Table 4, we present the estimates of the BVS CLR inclusion probabilities across the 25 chains for the biomarkers identified using the Bayesian FDR approach and their corresponding ranks, along with the corresponding coefficient estimates and their SDs. We also report the lasso CLR coefficient estimates and their SDs across the chains for the biomarkers identified as relevant under lasso CLR, along with their ranks based on the magnitude of their coefficient estimates.

There is 52% overlap among the biomarkers identified as relevant under BVS and lasso CLR, and only 25% overlap among the biomarkers identified as relevant under BVS CLR and R-PCLR. Relative to lasso CLR, the coefficient estimates under BVS CLR are larger in magnitude, along with their SDs, which may be due to the fact that the lasso CLR coefficient estimates are generally close to zero.

In Figure 3 and Figure 4, we display heatmaps of the correlation matrix plot for the pairwise differences among the biomarkers selected in Table 4 as relevant and those not selected, where we see that the correlation among biomarkers is generally low, so that the structure of this dataset most closely resembles Scenario 3 ($K = 600$ features) in our simulation study when all normal features were weakly correlated. The fact that we saw lower MSE in our application of BVS CLR to these normal datasets suggests that the coefficient estimates under BVS CLR are more likely to be more accurate relative to lasso CLR.

We note that the multivariate scale reduction factors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ were 3.01 and 2.95, respectively. Although we do not report the results, increasing the number of iterations and folds to a reasonable degree, considering the computational expense involved, did not substantially decrease these reduction factor values. This may result from the high number of biomarkers examined, relative to the number of subjects.

Table 4 (CVS Disease Biomarker Study) BVS CLR inclusion rates and corresponding ranks; BVS and lasso CLR CV coefficient estimates (and SDs).

BVS CLR					**Lasso CLR**			
Biomarker	**Rank**	**Inclusion Prob.**	**Coef. Est.**	**Coef. SD**	**Biomarker**	**Rank**	**Coef. Est.**	**Coef. SD**
V595	1	0.68	−0.76	0.15	V595	3	−0.27	0.05
V272	2	0.67	0.78	0.18	V272	5	0.18	0.06
V582	3	0.64	−0.66	0.15	V582	9	−0.13	0.01
V275[a]	4	0.63	0.64	0.19	V275[a]	12	0.10	0.01
V617[a]	5	0.63	−0.63	0.28	V617[a]	2	−0.41	0.05
V164	6	0.62	−0.58	0.13	V164	4	−0.19	0.02
V625[a]	7	0.61	0.63	0.18	V625[a]	1	0.54	0.06
V174	8	0.61	0.52	0.17	V174	6	0.17	0.03
V593	9	0.61	0.56	0.22	V593	10	0.11	0.03
V436	10	0.60	−0.51	0.25	V436	26	−0.01	0.01
V362	11	0.59	0.51	0.16	V362	13	0.09	0.04
V607	12	0.58	0.46	0.14	V607	7	0.16	0.04
V535	15	0.58	0.45	0.18	V535	18	0.06	0.03
V75	23	0.57	−0.44	0.20	V75	15	−0.08	0.03
V31	26	0.56	0.49	0.29	V31	28	<0.01	0.01
V464	30	0.56	−0.37	0.22	V464	8	−0.15	0.05
V158	13	0.58	−0.35	0.20	V185	11	0.10	0.01
V219	14	0.58	0.43	0.17	V66	14	0.08	0.01
V620	16	0.58	−0.34	0.18	V314	16	0.08	0.03
V24	17	0.58	0.46	0.17	V122	17	0.08	0.01
V151	18	0.57	−0.45	0.20	V605a	19	0.03	0.01
V96	19	0.57	0.41	0.13	V41	20	0.02	0.01
V223	20	0.57	0.22	0.23	V611[a]	21	0.02	0.02
V234	21	0.57	0.28	0.18	V563	22	0.01	0.01
V227	22	0.57	0.30	0.17	V495	23	0.01	0.01
V152	24	0.56	−0.38	0.17	V459	24	−0.01	0.02
V99	25	0.56	−0.44	0.20	V64	25	0.01	0.01
V54	27	0.56	0.47	0.14	V621	27	0.01	0.01
V89	28	0.56	−0.34	0.13	V128	29	<0.01	0.01
V228	29	0.56	0.36	0.14	V117	30	<0.01	0.01
V578	31	0.56	0.38	0.13	V597[a]	31	<0.01	0.01

[a] Note:Selected using R-PCLR.

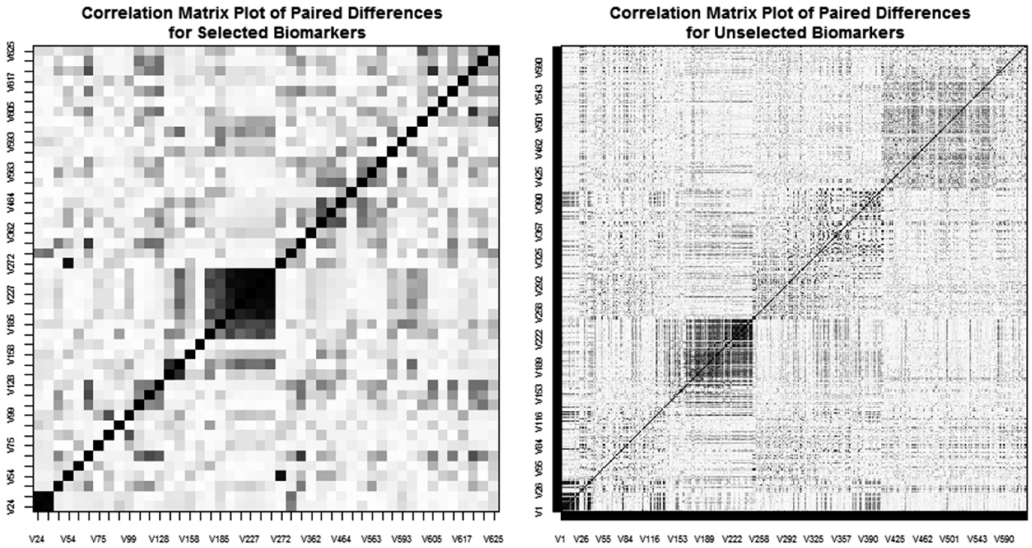


Figure 3: (CVS Disease Biomarker Study) Correlation matrix plots of paired differences for biomarkers selected (a) and not selected (b) in Table 4 (black denotes perfect positive or negative correlation, white denotes no correlation; light/dark gray denotes weak/strong correlation).

