Penalized logistic regression for high-dimensional DNA methylation data with case-control studies

Hokeun Sun and Shuang Wang

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

ABSTRACT

Motivation: DNA methylation is a molecular modification of DNA that plays crucial roles in regulation of gene expression. Particularly, CpG rich regions are frequently hypermethylated in cancer tissues, but not methylated in normal tissues. However, there are not many methodological literatures of case-control association studies for high-dimensional DNA methylation data, compared with those of microarray gene expression. One key feature of DNA methylation data is a grouped structure among CpG sites from a gene that are possibly highly correlated. In this article, we proposed a penalized logistic regression model for correlated DNA methylation CpG sites within genes from high-dimensional array data. Our regularization procedure is based on a combination of the $l_1$ penalty and squared $l_2$ penalty on degree-scaled differences of coefficients of CpG sites within one gene, so it induces both sparsity and smoothness with respect to the correlated regression coefficients. We combined the penalized procedure with a stability selection procedure such that a selection probability of each regression coefficient was provided which helps us make a stable and confident selection of methylation CpG sites that are possibly truly associated with the outcome.

Results: Using simulation studies we demonstrated that the proposed procedure outperforms existing main-stream regularization methods such as lasso and elastic-net when data is correlated within a group. We also applied our method to identify important CpG sites and corresponding genes for ovarian cancer from over 20,000 CpGs generated from Illumina HumanMethylation27K Beadchip. Some genes identified are potentially associated with cancers.

Contact: sw2206@columbia.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on December 12, 2011; revised on February 29, 2012; accepted on March 22, 2012

1 INTRODUCTION

DNA methylation, which is the addition of a methyl group to the $5'$ position of cytosine in the context of a CpG dinucleotide, is a molecular modification of DNA that plays crucial roles in regulation of gene expression. Particularly, CpG rich regions are frequently hypermethylated in cancer tissues, but not methylated in normal tissues. Tremendous amounts of DNA methylation data have recently been generated from high-throughput DNA methylation platforms. For example, the Illumina GoldenGate array, the Illumina Infinium HumanMethylation27K array and the most recent Illumina Infinium HumanMethylation450K array are popularly used. These platforms are based on genotyping bisulfite converted DNA. The results of the array, the methylation status of the interrogated CpG site are a sequence of $\beta$-values, one for each locus, calculated as the average of approximately 30 replicates of the quantity (Bibikova et al. 2004).

$$\beta = \frac{\max(M, 0)}{\max(U, 0) + \max(M, 0) + 100}$$

where $U$ is the fluorescent signal from an unmethylated allele on a single bead, and $M$ is that from a methylated allele. A maximum between signal intensity and 0 is chosen to compensate for negative signals due to background subtraction. The constant 100 is to regularize $\beta$-values where both $M$ and $U$ values are small. This $\beta$-value ranges continuously from 0 (unmethylated) to 1 (completely methylated) and reflects the methylation level of each CpG site.

Many researchers have applied statistical classification methods to select differently methylated loci (Houseman et al. 2008; Kuan et al. 2010; Siegmund et al. 2011). Statistical approaches developed for gene expression data may not be applied directly to methylation data since many genes are methylated, while only a few genes are differently expressed. However, disease related CpG regions should still be sparse, in which case the problem is equivalent to identify a few relevant genes from high-dimensional gene expression data. Compared with the case-control studies of gene expression data, there are currently not many methodological developments for methylation data (Wang et al. 2011) has recently proposed a likelihood-based uniform-normal mixture model to select differently methylated loci between case and control groups. One difference of methylation data from gene expression is that the former ranges between 0 and 1. But, this is not an issue in regression frameworks. Another key feature of DNA methylation data that has not been fully utilized is the group structure within a gene. With the Illumina HumanMethylation27K array, there are about 1–22 CpG sites per gene where methylation levels of CpG sites within a gene are usually correlated. Unlike genotype data with single nucleotide polymorphisms (SNPs), $\beta$-values $\leq 0.1$ are continuous thus correlations among them can be observed more precisely. Note that although CpG sites within a gene are correlated, some CpG sites might be neutral while some CpG sites might be causal. Therefore, considering these features of DNA methylation data, in this article we proposed a penalized logistic regression model for correlated CpG sites within a gene as predictors of a disease status. The proposed method can select CpG sites individually that associated with a disease while grouping effects are encouraged. The penalized logistic regression has been recently used to select SNPs or genes associated with a disease in high-dimensional
The rest of this article is organized as follows. In Section 2, we
penalty of Meier
association studies (GWAS). Their approaches basically combine
model that uses the mixture of group and lasso penalties in logistic
performance of lasso penalized logistic regression in case-control

\[ \beta \]

Let us denote the methylation

2.1 Penalized logistic regression

\[ \xi \]

\[ x_i \]

\[ p \]

\[ y_i \]

\[ \exp(\theta_0 + \xi^T \theta) \]

\[ 1 + \exp(\theta_0 + \xi^T \theta) \]

The intercept parameter \( \theta_0 \) and regression coefficients \( \theta = (\theta_1, \ldots, \theta_p)^T \) can be estimated by minimizing the objective function \( \bar{\theta} \).

