

Research Article

A Two-Stage Penalized Logistic Regression Approach to Case-Control Genome-Wide Association Studies

Jingyuan Zhao¹ and Zehua Chen²

¹ Human Genetics, Genome Institute of Singapore, 60 Biopolis, Genome No. 02-01, Singapore 138672

² Department of Statistics and Applied Probability, National University of Singapore,

3 Science Drive 2, Singapore 117546

Correspondence should be addressed to Zehua Chen, stachen@nus.edu.sg

Received 20 September 2011; Accepted 28 October 2011

Academic Editor: Yongzhao Shao

Copyright © 2012 J. Zhao and Z. Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We propose a two-stage penalized logistic regression approach to case-control genome-wide association studies. This approach consists of a screening stage and a selection stage. In the screening stage, main-effect and interaction-effect features are screened by using L_1 -penalized logistic like-lihoods. In the selection stage, the retained features are ranked by the logistic likelihood with the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) and Jeffrey's Prior penalty (Firth, 1993), a sequence of nested candidate models are formed, and the models are assessed by a family of extended Bayesian information criteria (J. Chen and Z. Chen, 2008). The proposed approach is applied to the analysis of the prostate cancer data of the Cancer Genetic Markers of Susceptibility (CGEMS) project in the National Cancer Institute, USA. Simulation studies are carried out to compare the approach with the pair-wise multiple testing approach (Marchini et al. 2005) and the LASSO-patternsearch algorithm (Shi et al. 2007).

1. Introduction

The case-control genome-wide association study (GWAS) with single-nucleotide polymorphism (SNP) data is a powerful approach to the research on common human diseases. There are two goals of GWAS: (1) to identify suitable SNPs for the construction of classification rules and (2) to discover SNPs which are etiologically important. The emphasis is on the prediction capacity of the SNPs for the first goal and on the etiological effect of the SNPs for the second goal. The phrase "an etiological SNP" is used in the sense that either the SNP itself is etiological or it is in high-linkage disequilibrium with an etiological locus. Well-developed classification methods in the literature can be used for the first goal. These methods include classification and regression trees [1], random forest [2], support vector machine [3], and logic regression [4]. In this article, we focus on statistical methods for the second goal.

The approach of multiple testing based on single or paired SNP models is commonly used for the detection of etiological SNPs. Either the Bonferroni correction is applied for the control of the overall Type I error rate, see, for example, Marchini et al. [5] or some methods are used to control the false discovery rate (FDR), see, Banjamini and Hochberg [6], Efron and Tibshirani [7], and Storey and Tibshirani [8]. Other variants of multiple testing have also been advocated, see Hoh and Ott [9]. The multiple test approach considers either a single SNP or a pair of SNPs at a time. It does not adjust for the effects of other markers. If there are many loci having high sample correlations with a true genetic variant, which is common in GWAS, it is prone to result in spurious etiological loci.

It is natural to seek alternative methods that overcome the drawback of multiple testing. Such methods must have the nature of considering many loci simultaneously and assessing the significance of the loci by their synergistic effect. When the synergistic effect is of concern, adding loci spuriously correlated to an etiological locus does not contribute to the synergistic effect while the etiological locus has already been considered. Thus the drawback of multiple testing can be avoided. In this paper, we propose a method of the abovementioned nature: a two-stage penalized logistic regression approach. In the first stage of this approach, L_1 -penalized logistic regression models are used together with a tournament procedure [10] to screen out apparently unimportant features (by features we refer to the covariates representing SNPs or their products). In the second stage, logistic models with the SCAD penalty [11] plus the Jeffrey's prior penalty [12] are used to rank the retained features and form a sequence of nested candidate models. The extended Bayesian information criteria (EBIC, [13, 14]) are used for the final model selection. In both stages of the approach, the features are assessed by their synergistic effects.