6 Discussion

In examining the simulation study results for BVS CLR in Section Section 3, we see that selection and prediction accuracy increase as the number of pairs I increases, while MSE (for the relevant features) decreases. Both selection accuracy and MSE increase with decreasing K and Q, while selection accuracy also decreases as the correlation between features increases. Prediction accuracy is higher and MSE is lower for normal datasets, while prediction accuracy is higher for datasets where the numbers of total and relevant features are both moderate relative to the other simulation scenarios considered, i.e., $K = 100$ where 25% of features are relevant. Relative to lasso CLR, BVS CLR generally has lower MSE. In addition, BVS CLR most often has higher selection accuracy when the total number of features is moderate ($K = 100$) and higher prediction accuracy when we have at least a moderate number of features, which is more pronounced in datasets with fewer pairs. We note that the increase in selection accuracy of BVS CLR, relative to lasso CLR, is especially pronounced when the β values for the relevant features is relatively small (i.e., less than 1). Based on the study results in Section Section 4 and Section Section 5, the degree of overlap in features identified as relevant is higher for the binary MGH imaging data.

In cases where selection accuracy is solely of interest, we can modify BVS CLR using the data augmentation approach in Sha et al. [24] based on the probit approximation to the logit link $[1 + exp(-\nu)]^{-1} \approx \Phi(\nu/1.7)$ [57]. This formulation allows for the integration of $\boldsymbol{\beta}$ out of the likelihood function, so that only the posterior inclusion probabilities are estimated. When applied to high dimensional datasets, this approximation approach can increase computational efficiency and improve mixing of the MCMC sampler [24]. Although we do not report its results, we also consider the probit approximation $p_{i1} = \Phi\left(\frac{1}{1.7}\sum_{k=1}^{K}\beta_k(X_{i1,k} - X_{i2,k})\right)$ to the case probability in eq. (1) when applying our proposed BVS approach and obtain nearly identical results compared with BVS CLR.

Possible future research directions include an extension of BVS CLR to account for interactions among the examined features. This can be done for two-way interactions by modifying our selection approach based on the methodology developed by Chipman [55]. Extensions of BVS CLR that incorporate specific correlation or grouping structures among different features, including extensions of the approaches of Smith and Fahrmeir [28] and Stingo et al. [29] to incorporate the spatial correlation and the correlation among features known to be involved in similar biological functions, can be considered for matched case-control data.

Although our selection methodology focuses on 1:1 case-control matching, we can extend our formulated models for BVS CLR to handle the more general case of 1:n case-control matching. Another extension would account for matching across multiple groups when an intrinsic ordering exists among the case and control groups, e.g., matching individuals across different disease severity levels. We will further investigate how to do so using the conditional adjacent categories logistic modeling approach developed by Mukherjee et al. [56] for matched 1:n case-control studies. Through appropriate transformations of the features and notational ex-

pansion of the matched sets, we can re-frame the modeling approach of Mukherjee et al. [56] for BVS CLR to handle ordinal-based matching.

Funding

This work was supported by the National Institutes of Health (Grant / Award Number: T32NS048005; Grant / Award Number: F32NS081904). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix

A Simulation Study Tables

Table 5 (Simulation Study: Selection Accuracy) Summary measures for AUCs obtained from BVS inclusion probabilities and magnitudes of lasso CLR coefficient estimates for $K = 20$ and $K = 100$ features ($\beta_1, \ldots, \beta_Q$ at least 1 in magnitude).

Data type	Correlation level	Number of pairs	Scenario 1 20 features Median(IQR) BVS CLR	Lasso CLR	Scenario 2 100 features Median(IQR) BVS CLR	Lasso CLR
Normal[a]	Low	50	1(<0.01)	1(<0.01)	0.89(0.1)	0.7(0.19)
		200	1(<0.01)	1(<0.01)	1(<0.01)	1(<0.01)
	High	50	1(<0.01)	1(0.08)	0.77(0.12)	0.61(0.07)
		200	1(<0.01)	1(<0.01)	0.87(0.07)	0.68(0.07)
Binary[a]	Low	50	0.94(0.12)	0.92(0.41)	0.81(0.1)	0.69(0.13)
		200	1(<0.01)	1(<0.01)	0.99(0.02)	0.99(0.01)
	High	50	0.94(0.14)	0.72(0.5)	0.78(0.11)	0.68(0.11)

		200	1(<0.01)	1(<0.01)	0.98(0.02)	0.99(0.03)
Normal[b]	Low	50	1(0.04)	0.99(0.04)	0.75(0.08)	0.58(0.12)
		200	1(<0.01)	1(<0.01)	0.98(0.02)	0.98(0.03)
	High	50	0.95(0.07)	0.73(0.07)	0.69(0.08)	0.54(0.05)
		200	1(0.03)	0.99(0.14)	0.84(0.06)	0.57(0.04)
Binary[b]	Low	50	0.91(0.12)	0.87(0.16)	0.73(0.07)	0.64(0.1)
		200	1(<0.01)	1(<0.01)	0.96(0.03)	0.95(0.04)
	High	50	0.92(0.11)	0.86(0.16)	0.75(0.08)	0.63(0.07)
		200	1(<0.01)	1(<0.01)	0.91(0.04)	0.89(0.05)
Normal[c]	Low	50	0.93(0.08)	0.88(0.12)	0.64(0.06)	0.5(0.01)
		200	1(<0.01)	1(<0.01)	0.9(0.04)	0.86(0.05)
	High	50	0.96(0.07)	0.68(0.09)	0.67(0.08)	0.51(0.05)
		200	0.98(0.05)	0.68(0.07)	0.84(0.07)	0.58(0.04)
Binary[c]	Low	50	0.87(0.13)	0.78(0.16)	0.66(0.07)	0.55(0.06)
		200	1(0.02)	1(0.01)	0.89(0.04)	0.84(0.05)
	High	50	0.84(0.12)	0.74(0.14)	0.67(0.08)	0.55(0.05)
		200	0.99(0.03)	0.99(0.03)	0.8(0.07)	0.77(0.06)

[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$);
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$);
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$).