In the work of \( \bar{\theta} \) and \( \bar{\theta} \) (2008, 2010), the \( k \)-dimensional Laplacian matrix \( L = \{ l_{uv} \} \) was used to represent a graph structure when the network information of predictors is provided. It is defined as

\[ l_{uv} = \begin{cases} 1 & \text{if } u = v \text{ and } d_u \neq 0 \\ -((d_u d_v)^{-\frac{1}{2}})^2 & \text{if } u \text{ and } v \text{ are linked with each other} \\ 0 & \text{otherwise,} \end{cases} \]

where \( d_u \) is the total number of links of the \( u \)-th predictor, and it is often called a degree of the vertex \( u \) in graph theory. Their penalty function is

\[ P(\theta) = |\theta_1| + |\lambda_1 L \theta|_1 \]

where \( | \cdot |_1 \) is a \( l_1 \) norm penalty and the \( \lambda_2 \) penalty in order to account for some smoothness of the coefficients.

The elastic-net (Enet) procedure \( \text{Zou and Hastie (2005)} \) is a compromise between a ridge regression penalty and a lasso penalty, which is to encourage a grouping effect of highly correlated variables. When genetic network information such as a genetic pathway is available, the Enet penalty becomes the Enet penalty. \( \text{Zou and Hastie (2005)} \) Also, the estimates of ridge regression for logistic regression can be obtained when \( \lambda_1 = 0 \) and \( \lambda_2 = 0 \).

This penalty is defined as a combination of the \( l_1 \) penalty and squared penalized logistic regression for the analysis of case-control DNA methylation data, where

\[ P_\lambda(\theta) = |\theta_1| + \lambda_1 |\lambda_2 L \theta|_1 \]

Here we extended it to the penalized logistic regression for the analysis of case-control DNA methylation data. First, we need to specify a network structure that describes correlation patterns for methylation measures of CpG sites within a group. There are two networks prevalently used in graph theory that fit our situation, namely, the ring network and the fully connected (F.con) network. Figure \( \text{H} \) depicts these network topologies in the scenario when there are 6 genes and each of which consists of 1, 2, 3, 4, 6 and 9 CpG sites. In the ring network, we assumed that the first and last CpG sites from a gene are connected with each other so that all CpG sites within the same gene have the same number of links. In contrast, a F.con network assumes that all CpG sites within the same gene are connected with each other. In the penalized regression model, both networks specify a group structure of predictors so that the coefficients of correlated predictors within the same group can shrink toward each other, allowing them to borrow information from each other.

Motivated by the graph-constrained procedure by \( \text{Li and Li (2008, 2010)} \), we extended it to the logistic regression model for the analysis of case-control DNA methylation data, where our Laplacian matrix represents CpG sites clustered within genes.

The rest of this article is organized as follows. In Section 2, we described our statistical model and regularization procedure for grouped and correlated predictors (CpG sites) together with the computational algorithm and the stability selection procedure. We then presented the simulation results in Section 3, where existing main-stream variable selection methods were compared with the proposed method when predictors within a group are correlated. Section 4 analyzed a real DNA methylation data in a case-control study of ovarian cancer. Finally, we gave a brief discussion of the method and future research in Section 5.

2 METHODS

2.1 Penalized logistic regression

Let us denote the methylation \( p \)-values of the \( i \)-th individual by \( x_i = (x_{i1}, \ldots, x_{ip})^T, i = 1, \ldots, n \) and \( p \) is the total number of CpG sites considered in the analysis. The penalized logistic regression is defined as

\[ -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) + (1-y_i) \log(1-p(x_i)) + P(\theta), \]

where \( P(\theta) \) is a penalty function, and the response \( y_i \) is 0 for controls and 1 for cases. The probability that the \( i \)-th individual is a case based on his/her DNA methylation information is denoted as

\[ P(\theta) = |\theta_1| + |\lambda_1 L \theta|_1 \]

The intercept parameter \( \theta_0 \) and regression coefficients \( \theta = (\theta_1, \ldots, \theta_p)^T \) can be estimated by minimizing the objective function \( \bar{\theta} \).

In the work of \( \lambda_1 \) and \( \lambda_2 \) (2008, 2010), the \( \lambda_2 \) proposed a model that uses the mixture of group and lasso penalties in logistic regression to identify common and rare variants in genome-wide association studies (GWAS). Their approaches basically combine the \( l_1 \) norm penalty of \( \text{Tibshirani (1996)} \) and the \( l_2 \) norm group penalty of \( \text{Meier et al. (2008)} \). These penalties induce the overall sparsity and group sparsity, respectively. Consequently, predictors with stronger effects on responses are more likely to be selected into the model. However, none of the methods above is designed to encourage selection effects for highly correlated predictors.