The two-stage strategy has been considered by other authors. For example, J. Fan and Y. Fan [15] adopted this strategy for high-dimensional classification, and Shi et al. [16] developed a two-stage procedure called LASSO-patternsearch. Sure independence screening (SIS, [17]) and its ramifications such as correlation screening and t -tests are commonly used in the screening stage. Compared with SIS approaches, the tournament screening with L_1 -penalized likelihood produces less spuriously correlated features while enjoying the sure screening property possessed by the SIS approaches, see Z. Chen and J. Chen [10] and the comprehensive simulation studies by Wu [18], which has an impact on the accuracy of feature selection in the second stage, see Koh [19]. The L_1 -penalized likelihood is easier to compute than that with the SCAD penalty. However, the SCAD penalty has an edge over the L_1 -penalty in ranking the features so that the ranks are more consistent with their actual effects. This has been observed in simulation studies, see Zhao [20]. It is possibly due to the fact that the L_1 penalty over-penalizes those features with large effects compared with SCAD penalty that does not penalize large effects at all. Jeffrey's prior penalty is added to handle the difficulty caused by separation of data that usually presents in logistic regression models with factor covariates, see Albert and Anderson [21]. If, within any of the categories determined by the levels of the factors, the responses are all 1 or 0, it is said that there is a complete data separation. When the responses within any of the categories are almost all 1 or 0, it is referred to as a quasicomplete data separation. When there is separation (complete or quasi-complete), the maximum likelihood estimate of the corresponding coefficients becomes infinite. Jeffrey's prior penalty plays the role to shrink the parameters toward zero in the case of separation.

Logistic regression models with various penalties have been considered for GWAS by a number of authors. Park and Hastie [22] considered logistic models with a L_2 -penalty. Wu et al. [23] considered logistic models with an L_1 -penalty. The LASSO-patternsearch

developed by Shi et al. [16] is also based on logistic regression models. However, the accuracy for identifying etiological SNPs was not fully addressed. Park and Hastie [22] introduced the L_2 -penalty mainly for computational reasons. Their method is essentially a classical stepwise procedure with AIC/BIC as model selection criteria. The method considered by Wu et al. [23] is in fact only a screening procedure. The numbers of main-effect and interaction features to be retained are predetermined and left as a subjective matter. The LASSO-patternsearch is closer to our approach. The procedure first screens the features by correlation screening based on single-feature (main-effect/interaction) models. Then a LASSO model is fitted to the retained features with its penalty parameter chosen by cross-validation. The features selected by LASSO are then refitted to a nonpenalized logistic regression model, and the coefficients of the features are subjected to hypothesis testing with varied level α . The α is again determined by cross-validation. By using cross-validation, this procedure addresses the prediction error of the selected model instead of the accuracy of the selected features. Our method is compared with the LASSO-patternsearch and the multiple test approach by simulation studies.

The two-stage penalized logistic regression approach is described in detail in Section 2. The approach is applied to a publically accessible CGEMS prostate cancer data in Section 3. Simulation studies are presented in Section 4. The paper is ended by some remarks. A supplementary document which contains some details omitted in the paper is provided at the website: <http://www.stat.nus.edu.sg/~stachenz/>, available online at doi: 10.1155/2012/642403.

2. The Two-Stage Penalized Logistic Regression Approach

We first give a brief account on the elements required in the approach: the logistic model for case-control study, the penalized likelihood, and the EBIC.

2.1. Logistic Model for Case-Control GWAS

Let y_i denote the disease status of individual i , 1 for case and 0 for control. Denote by x_{ij} , $j = 1, \dots, P$, the genotypes of individual i at the SNPs under study. The x_{ij} takes the value 0, 1, or 2, corresponding to the number of a particular allele in the genotype. Here, the additive genetic mode is assumed for all SNPs. The logistic model is as follows:

$$y_i \sim \text{Binomial}(1, \pi_i),$$

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \sum_{j=1}^P \beta_j x_{ij} + \sum_{j < k} \xi_{jk} x_{ij} x_{ik}, \quad i = 1, \dots, n, \quad (2.1)$$

where x_{ij} and $x_{ij}x_{ik}$ are referred to as main-effect and interaction features, respectively, hereafter. The validity of the logistic model for case-control experiments has been argued by Armitage [24] and Breslow and Day [25]. There are two fundamental facts about the above model for GWAS: (a) the number of features is much larger than the sample size n , since P is usually huge in GWAS, this situation is referred to as small- n -large- p ; (b) since there are only a few etiological SNPs, only a few of the coefficients in the model are nonzero, this phenomenon is referred to as sparsity.