Table 6 (Simulation Study: Selection Accuracy) Summary measures for AUCs obtained from BVS inclusion probabilities and magnitudes of lasso CLR coefficient estimates for $K = 600$ features ($\beta_1, \ldots, \beta_Q$ at least 1 in magnitude).

Data type	Correlation level	Number of pairs	Scenario 3 600 features Median(IQR) BVS CLR	Lasso CLR
Normal[a]	Low	50	0.52(0.05)	0.5(0.01)
		200	0.61(0.05)	0.65(0.06)
	High	50	0.57(0.05)	0.53(0.02)
		200	0.63(0.04)	0.57(0.02)
Binary[a]	Low	50	0.53(0.03)	0.51(0.02)
		200	0.69(0.04)	0.71(0.03)
	High	50	0.54(0.06)	0.52(0.02)
		200	0.66(0.04)	0.67(0.03)
Normal[b]	Low	50	0.5(0.03)	0.5(0.01)
		200	0.57(0.03)	0.56(0.03)
	High	50	0.53(0.06)	0.52(0.02)
		200	0.59(0.03)	0.57(0.01)
Binary[b]	Low	50	0.51(0.03)	0.5(0.01)
		200	0.61(0.05)	0.58(0.02)
	High	50	0.52(0.05)	0.51(0.01)
		200	0.62(0.04)	0.6(0.03)
Normal[c]	Low	50	0.51(0.04)	0.5(0.01)
		200	0.54(0.03)	0.51(0.02)
	High	50	0.52(0.04)	0.51(0.01)
		200	0.55(0.03)	0.55(0.01)
Binary[c]	Low	50	0.51(0.03)	0.5(<0.01)
		200	0.57(0.03)	0.54(0.01)
	High	50	0.54(0.04)	0.53(0.01)
		200	0.61(0.02)	0.6(0.01)

[a] 10% relevant ($Q = 60$).
[b] 25% relevant ($Q = 150$).
[c] 50% relevant ($Q = 300$).

Table 7 (Simulation Study: Selection Accuracy) Summary measures for AUCs obtained from BVS inclusion probabilities and magnitudes of lasso CLR coefficient estimates for $K = 20$ and $K = 100$ features ($\beta_1, \ldots, \beta_Q$ between 0.3 and 0.7 in magnitude).

	Scenario 1	Scenario 2

Data type	Correlation level	Number of pairs	20 features Median(IQR) BVS CLR	Lasso CLR	100 features Median(IQR) BVS CLR	Lasso CLR
Normal[a]	Low	50	0.83(0.25)	0.5(0.21)	0.75(0.09)	0.53(0.13)
		200	1(<0.01)	1(<0.01)	0.94(0.08)	0.94(0.07)
	High	50	0.89(0.22)	0.5(0.05)	0.68(0.14)	0.53(0.08)
		200	1(<0.01)	1(<0.01)	0.8(0.1)	0.66(0.08)
Binary[a]	Low	50	0.64(0.28)	0.5(<0.01)	0.62(0.11)	0.5(0.04)
		200	0.88(0.19)	0.73(0.3)	0.8(0.08)	0.73(0.14)
	High	50	0.69(0.28)	0.5(<0.01)	0.61(0.12)	0.5(0.05)
		200	0.83(0.21)	0.71(0.28)	0.82(0.08)	0.74(0.12)
Normal[b]	Low	50	0.88(0.14)	0.85(0.17)	0.7(0.07)	0.51(0.07)
		200	1(<0.01)	1(<0.01)	0.9(0.05)	0.89(0.07)
	High	50	0.83(0.12)	0.6(0.12)	0.71(0.07)	0.55(0.04)
		200	0.93(0.08)	0.68(0.11)	0.83(0.05)	0.62(0.05)
Binary[b]	Low	50	0.67(0.21)	0.5(0.03)	0.59(0.09)	0.51(0.04)
		200	0.84(0.11)	0.68(0.17)	0.78(0.09)	0.74(0.07)
	High	50	0.7(0.16)	0.5(0.05)	0.63(0.09)	0.52(0.07)
		200	0.89(0.09)	0.83(0.14)	0.75(0.1)	0.71(0.07)
Normal[c]	Low	50	0.82(0.13)	0.77(0.16)	0.63(0.07)	0.5(0.01)
		200	0.99(0.03)	0.99(0.03)	0.84(0.05)	0.81(0.05)
	High	50	0.85(0.12)	0.52(0.1)	0.63(0.08)	0.49(0.03)
		200	0.96(0.08)	0.67(0.08)	0.72(0.07)	0.49(0.05)
Binary[c]	Low	50	0.71(0.15)	0.5(0.11)	0.61(0.07)	0.51(0.04)
		200	0.88(0.1)	0.82(0.16)	0.76(0.06)	0.71(0.05)
	High	50	0.7(0.13)	0.5(0.07)	0.6(0.06)	0.51(0.04)
		200	0.88(0.08)	0.82(0.11)	0.64(0.12)	0.62(0.05)

[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$).
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$).
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$).