For the analysis of correlated microarray gene expression data, several variable selection methods have been developed in the past decade. The fused lasso \( \text{Tibshirani et al. (2005)} \) imposes the \( l_1 \) penalty on the absolute differences of the regression coefficients in order to account for some smoothness of the coefficients. The elastic-net (Enet) procedure \( \text{Zou and Hastie (2005)} \), which is a compromise between a ridge regression penalty and a lasso penalty, encourages a grouping effect of highly correlated variables. When genetic network information such as a genetic pathway is available, the Enet procedure \( \text{Li and Li (2008, 2010)} \) proposed a graph-constrained regularization procedure. In their method, a Laplacian matrix representing a graph structure was imposed on a ridge-regression penalty of the Enet procedure such that linked genes have a smoothness penalty on their regression coefficients. The authors demonstrated that when the information of gene networks is incorporated into the regularization procedure, it can select more relevant genes than the lasso and Enet procedures.

Motivated by the graph-constrained procedure by \( \text{Li and Li (2008, 2010)} \), we extended it to the logistic regression model for the analysis of case-control DNA methylation data, where our Laplacian matrix represents CpG sites clustered within genes.

The rest of this article is organized as follows. In Section 2, we described our statistical model and regularization procedure for grouped and correlated predictors (CpG sites) together with the computational algorithm and the stability selection procedure. We then presented the simulation results in Section 3, where existing main-stream variable selection methods were compared with the proposed method when predictors within a group are correlated. Section 4 analyzed a real DNA methylation data in a case-control study of ovarian cancer. Finally, we gave a brief discussion of the method and future research in Section 5.

2 METHODS

2.1 Penalized logistic regression

Let us denote the methylation \( p \)-values of the \( i \)-th individual by \( x_i = (x_{i1}, \ldots, x_{ip})^T, i = 1, \ldots, n \) and \( p \) is the total number of CpG sites considered in the analysis. The penalized logistic regression is defined as

\[ -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) + (1-y_i) \log(1-p(x_i)) + P(\theta), \]

where \( P(\theta) \) is a penalty function, and the response \( y_i \) is 0 for controls and 1 for cases. The probability that the \( i \)-th individual is a case based on his/her DNA methylation information is denoted as

\[ P(\theta) = |\theta_1| + |\lambda_1 L \theta|_1 \]

The intercept parameter \( \theta_0 \) and regression coefficients \( \theta = (\theta_1, \ldots, \theta_p)^T \) can be estimated by minimizing the objective function \( \bar{\theta} \).

In the work of \( \text{Li and Li (2008, 2010)} \), we extended it to the logistic regression model for the analysis of case-control DNA methylation data, where

\[ P(\theta) = |\theta_1| + |\lambda_1 L \theta|_1 \]

This penalty is defined as a combination of the \( l_1 \) penalty and squared penalized logistic regression for the analysis of case-control DNA methylation data. First, we need to specify a network structure that describes correlation patterns for methylation measures of CpG sites within a group. There are two networks prevalently used in graph theory that fit our situation, namely, the ring network and the fully connected (F.con) network. Figure \( \text{H} \) depicts these network topologies in the scenario when there are 6 genes and each of which consists of 1, 2, 3, 4, 6 and 9 CpG sites. In the ring network, we assumed that the first and last CpG sites from a gene are connected with each other so that all CpG sites within the same gene have the same number of links. In contrast, a F.con network assumes that all CpG sites within the same gene are connected with each other. In the penalized regression model, both networks specify a group structure of predictors so that the coefficients of correlated predictors within the same group can shrink toward each other, allowing them to borrow information from each other.
from each other. Unlike the group penalized method of [Lepage et al. 2011], our model still performs individual selection, so neutral CpG sites within significant genes are not forced to remain in the final model.

To compare the biological basis of the ring and the F-con networks, although it depends on the underlying true correlation patterns of CpG sites within genes, the later might be more appropriate for DNA methylation data since one CpG site might be correlated with the rest sites within the same gene. The ring network assumes correlation of just flanking sites except for the first and last sites. Note that the link between the first and the last sites in the ring network is less possible biologically but is imposed to apply the ring network. In terms of variable selection, the ring network will induce a mild grouping effect on a large group since the number of links per CpG site is fixed at 2 for all genes that have >2 CpG sites. In contrast, the total number of links per CpG site in the F-con network model is increased by the number of CpG sites within a gene. So, the fully connected network usually produces a strong grouping effect since variables with more links are more likely to be selected in the network-based regularized procedure. This is also discussed by [Huang et al. 2010], if all CpG sites in a gene are true signals, the F-con network are desirable. But, if there are both neutral sites and causal sites in a gene, the ring network might be preferable to weaken the grouping effect. However, in simulation study and real data application we showed that the selection results using the two networks are almost identical. Also note that the Laplacian matrix in our model \( \mathcal{B} \) has the form of a blockwise diagonal matrix for both network models.