2.2. Penalized Likelihood

Penalized likelihood makes the fitting of a logistic model with small- n -large- p computationally feasible. It also provides a mechanism for feature selection. Let s be the index set of a subset of the features. Let $L(\boldsymbol{\theta}(s) \mid s)$ denote the likelihood function of the logistic model consisting of features with indices in s , where $\boldsymbol{\theta}(s)$ consists of those β and ξ with their indices in s . The penalized log likelihood is defined as

$$l_p(\boldsymbol{\theta}(s) \mid \lambda) = -2 \log L(\boldsymbol{\theta}(s) \mid s) + \sum_{j \in s} p_\lambda(\theta_j), \quad (2.2)$$

where $p_\lambda(\cdot)$ is a penalty function and λ is called the penalty parameter. The following penalty functions are used in our approach:

$$\begin{aligned} L_1\text{-penalty : } p_\lambda(\theta_j) &= \lambda |\theta_j|, \\ \text{SCAD penalty : } p'_\lambda(|\theta|) &= \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\}, \end{aligned} \quad (2.3)$$

where a is a fixed constant bigger than 2. The penalized log likelihood with L_1 -penalty is used together with the tournament procedure [10] in the screening stage. At each application of the penalized likelihood, the parameter λ is tuned such that the minimization of the penalized likelihood yields a predetermined number of nonzero coefficients. The R package `glm` developed by Park and Hastie [26] is used for the computation, the tuning on λ is equivalent to setting the maximum steps in the `glm` function to the predetermined number of nonzero coefficients. The SCAD penalty is used in the second stage for ranking the features.

2.3. The Extended BIC

In small- n -large- p problems, the AIC and BIC are not selection consistent. To tackle the issue of feature selection in small- n -large- p problems, J. Chen and Z. Chen [13] developed a family of extended Bayes information criteria (EBIC). In the context of the logistic model described above, the EBIC is given as

$$\text{EBIC}(\gamma) = -2 \log L(\hat{\boldsymbol{\theta}}(s) \mid s) + (v_1 + v_2) \log(n) + 2\gamma \log \binom{P}{v_1} \binom{P(P-1)/2}{v_2}, \quad \gamma \geq 0, \quad (2.4)$$

where v_1 and v_2 are, respectively, the numbers of main-effect and interaction features and $\hat{\boldsymbol{\theta}}(s)$ is the maximum likelihood estimate of the parameter vector in the model. It has been shown that, under certain conditions, the EBIC is selection consistent when γ is larger than $1 - \ln n / (2 \ln p)$, see J. Chen and Z. Chen [13, 14]. The original BIC, which corresponds to the EBIC with $\gamma = 0$, fails to be selection consistent when p has a \sqrt{n} order.

We now describe the two-stage penalized logistic regression (TPLR) approach as follows.

2.4. Screening Stage

Let n_M and n_I be two predetermined numbers, respectively, for main-effect and interaction features to be retained. The screening stage consists of two steps: a main-effect screening step and an interaction screening step.

In the main-effect screening step, only the main-effect features are considered. Let S_M denote the index set of the main features, that is, $S_M = \{1, \dots, p\}$. If $|S_M|$, the number of members in S_M , is not too large, minimize

$$-2 \log L(\boldsymbol{\beta} \mid S_M) + \lambda \sum_{j=1}^p |\beta_j| \quad (2.5)$$

by tuning the value of λ to retain n_M features. If $|S_M|$ is very large. The following tournament procedure proposed in Z. Chen and J. Chen [10] is applied. Partition S_M into $S_M = \cup_k s_k$ with $|s_k|$ equal to an appropriate group size n_G , where n_G is chosen such that the minimization of the penalized likelihood with n_G features can be efficiently carried out. For each k , minimize

$$-2 \log L(\boldsymbol{\beta}(s_k) \mid s_k) + \lambda \sum_{j \in s_k} |\beta_j| \quad (2.6)$$

by tuning the value of λ to retain $n_k (\approx n_M)$ features. If $\sum_j n_j > n_G$, repeat the above process with all retained features; otherwise, apply the L_1 -penalized logistic model to the retained features to reduce the number to n_M . Let s_M denote the indices of these n_M features.