Table 8 (Simulation Study: Prediction Accuracy) Summary measures for AUCs obtained from true case-control status and BVS CLR/lasso CLR predicted case probabilities in independent test sets for $K = 20$ and $K = 100$ features ($\beta_1, \ldots, \beta_Q$ at least 1 in magnitude).

Data type	Correlation level	Number of pairs	Scenario 1 20 features Median(IQR) BVS CLR	Lasso CLR	Scenario 2 100 features Median(IQR) BVS CLR	Lasso CLR
Normal[a]	Low	50	0.66(0.13)	0.65(0.15)	0.71(0.13)	0.65(0.24)
		200	0.67(0.08)	0.66(0.09)	0.79(0.07)	0.79(0.1)
	High	50	0.6(0.14)	0.55(0.15)	0.61(0.14)	0.7(0.13)
		200	0.64(0.08)	0.63(0.09)	0.72(0.08)	0.73(0.07)
Binary[a]	Low	50	0.55(0.09)	0.51(0.07)	0.6(0.1)	0.61(0.16)
		200	0.53(0.06)	0.53(0.06)	0.69(0.08)	0.69(0.09)
	High	50	0.53(0.11)	0.5(0.05)	0.63(0.08)	0.66(0.13)
		200	0.51(0.06)	0.51(0.05)	0.69(0.07)	0.69(0.07)
Normal[b]	Low	50	0.87(0.13)	0.89(0.1)	0.79(0.16)	0.64(0.29)
		200	0.88(0.04)	0.88(0.05)	0.89(0.11)	0.91(0.08)
	High	50	0.89(0.04)	0.9(0.03)	0.91(0.02)	0.91(0.02)
		200	0.9(0.02)	0.9(0.02)	0.94(0.01)	0.94(0.01)
Binary[b]	Low	50	0.72(0.11)	0.72(0.1)	0.72(0.09)	0.69(0.15)
		200	0.73(0.06)	0.74(0.06)	0.83(0.05)	0.84(0.06)
	High	50	0.74(0.1)	0.74(0.1)	0.68(0.1)	0.63(0.12)
		200	0.75(0.03)	0.75(0.04)	0.8(0.05)	0.79(0.06)
Normal[c]	Low	50	0.77(0.16)	0.76(0.19)	0.76(0.09)	0.5(0.06)
		200	0.82(0.08)	0.83(0.09)	0.87(0.11)	0.86(0.1)
	High	50	0.79(0.08)	0.78(0.12)	0.87(0.03)	0.79(0.04)

		200	0.83(0.05)	0.84(0.05)	0.91(0.02)	0.86(0.03)
Binary[c]	Low	50	0.69(0.1)	0.67(0.17)	0.69(0.07)	0.58(0.12)
		200	0.71(0.06)	0.71(0.07)	0.78(0.06)	0.77(0.08)
	High	50	0.63(0.1)	0.61(0.14)	0.66(0.09)	0.58(0.1)
		200	0.63(0.06)	0.62(0.06)	0.78(0.05)	0.78(0.06)

[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$).
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$);
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$).

Table 9 (Simulation Study: Prediction Accuracy) Summary measures for AUCs obtained from true case-control status and BVS CLR/lasso CLR predicted case probabilities in independent test sets for $K = 600$ features ($\beta_1, \ldots, \beta_Q$ at least 1 in magnitude).

Data type	Correlation level	Number of pairs	Scenario 3 600 features Median(IQR) BVS CLR	Lasso CLR
Normal[a]	Low	50	0.66(0.03)	0.55(0.07)
		200	0.79(0.02)	0.76(0.02)
	High	50	0.77(0.07)	0.75(0.03)
		200	0.84(0.04)	0.81(0.03)
Binary[a]	Low	50	0.56(0.03)	0.52(0.04)
		200	0.64(0.03)	0.66(0.08)
	High	50	0.61(0.03)	0.58(0.03)
		200	0.67(0.04)	0.68(0.03)
Normal[b]	Low	50	0.67(0.04)	0.57(0.06)
		200	0.81(0.02)	0.75(0.04)
	High	50	0.86(0.06)	0.69(0.07)
		200	0.9(0.03)	0.81(0.04)
Binary[b]	Low	50	0.57(0.06)	0.51(0.09)
		200	0.63(0.05)	0.6(0.1)
	High	50	0.68(0.06)	0.6(0.07)
		200	0.78(0.02)	0.76(0.03)
Normal[c]	Low	50	0.66(0.04)	0.57(0.07)
		200	0.79(0.03)	0.75(0.03)
	High	50	0.91(0.03)	0.8(0.05)
		200	0.96(0.01)	0.88(0.04)
Binary[c]	Low	50	0.59(0.05)	0.5(0.04)
		200	0.73(0.04)	0.67(0.04)
	High	50	0.94(0.03)	0.81(0.04)
		200	0.97(0.01)	0.92(0.01)

[a] 10% relevant ($Q = 60$).
[b] 25% relevant ($Q = 150$)
[c] 50% relevant ($Q = 300$).

Table 10 (Simulation Study: Prediction Accuracy) Summary measures for AUCs obtained from true case-control status and BVS CLR/lasso CLR predicted case probabilities in independent test sets for $K = 20$ and $K = 100$ features ($\beta_1, \ldots, \beta_Q$ between 0.3 and 0.7 in magnitude).