2.2 Computational algorithms

Let and [Huang et al. 2010] has applied the coordinate descent algorithm of Friedman et al. [2007] to obtain the minimizer of the function \( Q(\theta_0) \) when a response variable follows a Gaussian distribution. But this cannot be directly applied to the logistic regression. Moreover, recent publication of [Friedman et al. 2010] has drastically improved the computational efficiency, and a fast package glmnet was developed for the Enet regularization procedure of the logistic regression. Since the Enet simply replaces the Laplacian matrix by an identity matrix in the penalty function \( \mathcal{B} \), we could impose the Laplacian matrix in the glmnet code and be benefited from the efficient computation algorithm already implemented. In this section, we briefly explained how the Enet procedure and our method differ in terms of algorithm.

Let us denote the negative log likelihood of the logistic regression model by \( -Q(\theta_0) \), which is equivalent to the first term of the function \( Q(\theta) \). Our objective function is then

\[
Q(\theta_0) = -Q(\theta_0) + P(\theta),
\]

where

\[
P(\theta) = \sum_{i=1}^{n} \left( y_i \log \left( 1 + e^{-\theta^T x_i} \right) + (1 - y_i) \log \left( 1 + e^{\theta^T x_i} \right) \right),
\]

and

\[
\lambda = \lambda_1 + 2 \lambda_2 \quad \text{and} \quad u = \frac{\lambda_1}{\lambda_1 + 2 \lambda_2}.
\]

As \( Q(\theta_0) \) is approximated by a second-order Taylor series expansion at current estimates \( (\theta_0^*, \lambda^*) \), the function \( Q(\theta) \) can be re-written as

\[
Q(\theta_0) = \sum_{i=1}^{n} w_i \left( z_i + \gamma_{-i} \theta_0 - x_i^T \theta \right)^2 + P(\theta),
\]

where

\[
z_i = \frac{z_i}{\sqrt{\gamma_{-i}}} + \gamma_{-i} \theta_0^t + w_i^{-1} \gamma_{-i} \varepsilon_i, \quad \theta = \sum_{i=1}^{n} P^t(\gamma_{-i}) (1 - P^t(\gamma_{-i}))^{-1}.
\]

We refer the readers to [Friedman et al. 2010] for the details of this derivation.

Next, we can compute the gradient at \( \theta_0 = \theta_0^* \) while the other estimates \( \theta_i^* \) for all \( i \neq 0 \) are fixed. By setting the gradient to 0 and solve for \( \theta_0 \), we can get the current estimate of \( \theta_0^* \) using the following formula,

\[
\theta_0^t = \left( \gamma_{-i}^{-1} \sum_{i=1}^{n} w_i x_i^T z_i + (1 - \alpha_0^t) x_i^T \gamma_{-i} \theta_0^t \right) \gamma_{-i}^{-1} \sum_{i=1}^{n} w_i x_i^T x_i + (1 - \alpha_0^t),
\]

where \( \theta_0^t = \theta_0^t + \sum_{i=1}^{n} w_i x_i^T \theta_i^t \),

\[
g(\theta) = \sum_{i=1}^{n} w_i \theta_i^2,
\]

and \( S(\gamma) \) is a soft thresholding operator with value

\[
sign(\gamma)(\mid \gamma \mid - \gamma) = \begin{cases} 
\gamma - \sigma & \text{if } \gamma > 0 \text{ and } \gamma < \sigma \\gamma + \sigma & \text{if } \gamma < 0 \text{ and } \gamma < -\sigma \\0 & \text{otherwise.}
\end{cases}
\]

If the \( \theta \)-th predictor is a unique member of a group, i.e. no links to any other predictors, the function \( g(\theta) \) is equal to 0. If all predictors are isolated, the penalty function \( \mathcal{B} \) is reduced to

\[
P(\theta) = \sum_{i=1}^{n} \left( \theta_i^2 \right),
\]

and thus \( g(\theta) = 0 \) for all \( n = 1, \ldots, p \). In this case the solution becomes exactly the same as that of the Enet procedure. In other words, we only need to include the term \( \lambda (1 - u) g(\theta) \) in the formula of the Enet when we create a group of CpG sites by genes.

2.3 Selection probabilities

In our penalty function \( \mathcal{B} \) we have two tuning parameters to be selected. The parameter \( \lambda \in (0,1) \) induces the model to have the estimates between ridge \( \lambda_1 = 1 \) and lasso \( \lambda_1 = 0 \) and thus \( \lambda_1 = 1 \) becomes indifferent to correlated predictors. The Enet procedure usually uses a small value of \( \lambda \) to overcome collinearity problems in regressions with high-dimensional predictors. However, results of variable selection are very similar with a small perturbation of \( \lambda \). In contrast, the tuning parameter \( \lambda \) forces the model to select more variables as \( \lambda \) decreases with a fixed \( \alpha \). Since the value of \( \lambda \) directly determines the total number of relevant predictors, the choice of \( \lambda \) is crucial in variable selection problems. Cross-validation (CV) is commonly used to find the optimal value of \( \lambda \) in variable selection literatures, although it is known that for high-dimensional data CV generally selects too many variables including some truly unrelated variables [Meinshausen and Buhlmann 2010] in their recent work proposed to compute selection probabilities for all variables, and to include only the variables with high selection probabilities in the model. Their method is based on resampling, and provides a more stable selection for high-dimensional data compared with CV. Recently [Alexander and Lange 2011] have applied it to select relevant SNPs in GWAS. We employed their method to determine important CpG sites in our regularization procedure.