The interaction screening is similar to the main-effect screening step. However, the main-effect features retained in the main-effect screening step are built in the models for interaction screening. Let S_I denote the set of pairs of the indices for all the interaction features, that is, $S_I = \{(i, j) : i < j, i, j = 1, \dots, p\}$. Since $|S_I|$ is large in general, the tournament procedure is applied for interaction screening. Let S_I be partitioned as $S_I = \cup_k T_k$ with $|T_k| \approx n_G$. For each k , minimize

$$-2 \log L(\boldsymbol{\beta}(s_M), \boldsymbol{\xi}(T_k) \mid s_M, T_k) + \lambda \sum_{(i,j) \in T_k} |\xi_{ij}| \quad (2.7)$$

by tuning the value of λ to retain $n_k (\approx n_I)$ interaction features. Note that, in the above penalized likelihood, both the main-effect features in s_M and the interaction features in T_k are involved in the likelihood part. However, only the parameters associated with the interaction features are penalized. Since no penalty is put on the parameters associated with the main-effect features, the main-effect features are always retained in the process of interaction screening. If $\sum_k n_k > n_G$, repeat the above process with the set of all retained features; otherwise, reduce the retained features to n_I of them by one run of the minimization using an L_1 -penalized likelihood.

2.5. Selection Stage

The selection stage consists of a ranking step and a model selection step. In the ranking step, the retained features (main-effect and interaction) are ranked together by a penalized

likelihood with SCAD penalty plus an additional Jeffrey's prior penalty. In the model selection step, a sequence of nested models are formed and evaluated by the EBIC.

For convenience, let the retained interaction features be referred to by a single index. Let S^* be the index set of all the main-effect and interaction features retained in the screening stage. Let $K = |S^*|$. Denote by $\boldsymbol{\theta}(S^*)$ the vector of coefficients corresponding to these features (the components of $\boldsymbol{\theta}(S^*)$ are the β 's and ξ 's corresponding to the retained main-effects and interactions). Jeffrey's prior penalty is the log determinant of the Fisher information matrix. Thus the penalized likelihood in the selection stage is given by

$$l_p(\boldsymbol{\theta}(S^*) | \lambda) = -2 \log L(\boldsymbol{\theta}(S^*) | S^*) - \log |I(\boldsymbol{\theta}(S^*))| + \sum_{j \in S^*} p_\lambda(|\theta_j|), \quad (2.8)$$

where p_λ is the SCAD penalty and $I(\boldsymbol{\theta}(S^*))$ is the Fisher information matrix. The ranking step is done as follows. The parameter λ is tuned to a value λ_1 such that it is the smallest to make at least one component of $\boldsymbol{\beta}(S^*)$ zero by minimizing $l_p(\boldsymbol{\beta}(S^*) | \lambda_1)$. Let $j_K \in S^*$ be the index corresponding to the zero component. Update S^* to $S^*/\{j_K\}$, that is, the feature with index j_K is eliminated from further consideration. With the updated S^* , the above process is repeated, and another feature is eliminated. Continuing this way, eventually, we obtain an ordered sequence of the indices in S^* : j_1, j_2, \dots, j_K . From the ordered sequence above, a sequence of nested models is formed as $S_k = \{j_1, \dots, j_k\}, k = 1, \dots, K$. For each S_k , the un-penalized likelihood $\log L(\boldsymbol{\theta}(S_k) | S_k)$ is maximized. The EBIC with γ values in a range $[1 - \ln n / \ln p, \gamma_{\max}]$ is computed for all these models. For each γ , the model with the smallest EBIC(γ) is identified. The upper bound of the range, γ_{\max} , is taken as a value such that no feature can be selected by the EBIC with that value. Only a few models can be identified when γ varies in the range. Each identified model corresponds to a subinterval of $[1 - \ln n / \ln p, \gamma_{\max}]$. The identified models together with their corresponding subintervals are then reported.