Data type	Correlation level	Number of pairs	Scenario 1 20 features Median(IQR) BVS CLR	Lasso CLR	Scenario 2 100 features Median(IQR) BVS CLR	Lasso CLR
Normal[a]	Low	50	0.45(0.19)	0.5(0.02)	0.54(0.17)	0.5(0.08)
		200	0.48(0.14)	0.46(0.15)	0.58(0.15)	0.56(0.16)
	High	50	0.49(0.24)	0.5(<0.01)	0.45(0.1)	0.48(0.12)
		200	0.49(0.1)	0.46(0.14)	0.52(0.1)	0.48(0.09)
Binary[a]	Low	50	0.5(0.11)	0.5(<0.01)	0.52(0.09)	0.5(0.01)
		200	0.47(0.13)	0.48(0.13)	0.52(0.1)	0.54(0.16)

	High	50	0.49(0.11)	0.5(<0.01)	0.52(0.1)	0.5(0.01)
		200	0.47(0.12)	0.48(0.11)	0.5(0.1)	0.53(0.16)
Normal[b]	Low	50	0.79(0.17)	0.8(0.25)	0.65(0.19)	0.5(0.13)
		200	0.85(0.05)	0.86(0.05)	0.68(0.19)	0.67(0.22)
	High	50	0.66(0.21)	0.71(0.3)	0.87(0.04)	0.87(0.04)
		200	0.71(0.1)	0.74(0.09)	0.88(0.03)	0.89(0.02)
Binary[b]	Low	50	0.51(0.14)	0.5(<0.01)	0.56(0.12)	0.5(0.08)
		200	0.52(0.14)	0.52(0.17)	0.58(0.09)	0.58(0.15)
	High	50	0.59(0.14)	0.5(0.06)	0.54(0.1)	0.5(0.09)
		200	0.64(0.11)	0.65(0.11)	0.57(0.09)	0.53(0.1)
Normal[c]	Low	50	0.65(0.23)	0.65(0.28)	0.71(0.17)	0.5(0.14)
		200	0.63(0.18)	0.64(0.18)	0.77(0.18)	0.75(0.23)
	High	50	0.62(0.18)	0.5(0.15)	0.9(0.02)	0.89(0.02)
		200	0.64(0.1)	0.6(0.1)	0.92(0.01)	0.91(0.01)
Binary[c]	Low	50	0.62(0.15)	0.5(0.18)	0.58(0.12)	0.5(0.13)
		200	0.66(0.09)	0.66(0.1)	0.63(0.11)	0.6(0.16)
	High	50	0.6(0.13)	0.5(0.14)	0.5(0.08)	0.46(0.09)
		200	0.65(0.1)	0.66(0.09)	0.55(0.07)	0.51(0.08)

[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$).
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$).
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$).

Table 11 (Simulation Study: MSE for relevant predictors) Summary measures for BVS CLR/lasso CLR MSE estimates for sets of relevant predictors computed across predictors ($\beta_1, \ldots, \beta_Q$ at least 1 in magnitude).

Data type	Correlation level	Relevant	Number of pairs	Scenario 1 20 features Median BVS CLR	Lasso CLR	Scenario 2 100 features Median BVS CLR	Lasso CLR	Scenario 3 600 features Median BVS CLR	Lasso CLR
Normal[a]	Low	Yes	50	0.29	0.22	0.46	1.02	1.15	1.30
			200	0.06	0.07	0.47	0.31	0.81	1.26
	High	Yes	50	0.19	0.53	0.78	1.11	1.17	1.28
			200	0.05	0.10	0.75	0.96	1.03	1.24
Binary[a]	Low	Yes	50	0.56	0.84	1.01	1.40	1.92	2.16
			200	0.09	0.17	0.29	0.40	1.34	1.83
	High	Yes	50	0.65	0.99	0.84	1.11	2.08	2.13
			200	0.10	0.18	0.32	0.51	1.49	1.85
Normal[b]	Low	Yes	50	0.27	0.53	0.83	1.28	1.26	1.34
			200	0.07	0.12	0.27	0.70	1.07	1.30
	High	Yes	50	0.39	0.60	1.09	1.25	1.17	1.31
			200	0.20	0.26	1.10	1.18	1	1.28
Binary[b]	Low	Yes	50	0.73	0.76	1.68	1.97	2.13	2.25
			200	0.13	0.18	0.64	0.78	1.80	2.12
	High	Yes	50	1.11	1.22	1.83	2.08	2.07	2.23
			200	0.14	0.26	0.63	0.78	1.70	2.05
Normal[c]	Low	Yes	50	0.47	0.76	0.92	1.24	1.25	1.31
			200	0.07	0.26	0.53	0.92	1.12	1.30
	High	Yes	50	0.56	1.03	1.07	1.29	1.18	1.31
			200	0.40	0.88	0.96	1.19	1.07	1.30
Binary[c]	Low	Yes	50	0.97	0.90	1.94	2.17	2.20	2.24
			200	0.26	0.21	1.20	1.24	1.90	2.20
	High	Yes	50	0.74	0.80	1.71	1.90	2.04	2.18
			200	0.26	0.19	1.04	1.13	1.88	2.10

[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$; $Q = 60$ for $K = 600$).
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$; $Q = 150$ for $K = 600$).
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$; $Q = 300$ for $K = 600$).

Table 12 (Simulation Study: MSE for relevant predictors) Summary measures for BVS CLR/lasso CLR MSE estimates for sets of relevant predictors computed across predictors ($\beta_1, \ldots, \beta_Q$ between 0.3 and 0.7 in magnitude).

Data type	Correlation level	Relevant	Number of pairs	Scenario 1 20 features Median BVS CLR	Lasso CLR	Scenario 2 100 features Median BVS CLR	Lasso CLR
Normal[a]	Low	Yes	50	0.16	0.14	0.18	0.19
			200	0.03	0.03	0.72	0.06
	High	Yes	50	0.18	0.27	0.18	0.17
			200	0.02	0.06	0.21	0.14
Binary[a]	Low	Yes	50	0.15	0.18	0.15	0.20
			200	0.10	0.11	0.23	0.12
	High	Yes	50	0.19	0.22	0.17	0.20
			200	0.11	0.12	0.22	0.12
Normal[b]	Low	Yes	50	0.25	0.14	0.15	0.25
			200	0.04	0.03	0.46	0.09
	High	Yes	50	0.14	0.16	0.18	0.20
			200	0.10	0.12	0.30	0.18
Binary[b]	Low	Yes	50	0.16	0.15	0.16	0.21
			200	0.10	0.11	0.24	0.13
	High	Yes	50	0.22	0.27	0.17	0.21
			200	0.14	0.13	0.22	0.13
Normal[c]	Low	Yes	50	0.20	0.13	0.20	0.29
			200	0.07	0.03	0.24	0.15
	High	Yes	50	0.19	0.20	0.18	0.24
			200	0.13	0.07	0.16	0.22
Binary[c]	Low	Yes	50	0.19	0.23	0.20	0.25
			200	0.13	0.10	0.20	0.14
	High	Yes	50	0.25	0.31	0.21	0.24
			200	0.14	0.11	0.20	0.19

[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$).
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$);
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$).