Let us denote \( \Lambda \) as a regularization parameter space, e.g. \( \lambda \in \Lambda \). Let \( \mathcal{I} \) be the \( k \)-th random subsample of \( \{1, \ldots, n\} \) of size \( \{n/2\} \) without replacement, where \( \{x\} \) is the largest integer not greater than \( x \). The selection probability of the \( \alpha \)-th predictor is then defined as

\[
SP(\alpha) = \max_{\lambda \in \Lambda} K \{ \lambda \leq K \} \{ \alpha \neq 0 \},
\]

where \( \hat{g}_i(\alpha) \) is the estimator of \( g_\alpha \) using a regularization procedure based on the subsample \( \mathcal{I} \), given \( \lambda \), and \( K \) is the total number of resampling. Then, the variables whose selection probabilities are greater than some cutoff value, say \( \pi \), are selected in the model. According to [Meinshausen and Buhlmann 2010], selection results do not rely much on the choice of \( \Lambda \), and tend to be very similar with different values of the cutoff \( \pi \). They also provided the formula to select both \( \Lambda \) and \( \pi \) so that the expected number of falsely
selected variables can be controlled when the exchangeability assumption of predictors are met. However, this assumption is a weaker version of independence and is hard to be satisfied with highly correlated genomic data. We thus focused on identifying the CpG sites which have high selection probabilities in the analysis of methylation data.

2.4 Adaptive regularization procedure

Li and Li (2010) have pointed out that their procedure does not perform well when two predictors that are linked with each other are negatively correlated with the response because in this situation the corresponding regression coefficients have different signs, so they are not expected to be locally smooth. Their solution to this problem is to estimate the signs of coefficients first, and then refit the model with the estimated signs. This change does not affect the L2 norm of the penalty function, so we can simply modify it in the following way:

\[ l_{k,g} = \begin{cases} 1 & \text{if } u=v \text{ and } d_k \neq 0 \\ s_{u,v}(d_k)^{-1} & \text{if } u \neq v \text{ and } v \neq 0 \\ 0 & \text{otherwise} \end{cases} \]

where \( s_{u,v} \) is the estimated sign of the \( u \)-th predictor, which can be obtained by ordinary regression estimates when \( p<n \), and ridge estimates for \( p \geq n \). Then, the function \( g(u) \) should be updated in the formula,

\[ g(u) = \sum_{k=1}^{p} \frac{u_k l_{k,g}}{d_k} \]

Li and Li (2010) have demonstrated that this adaptive regularization procedure leads to improved estimates and selection results compared with the ordinary procedure in their simulation studies. Therefore, we also have included the sign estimates in our regularized estimates for both simulation studies and the real data application.

3 SIMULATION STUDIES

We conducted some simulation studies to compare the performance of our proposed method to existing main-stream regularization procedures. In our simulation models, we have 600 genes which consist of 1–9 CpG sites such that 100 genes have 1 CpG site, 150 genes have 2 CpG sites and each 50 genes of the others have 3–9 CpG sites, respectively. Therefore, a total of 2500 CpG sites were simulated.

To mimic real DNA methylation data, we need to generate group correlated variables within the [0,1] range. To do so, we employed the inverse logit transformation of multivariate normal random variables. Thus, the methylation \( \beta \)-values of the \( g \)-th gene for the \( i \)-th individual is calculated as

\[ x_{i,g} = \frac{\exp(t_{i,g})}{1+\exp(t_{i,g})} \]

where \( x \) is a scale parameter and \( p_k \) is the size of the \( g \)-th gene, i.e. \( 1 \leq p_k \leq 9 \). In this study, we set \( x = 0.1 \) and \( x = 0.1 \) so that the distribution of the methylation values have an enriched ‘0’ (unmethylated) and enriched ‘1’ (completely methylated) as previously observed (Wand, 2010).