The choice of γ in the EBIC affects the positive discovery rate (PDR) and the false discovery rate (FDR). In the context of GWAS, the PDR is the proportion of correctly identified SNPs among all etiological SNPs, and the FDR is the proportion of incorrectly identified SNPs among all identified SNPs. A larger γ results in a smaller FDR and also a lower PDR. A smaller γ results in a higher PDR and also a higher FDR. A balance must be stricken between PDR and FDR according to the purpose of the study. If the purpose is to confirm the etiological effect of certain well-studied loci or regions, one should emphasize more on a desirably low FDR rather than a high PDR. If the purpose is to discover candidate loci or regions for further study, one should emphasize more on a high PDR with only a reasonable FDR. The FDR is related to the statistical significance of the features. Measures on the statistical significance can be obtained from the final identified models and their corresponding subintervals. The upper bound of the subinterval determines the largest threshold which the effects of the features in the model must exceed. Likelihood ratio test (LRT) statistics can be used to assess the significance of the feature effects. For example, suppose a model consisting of ν_1 main-effect features and ν_2 interaction features is selected with γ in a sub-interval $(\underline{\gamma}, \bar{\gamma}]$. The LRT statistic for the significance of the feature with the lowest rank in the model must exceed the threshold $\log n + 2\bar{\gamma} \log((P - \nu_1 + 1)/\nu_1)$, if the feature is a main-effect one, and $\log n + 2\bar{\gamma} \log((P(P - 1)/2 - \nu_2 + 1)/\nu_2)$, if the feature is an interaction one. The probability for the LRT to exceed the threshold is at most $Pr(\chi_1^2 > \log n + 2\bar{\gamma} \log((P - \nu_1 + 1)/\nu_1))$ or $Pr(\chi_1^2 > \log n + 2\bar{\gamma} \log((P(P - 1)/2 - \nu_2 + 1)/\nu_2))$ for a main-effect

or interaction feature if the feature does not actually have any effect. These probabilities, like the P -values in classical hypothesis testing, provide statistical basis for the user to determine which model should be taken as the selected model.

A final issue on the two-stage logistic regression procedure is how to determine n_M and n_I . If they are large enough, usually several times of the actual numbers, their choice will not affect the final model selection. Since the actual numbers are unknown, a strategy is to consider several different n_M and n_I . First, run the procedure with some educated guess on n_M and n_I . Then, run the procedure again using larger n_M and n_I . If the identified models by using these n_M and n_I are almost the same, the choice of n_M and n_I is appropriate. Otherwise, further values of n_M and n_I should be considered, until eventually different n_M, n_I result in the same results.

3. Analysis of CGEMS Prostate Cancer Data

The CGEMS data portal of National Cancer Institute, USA, provides public access to the summary results of approximately 550,000 SNPs genotyped in the CGEMS prostate cancer whole genome scan, see <http://cgems.cancer.gov>. We applied the two-stage penalized regression approach to the prostate cancer Phase 1A data in the prostate, lung, colon, and ovarian (PLCO) cancer screening trial. The dataset contains 294,179 autosomal SNPs which passed the quality controls on 1,111 controls and 1,148 cases (673 cases are aggressive, Gleason ≥ 7 or stage \geq III; 475 cases are nonaggressive, Gleason < 7 and stage $<$ III). In our analysis, we put all the cases together without distinguishing aggressive and non-aggressive ones. We assumed additive genetic mode for all the SNPs.

The application of the screening stage to all the 294,179 SNPs directly is not only time consuming but also unnecessary. Therefore, we did a preliminary screening by using single-SNP logistic models. For each SNP, a logistic model is fitted and the P -value of the significance test of the SNP effect is obtained. Those SNPs with a P -value bigger than 0.05 are discarded. There are 17,387 SNPs which have a P -value less than 0.05 and are retained.

Because of the sheer huge number of features, 17,387 main features and $17,387 \times (1,7387-1)/2$ interaction features, the tournament procedure is applied in the screening stage. At the main-effect feature screening step, the main-effect features are randomly partitioned into groups of size 1,000, except one group of size 1,387, and 100 features are selected from each group. A second round of screening is applied to the selected 1,700 features out of which 100 features are retained. The interaction feature screening is applied to $17,387 \times (1,7387-1)/2$ interaction features. Each round, the retained features are partitioned into groups of size 1,000, and 50 features are selected from each group. The procedure continues until 300 interaction features are finally selected. Eventually, the 100 main-effect features and 300 interaction features are put together and screened to retain a total of 100 features (main-effect or interaction). The eventual 100 features are then subjected to the selection procedure.