Table 13 (Simulation Study: MSE for non-relevant predictors) Summary measures for BVS CLR/lasso CLR MSE estimates for sets of non-relevant predictors computed across predictors ($\beta_1, \ldots, \beta_Q$ at least 1 in magnitude).

Data type	Correlation level	Relevant	Number of pairs	Scenario 1 20 features Median BVS CLR	Lasso CLR	Scenario 2 100 features Median BVS CLR	Lasso CLR	Scenario 3 600 features Median BVS CLR	Lasso CLR
Normal[c]	Low	No	50	0.03	0.01	0.02	<0.01	0.02	<0.01
			200	<0.01	<0.01	0.01	<0.01	0.04	<0.01
	High	No	50	0.03	0.01	0.03	<0.01	0.01	<0.01
			200	<0.01	0.01	0.03	<0.01	0.01	<0.01
Binary[a]	Low	No	50	0.04	0.02	0.04	0.01	0.01	<0.01
			200	0.01	0.01	0.03	0.01	0.03	<0.01
	High	No	50	0.04	0.03	0.04	0.01	0.01	<0.01
			200	0.01	0.01	0.03	0.01	0.03	<0.01
Normal[b]	Low	No	50	0.02	0.01	0.02	<0.01	0.01	<0.01
			200	0.01	0.01	0.01	<0.01	0.04	<0.01
	High	No	50	0.02	0.01	0.01	<0.01	0.01	<0.01
			200	<0.01	0.01	0.01	<0.01	0.01	<0.01
Binary[b]	Low	No	50	0.04	0.06	0.02	0.01	0.01	<0.01
			200	0.01	0.02	0.02	0.02	0.03	<0.01
	High	No	50	0.03	0.05	0.04	0.01	0.01	<0.01
			200	0.01	0.03	0.01	0.02	0.02	<0.01
Normal[c]	Low	No	50	0.02	0.01	0.02	<0.01	0.02	<0.01
			200	<0.01	0.01	0.01	<0.01	0.04	<0.01

	High	No	50	0.01	0.01	0.01	<0.01	<0.01	<0.01
			200	<0.01	0.01	0.01	<0.01	0.01	<0.01
Binary[c]	Low	No	50	0.03	0.06	0.03	0.01	0.02	<0.01
			200	0.01	0.04	0.01	0.02	0.03	<0.01
	High	No	50	0.04	0.07	0.04	0.01	0.01	<0.01
			200	0.01	0.04	0.01	0.02	0.01	<0.01

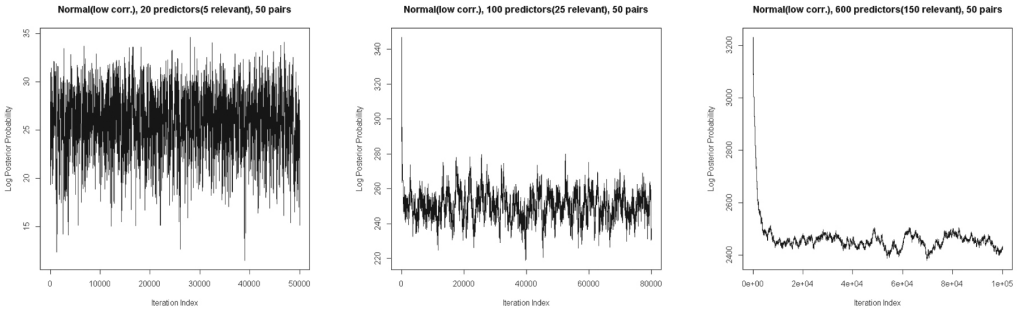
[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$; $Q = 60$ for $K = 600$).
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$; $Q = 150$ for $K = 600$);
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$; $Q = 300$ for $K = 600$).

Table 14 (Simulation Study: MSE for non-relevant predictors) Summary measures for BVS CLR/lasso CLR MSE estimates for sets of relevant predictors computed across predictors ($\beta_1, \ldots, \beta_Q$ between 0.3 and 0.7 in magnitude).

				Scenario 1 20 features Median		**Scenario 2 100 features Median**	
Data type	**Correaltion level**	**Relevant**	**Number of pairs**	**BVS CLR**	**Lasso CLR**	**BVS CLR**	**Lasso CLR**
Normal[a]	Low	Yes	50	0.03	<0.01	0.04	<0.01
			200	<0.01	<0.01	0.07	<0.01
	High	Yes	50	0.04	<0.01	0.05	<0.01
			200	<0.01	<0.01	0.04	<0.01
Binary[a]	Low	Yes	50	0.05	0.01	0.05	<0.01
			200	0.01	0.01	0.04	<0.01
	High	Yes	50	0.05	0.01	0.05	<0.01
			200	0.01	<0.01	0.04	<0.01
Normal[b]	Low	Yes	50	0.04	0.01	0.03	<0.01
			200	<0.01	0.01	0.04	<0.01
	High	Yes	50	0.04	0.01	0.02	<0.01
			200	<0.01	0.01	0.04	<0.01
Binary[b]	Low	Yes	50	0.05	0.01	0.05	<0.01
			200	0.01	0.01	0.05	0.01
	High	Yes	50	0.05	0.01	0.05	0.01
			200	0.01	0.01	0.04	0.01
Normal[c]	Low	Yes	50	0.04	0.02	0.03	<0.01
			200	<0.01	0.01	0.03	0.01
	High	Yes	50	0.03	0.01	0.01	<0.01
			200	<0.01	0.01	0.02	0.01
Binary[c]	Low	Yes	50	0.04	0.02	0.04	0.01
			200	0.01	0.02	0.04	0.01
	High	Yes	50	0.04	0.02	0.04	0.01
			200	0.01	0.02	0.04	0.01