Next, we specified the true regression coefficients \( \theta \) similar as Li and Li (2010). Let us first denote the coefficients of the \( g \)-th gene by \( \theta_{g}=(\theta_{1,g},...,\theta_{pg,g})^{T} \). Since 600 genes have 1 to 9 CpG sites, we grouped the genes by the number of CpG sites. For example, the first gene group has only one CpG site, and the second group has only two CpG sites, and so on. Then, we selected one gene from each of the 9 different gene groups, which leads to a total of 45 CpG sites. We denoted the regression coefficients of these 45 CpG sites as

\[ \theta_{k,g} = (-1)^{p_k} \frac{\delta}{\sqrt{p_k}} \]

for \( g=1,\ldots,9 \) and referred them as CpG set-1. Similarly, we selected another nine genes from each of the nine different gene groups, and denoted their coefficients as

\[ \theta_{k,g} = (-1)^{p_k} \frac{\delta}{\sqrt{p_k}} \]

for \( g=1,\ldots,9 \). We then simulated the data with different correlation coefficient \( \delta \) and \( \rho \) within \( [0,1] \). Finally, the corresponding responses \( y_i \) was simulated according to a Bernoulli distribution with the following model-based probabilities,

\[ y_i \sim \text{Bernoulli}(p(x_{i1}), p(x_i) = \frac{\exp(x_{i1}^T S_{1}\theta)}{1+\exp(x_{i1}^T S_{1}\theta)}) \]

where \( x_i = (x_{i1},...,x_{i600})^{T} \) and \( S_{1}=(\theta_{1,1},...,\theta_{1,600})^{T} \). For each simulation set, samples were generated until we have 200 cases. We then randomly selected 200 controls from the control pool already generated.

We considered nine different simulation models, differing the strength of the true signals \( \delta \) and the covariance matrix \( S \) within genes. First, three models are defined in the following way,

(1) \( \delta = 1 \) and \( S_{100} = \rho^{i=1} \)
(2) \( \delta = 2 \) and \( S_{100} = \rho^{i=1} \)
(3) \( \delta = 2 \) and \( S_{100} = \rho \) for \( u \neq v \) and \( S_{uv} = 1 \) for \( u = v \),

where the correlation of the first two models is AR(1), and the third model has compound symmetric correlation structure. We then simulated the data with different correlation coefficient \( \rho \) of 0.2, 0.5 and 0.7 for all three models. For each model we repeated simulations 100 times, and selection probabilities for each simulation set was computed based on 100 resamplings.

We compared the performance of the proposed ring and F econ network-based method to that of the group MCP (gMCP) procedure using the R package qgrror (Breheny and Huang, 2009), the lasso and Enet procedures using glmnet (Friedman et al., 2010). Figure 2 shows the box plots of averaged selection probabilities of 70 true signals in 9 different simulation models. It appears that the lasso has the lowest selection probabilities, and the Enet and our methods have similar selection probabilities through all simulation models. The selection probabilities of gMCP are slightly lower than that of the Enet but much higher than that of the lasso. However, the model with the highest selection probabilities does not always lead to the best model, when the model might have selected too many false positives if the selection probabilities of the disease unrelated predictors are also high. Thus, we computed both true positive rates and false positive rates of each procedure while varying the cutoff \( \pi \) of selection probabilities from 0 to 1.
Fig. 2. The averaged selection probabilities of 70 true signals via Lasso, gMCP, Enet, and the ring and F.con network-based procedures are present. The signal strength is set at $\delta = 1$ in A–C, and $\delta = 2$ in D–J. The AR(1) covariance is used in A–F, and the compound symmetric correlation for G–J. The correlation is set at $\rho = 0.2$ for A, D and G, $\rho = 0.5$ for B, E and H, and $\rho = 0.7$ for the others.

Table 1. The area under the averaged ROC curves of Lasso, gMCP, Enet, and the ring and F.con network-based procedures along with the different signal strength $\delta$, covariance $\Sigma$ and correlation coefficient $\rho$.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$\Sigma$</th>
<th>$\rho$</th>
<th>Lasso</th>
<th>gMCP</th>
<th>Enet</th>
<th>Ring</th>
<th>F.con</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1) 0.2</td>
<td>0.6896</td>
<td>0.6980</td>
<td>0.7000</td>
<td>0.7472</td>
<td>0.7470</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1) 0.5</td>
<td>0.7707</td>
<td>0.7963</td>
<td>0.8019</td>
<td>0.8538</td>
<td>0.8562</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1) 0.7</td>
<td>0.8202</td>
<td>0.8632</td>
<td>0.8728</td>
<td>0.9083</td>
<td>0.9127</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1) 0.2</td>
<td>0.7943</td>
<td>0.8012</td>
<td>0.8023</td>
<td>0.8574</td>
<td>0.8582</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1) 0.5</td>
<td>0.8619</td>
<td>0.8831</td>
<td>0.8862</td>
<td>0.9310</td>
<td>0.9321</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(1) 0.7</td>
<td>0.9037</td>
<td>0.9360</td>
<td>0.9410</td>
<td>0.9655</td>
<td>0.9669</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS 0.2</td>
<td>0.8468</td>
<td>0.8592</td>
<td>0.8615</td>
<td>0.9113</td>
<td>0.9124</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS 0.5</td>
<td>0.9010</td>
<td>0.9279</td>
<td>0.9328</td>
<td>0.9563</td>
<td>0.9587</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS 0.7</td>
<td>0.9088</td>
<td>0.9490</td>
<td>0.9549</td>
<td>0.9700</td>
<td>0.9707</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Compound symmetry covariance.

The averaged receiver operating characteristic (ROC) curves of selection results of the procedures are given in Supplementary Materials (Fig. S1). The corresponding area under the ROC curves (AUC) are shown in Table 1. It is obvious that our proposed network-based methods outperform other procedures in all simulation scenarios. Surprisingly, both versions of the proposed network-based methods (Ring and F.con) have very similar AUCs and almost identical ROC curves in all simulation models. This may suggest that the penalty of misidentifying correlation structures of CpG sites within a gene is negligible, but instead the group selection effects of clustering CpG sites is strong enough to enhance overall selection performance and overwhelm the others.

The null simulation (with $\delta = 0$) were also conducted for these five procedures for validation. The results are provided in Supplementary Materials (Figs S2 and S3), where averaged selection probabilities over 100 simulation replications are displayed for all 2500 CpG sites. It can be seen that all procedures produced roughly uniform selection probabilities across all CpG sites. Note that the comparison of selection probabilities across the five procedures in the null simulation is not meaningful, but we should compare selection probabilities across CpG sites.

Since the lasso procedure handles neither group structure nor correlated predictors, their selection performance is always poor for group correlated predictors. The gMCP is better than the lasso as it specifies grouped variables but does not differentiate their correlations. In contrast, the Enet is designed to have grouping effects for correlated predictors without specifying defined groups. Only our proposed methods account for both group structure and correlated variables thus lead to a superior selection performance over the other regularization procedures.

4 DATA ANALYSIS

We applied our proposed method to select differentially methylated CpG sites between ovarian cancer cases and healthy controls using the DNA methylation data generated from Illumina Infinium HumanMethylation27K Beadchip [Teschendorff et al. 2010]. The data is available at the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

We first performed the quality control procedure of the methylation data similar as [Teschendorff et al. 2010] and [Wang 2011], which include the removal of samples with either a low BS conversion efficiency or low CpG coverage. We also left out CpG sites with any missing $\beta$-values. Since our procedure assumes that every CpG site belongs to a single gene, CpG sites without corresponding gene information were removed. We ended up with 20 461 CpG sites from 12 770 genes where we have 152 controls, 119...
Table 2. The CpG sites and corresponding genes with top 20 selection probabilities identified by Enet, and the ring and F.con network-based procedures from the comparison between pre-treatment cases and normal controls

<table>
<thead>
<tr>
<th>Gene</th>
<th>Prob</th>
<th>IlmnID</th>
<th>Gene</th>
<th>Prob</th>
<th>IlmnID</th>
<th>Gene</th>
<th>Prob</th>
<th>IlmnID</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg11009736</td>
<td>0.990</td>
<td>(MARCO)</td>
<td>cg20792833</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.990</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
<tr>
<td>cg109114098</td>
<td>0.985</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td>cg109114098</td>
<td>0.980</td>
<td>(PTPRCAP)</td>
<td></td>
</tr>
</tbody>
</table>

*Indicates the overlapped genes in the top 20 lists of all 5 procedures.

Analysis of high-dimensional DNA methylation data
Beadchip. Some genes in our findings are known to be associated with other types of cancers. Since we mainly focused on identifying the most likely associated Cpg sites through selection probabilities, we did not discuss about estimation performance of our proposed method. In the case that prediction is desirable, we strongly recommend to reobtain unpenalized MLE is often adopted for better prediction (Meier et al., 2008; Wu et al., 2009). To incorporate a prior knowledge of genetic pathways may further improve the selection accuracy. We are currently pursuing the usage of genetic pathways information for the DNA methylation data analysis.

ACKNOWLEDGEMENTS
The authors want to thank the Yale High Performance Computing Center for providing computing facility (supported by NIH grant RR19895) and also want to thank Dr Yanjuan Wang for helpful discussions.

Funding: NCI (grant R03 CA150140) and NIH (grant R01 ES05115).

Conflict of Interest: none declared.

REFERENCES


Table 3. The Cpg sites and corresponding genes with top 20 selection probabilities identified by Enet, and the ring and FCon network-based procedures from the comparison between post-treatment cases and normal controls.