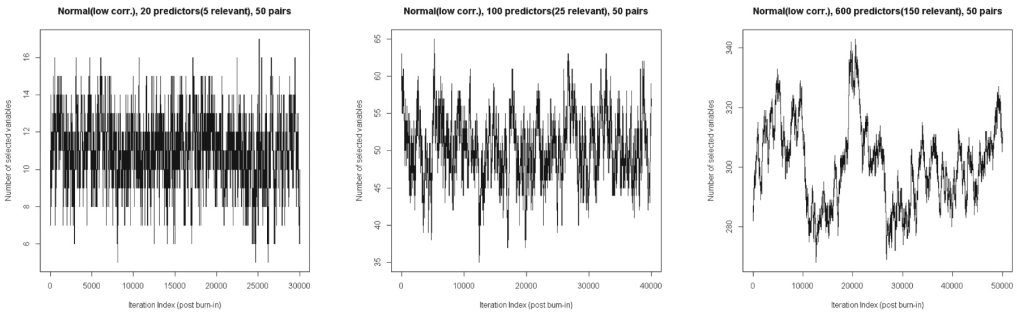
[a] 10% relevant ($Q = 2$ for $K = 20$; $Q = 10$ for $K = 100$).
[b] 25% relevant ($Q = 5$ for $K = 20$; $Q = 25$ for $K = 100$).
[c] 50% relevant ($Q = 10$ for $K = 20$; $Q = 50$ for $K = 100$).

B Trace Plots



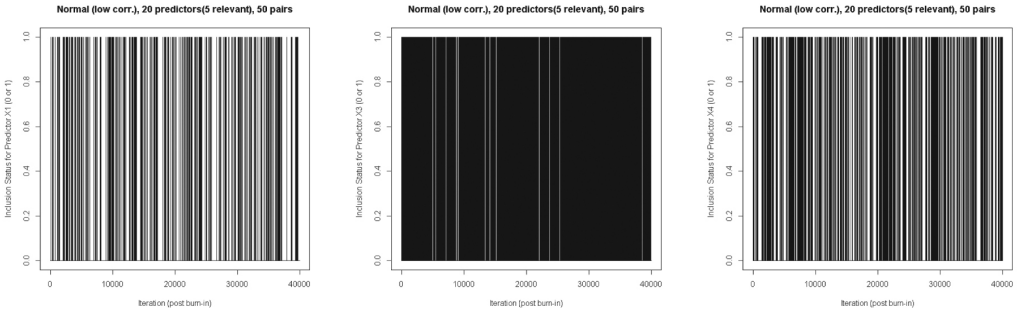
(a) $K = 20$, $Q = 5$ (b) $K = 100$, $Q = 25$ (c) $K = 600$, $Q = 150$

Figure 4: Log posterior probability plots for simulation settings with K normally distributed features of which Q are relevant and weakly correlated ($I = 50$ pairs).



(a) $K = 20$, $Q = 5$ (b) $K = 100$, $Q = 25$ (c) $K = 600$, $Q = 150$

Figure 5: Number of selected features across post burn-in iterations for simulation settings with K normally distributed features of which Q are relevant and weakly correlated ($I = 50$ pairs).



(a) Feature 1 (X_1) (b) Feature 2 (X_3) (c) Feature 3 (X_4)

Figure 6: Inclusion status (0 = no, 1 = yes) across burn-in iterations for features with top 3 inclusion probabilities for simulation settings with $K = 20$ normally distributed features of which $Q = 5$ are relevant and weakly correlated ($I = 50$ pairs).

References

[1]Balasubramanian R, Houseman EA, Coull BA, Lev M, Schwamm LH, Betensky RA. Variable importance in matched case-control studies in settings of high dimensional data. J Roy Stat Soc Ser C 2014;63:639–655.
[2]Anglim PP, Galler JS, Koss MN, Hagen JA, Turla S, Campan M, et al. Identification of a panel of sensitive and specific DNA methylation markers for squamous cell lung cancer. Mol Cancer 2008;7:62.
[3]Westman E, Simmons A, Zhang Y, Muehlboeck JS, Tunnard C, Liu Y, et al. Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls. Neuroimage 2011;54:1178–1187.
[4]Breslow N, Day N. Statistical methods in cancer research (vol. 1): the analysisof case-control studies. Lyon: IARC, 1980.
[5]Tan Q, Thomassen M, Kruse TA. Feature selection for predicting tumor metastases in microarray experiments using paired design. Cancer Inf 2007;3:213–218.
[6]Adewale AJ, Dinu I, Yasui Y. Boosting for correlated binary classification. J Comput Graph Stat 2010;19:140–153.
[7]Friedman J. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–1232.