<table>
<thead>
<tr>
<th>Prob</th>
<th>IlmnID</th>
<th>Gene</th>
<th>Prob</th>
<th>IlmnID</th>
<th>Gene</th>
<th>Prob</th>
<th>IlmnID</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cg25580000</td>
<td>(ADCY)*</td>
<td>1.00</td>
<td>cg06653796</td>
<td>(LIME1)*</td>
<td>1.00</td>
<td>cg06653796</td>
<td>(LIME1)*</td>
</tr>
<tr>
<td>2</td>
<td>cg06653796</td>
<td>(LIME1)*</td>
<td>1.00</td>
<td>cg06654034</td>
<td>(TPAP)*</td>
<td>1.00</td>
<td>cg06654034</td>
<td>(TPAP)*</td>
</tr>
<tr>
<td>3</td>
<td>cg10986043</td>
<td>(ADCY)*</td>
<td>0.975</td>
<td>cg25580000</td>
<td>(ADCY)*</td>
<td>0.975</td>
<td>cg25580000</td>
<td>(ADCY)*</td>
</tr>
<tr>
<td>4</td>
<td>cg13379236</td>
<td>(EGF)*</td>
<td>0.95</td>
<td>cg13379236</td>
<td>(EGF)*</td>
<td>0.95</td>
<td>cg13379236</td>
<td>(EGF)*</td>
</tr>
<tr>
<td>5</td>
<td>cg06547797</td>
<td>(GAS2)*</td>
<td>0.94</td>
<td>cg06547797</td>
<td>(GAS2)*</td>
<td>0.94</td>
<td>cg06547797</td>
<td>(GAS2)*</td>
</tr>
<tr>
<td>6</td>
<td>cg05135288</td>
<td>(RIOT)*</td>
<td>0.935</td>
<td>cg05135288</td>
<td>(RIOT)*</td>
<td>0.935</td>
<td>cg05135288</td>
<td>(RIOT)*</td>
</tr>
<tr>
<td>7</td>
<td>cg20357806</td>
<td>(PPBP)*</td>
<td>0.87</td>
<td>cg12006284</td>
<td>(WTI)*</td>
<td>0.87</td>
<td>cg12006284</td>
<td>(WTI)*</td>
</tr>
<tr>
<td>8</td>
<td>cg12006284</td>
<td>(WTI)*</td>
<td>0.865</td>
<td>cg20357806</td>
<td>(PPBP)*</td>
<td>0.86</td>
<td>cg20357806</td>
<td>(PPBP)*</td>
</tr>
<tr>
<td>9</td>
<td>cg21640749</td>
<td>(CD300LF)*</td>
<td>0.84</td>
<td>cg2435895</td>
<td>(C0XTA1)*</td>
<td>0.84</td>
<td>cg2435895</td>
<td>(C0XTA1)*</td>
</tr>
<tr>
<td>10</td>
<td>cg12243272</td>
<td>(CBP)*</td>
<td>0.81</td>
<td>cg21640749</td>
<td>(CD300LF)*</td>
<td>0.81</td>
<td>cg21640749</td>
<td>(CD300LF)*</td>
</tr>
<tr>
<td>11</td>
<td>cg09626634</td>
<td>(EBI2)</td>
<td>0.805</td>
<td>cg12243272</td>
<td>(CBP)*</td>
<td>0.815</td>
<td>cg12243272</td>
<td>(CBP)*</td>
</tr>
<tr>
<td>12</td>
<td>cg29888566</td>
<td>(WTDC1B)*</td>
<td>0.80</td>
<td>cg19573166</td>
<td>(SLC22A17)</td>
<td>0.80</td>
<td>cg19573166</td>
<td>(SLC22A17)</td>
</tr>
<tr>
<td>13</td>
<td>cg00134539</td>
<td>(UBASH3A)</td>
<td>0.795</td>
<td>cg10467098</td>
<td>(Blox3)*</td>
<td>0.80</td>
<td>cg19573166</td>
<td>(SLC22A17)</td>
</tr>
<tr>
<td>14</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
<td>0.785</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
<td>0.785</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
</tr>
<tr>
<td>15</td>
<td>cg22988566</td>
<td>(WTDC1B)*</td>
<td>0.76</td>
<td>cg15096140</td>
<td>(MYOIV)*</td>
<td>0.76</td>
<td>cg15096140</td>
<td>(MYOIV)*</td>
</tr>
<tr>
<td>16</td>
<td>cg05135288</td>
<td>(RHOT2)*</td>
<td>0.75</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
<td>0.75</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
</tr>
<tr>
<td>17</td>
<td>cg00134539</td>
<td>(UBASH3A)</td>
<td>0.755</td>
<td>cg23506842</td>
<td>(PTPP77)</td>
<td>0.755</td>
<td>cg23506842</td>
<td>(PTPP77)</td>
</tr>
<tr>
<td>18</td>
<td>cg16853982</td>
<td>(ACTN2)</td>
<td>0.75</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
<td>0.75</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
</tr>
<tr>
<td>19</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
<td>0.745</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
<td>0.75</td>
<td>cg13745870</td>
<td>(SLC22A17)</td>
</tr>
<tr>
<td>20</td>
<td>cg13247990</td>
<td>(MLCK)</td>
<td>0.735</td>
<td>cg23917399</td>
<td>(TFNAP8)</td>
<td>0.735</td>
<td>cg23917399</td>
<td>(TFNAP8)</td>
</tr>
</tbody>
</table>

*Indicates the overlapped genes in the top 20 lists of all 5 procedures.
Analysis of high-dimensional DNA methylation data