[8]Freund Y, Schapire R. A short introduction to boosting. J Jpn Soc Artif Intell 1999;14:771–780.
[9]Frank I, Friedman J. A statistical view of some chemometrics regression tools. Technometrics 1993;35:109–135.
[10]Breiman L. Random forests. Mach Learn 2001;45:5–32.
[11]Qian J, Payabvash S, Kemmling A, Lev M, Schwamm L, Betensky RA. Variable selection and prediction using a nested, matched case-control study: application to hospital acquired pneumonia in stroke patients. Biometrics, 2013; 70:153-163. doi:10.1111/biom.12113.
[12]Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B Methodol 1996;58:267–288.
[13]Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc Ser B 2005;67:301–320.
[14]Mitchell T, Beauchamp J. Bayesian variable selection in linear regression. J Am Stat Assoc 1988;83:1023–1032.
[15]George E, McCulloch R. Variable selection via Gibbs sampling. J Am Stat Assoc 1993;88:881–889.
[16]Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. Ann Stat 2005;33:730–773.
[17]George E, McCulloch R. Approaches for Bayesian variable selection. Stat Sin 1997;7:339–373.
[18]Barbieri M, Berger J. Optimal predictive model selection. Ann Stat 2004;32:870–897.
[19]Ishwaran H, Rao S. Clustering gene expression profile data by selective shrinkage. Stat Probab Lett 2008;78:1490–1497.
[20]Rockova V, George E. The spike-and-slab LASSO. Submitted manuscript, 2015:1–39.
[21]Rockova V. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. Ann Stat 2015:1–44. (under revision).
[22]Pang X, Gill J. Spike and slab prior distributions for simultaneous Bayesian hypothesis testing, model selection, and prediction, of nonlinear outcomes. Washington University in St. Louis 2009.
[23]Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003;19:90–97.
[24]Sha N, Vannucci M, Tadesse MG, Brown PJ, Dragoni I, Davies N, et al. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. Biometrics 2004;60:812–819.
[25]Zhou X, Liu KY, Wong ST. Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inf 2004;37:249–259.
[26]Celeux G, Anbari ME, Marin JM, Robert CP. Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. Bayesian Anal 2012;7:477–502.
[27]Johnson V. On numerical aspects of Bayesian model selection in high and ultrahigh-dimensional settings. Bayesian Anal 2004;1:1–17.
[28]Smith M, Fahrmeir L. Spatial Bayesian variable selection with application to functional magnetic resonance imaging. J Am Stat Assoc 2007;102:417–431.
[29]Stingo F, Chen Y, Tadesse M, Vannucci M. Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. Ann Appl Stat 2011;5:1978–2002.
[30]Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J Roy Stat Soc Ser B Stat Methodol 2006;68:49–67.
[31]Meier L, van de Geer S, Bühlmann P. The group lasso for logistic regression. J Roy Stat Soc Ser B Stat Methodol 2008;70:53–71.
[32]Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. Technical report. Department of Statistics, Stanford University, 2010.
[33]Fan J, Li R. Variable selection via nonconcave penalized likelihood and its Oracle properties. J Am Stat Assoc 2001;96:1348–1360.
[34]Zou H. The adaptive lasso and its Oracle properties. J Am Stat Assoc 2006;101:1418–1429.
[35]Gelman A, Roberts G, Gilks W. Efficient metropolis jumping rules. Bayesian Stat 1996;5:599–607.
[36]Roberts G, Gelman A, Gilks W. Weak convergence and optimal scaling of random walk metropolis algorithms. Ann Appl Probab 1997;7:110–120.
[37]Lamnisos D, Griffin JE, Steel MF. Adaptive Monte Carlo for Bayesian variable selection in regression models. J Comput Graph Stat 2013;22:729–748. doi:10.1080/10618600.2012.694756.
[38]Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. Stat Sci 1992;7:457–511.
[39]Brooks S, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 1998;7:434–455.
[40]Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 2014;15:2869–2909.
[41]Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regression. Stat Sin 26 (2016), 35-67. doi:10.5705/ss.2014.042.
[42]Core Team R, Language R: A. and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014;Available at http://www.R-project.org/.
[43]Kemmling A, Lev M, Payabvash S, Betensky R, Qian J, Masrur S, et al. Hospital acquired pneumonia is linked to right peri-insular stroke. PLoS ONE 2013;8:e71141. doi:10.1371/journal.pone.0071141.
[44]Desikan R, Ségonne F, Fischl B, Quinn B, Dickerson B, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006;31:968–980.
[45]Hua K, Zhang J, Wakana S, Jiang H, Li X, Reich D, et al. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. Neuroimage 2008;39:336–347.
[46]Cechetto D, Chen S. Subcortical sites mediating sympathetic responses from insular cortex in rats. Am J Physiol 1990;258:R245–255.
[47]Sander D, Klingelhöfer J. Changes of circadian blood pressure patterns and cardiovascular parameters indicate lateralization of sympathetic activation following hemispheric brain infarction. J Neurol 1995;242:313–318.
[48]Meyer S, Strittmatter M, Fischer C, Georg T, Schmitz B. Lateralization in autononic dysfunction in ischemic stroke involving the insular cortex. Neuroreport 2004;15:357–361.
[49]Colivicchi F, Bassi A, Santini M, Caltagirone C. Cardiac autonomic derangement and arrhythmias in right-sided stroke with insular involvement. Stroke J Cereb Circ 2004;35:2094–2098.
[50]Muller P, Parmigiani G, FDR Rice K. Bayesian multiple comparisons rules. Technical report. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 115. 2006;Available at http://biostats.bepress.com/jhubiostat/paper115.
[51]Cassese A, Guindani M, Tadesse MG, Falciani F, Vannucci M. A hierarchical Bayesian model for inference of copy number variants and their association to gene expression. Ann Appl Stat 2014;8:148–175. doi:10.1214/13-AOAS705.

[52]Lewin A, Saadi H, Peters JE, Moreno-Moral A, Lee JC, Smith KG, et al. MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. Bioinformatics 2016;32:523–532. doi:10.1093/bioinformatics/btv568.
[53]Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B 1995;57:289–300.
[54]Guo Y , Balasubramanian R. Comparative evaluation of classifiers in the presence of statistical interactions between features in high dimensional data settings. Int J Biostat 2012;8. Article 17. doi:10.1515/1557-4679.1373.
[55]Chipman H. Bayesian variable selection with related predictors. Canadian J Stat 1996;24:17–36.
[56]Mukherjee B, Liu I, Sinha S. Analysis of matched case-control data with multiple ordered disease states: possible choices and comparisons. Stat Med 2007;26:3240–3257.
[57]Carroll R, Ruppert D, Stefanski L. Monographs on statistics and applied probability. Measurement error in nonlinear models, 1st ed. Vol. 63. Boca Raton: Chapman & Hall/ CRC, 1995.