The features selected by EBIC with γ in the subintervals $(0.70, 0.73]$, $(0.73, 0.77]$, and $(0.77, 0.8]$ are given in Table 1. With $\gamma = 0.8$, the largest value at which at least one feature can be selected, the following three interaction features are selected: rs1885693-rs12537363, rs7837688-rs2256142 and rs1721525-rs2243988. The effects of these features have a significance level at least $1.8250e-11$. The next largest γ value, 0.77, selects 7 additional interaction features which have a significance level at least $9.6435e-11$. The third largest γ value, 0.73, selects still 2 additional interaction features which have a significance level at least $2.7407e-10$. The chromosomal region 8q24 is the one where many previous prostate cancer studies are concentrated. It has been reported in a number of studies that rs1447295, one of the 4 tightly

Table 1: Features associated with prostate cancer from the analysis of CGEMS data (“rsXXX” denotes SNP reference).

Chromosome	Feature	Maximum γ	Significance Level
6, 7	rs1885693-rs12537363	0.80	1.824985e-11
8, 13	rs7837688 -rs2256142	0.80	1.824985e-11
1, 21	rs1721525-rs2243988	0.80	1.824985e-11
10, 16	rs11595532-rs8055313	0.77	9.64352e-11
12, 12	rs10842794-rs10848967	0.77	9.64352e-11
9, 12	rs3802357-rs10880221	0.77	9.64352e-11
1, 2	rs3900628-rs642501	0.77	9.64352e-11
1, 16	rs10518441-rs2663158	0.77	9.64352e-11
3, 13	rs1880589-rs1999494	0.77	9.64352e-11
5, 18	rs6883810-rs11874224	0.77	9.64352e-11
13, 19	rs4274307-rs3745180	0.73	2.740672e-10
5, 19	rs672413-rs3915790	0.73	2.740672e-10

linked SNPs in the “locus 1” region of 8q24, is associated with prostate cancer, and it has been established as a benchmark for prostate cancer association studies. In the current data set, we found that rs7837688 is highly correlated with rs1447295 ($r^2 = 0.9$) and is more significant than rs1447295 based on single-SNP models. These two SNPs, which are in the same recombination block, are also physically close.

An older and slightly different version of the CGEMS prostate data has been analyzed by Yeager et al. [27] using single-SNP multiple testing approach. In their analysis, they distinguished between aggressive and non-aggressive status and assumed no structure on genetic modes. For each SNP, they considered four tests: a χ^2 -test with 4 degrees of freedom based on a 3×3 contingency table, a score test with 4 degrees of freedom based on a polytomous logistic regression model adjusted for age group, region of recruitment, and whether a case is diagnosed within one year of entry to the trial, as well as the other two which are the same as the χ^2 and score tests but take into account incidence-density sampling. They identified two physically close but genetically independent regions (in a distance 0.65 centi-Morgans) within 8q24. One of the regions is where the benchmark SNP rs1447295 is located. They reported three SNPs: rs1447295 (P -value: $9.75e-05$), rs7837688 (P -value: $6.52e-06$) and rs6983267 (P -value: $2.43e-05$), where rs7837688 is in the same region as rs1447295 and rs6983267 is in the other region. The P -values are computed from the score statistic based on incidence-density sampling polytomous logistic regression model adjusted for other covariates.

In our analysis, we identified rs7837688 but not rs1447295. This is because the penalized likelihood tends to select only one feature among several highly correlated features, which is a contrast to the multiple testing that selects all the correlated features if any of them is associated with the disease status. We failed to identify rs6983267. The possible reason could be that its effect is masked by other more significant features which are identified in our analysis. We also carried out the selection procedure with only the 100 main-effect features retained from the screening stage. It is found that rs6983267 is among the top 20 selected main-effect features with a significance level $2.3278e-05$. It is interesting to notice that the two SNPs rs7837688 and rs1721525 appearing in the top three interaction features are also among the top four features selected with a maximum γ value 0.7185 when only

main-effect features are considered. Since no SNP on chromosomes other than 8q24 has been reported in other studies, we wonder whether statistically significant SNPs on other chromosomes can be ignored due to biological reasons: if not, our analysis strongly suggests that rs1721525 located on chromosome 1 could represent another region in the genome which is associated with prostate cancer, if it holds, biologically, chromosome 1 cannot be excluded in the consideration of genetic variants for prostate cancer.

4. Simulation Studies

We present results of two simulation studies in this section. In the first study, we compare the two-stage penalized logistic regression (TPLR) approach with the paired-SNP multiple testing (PMT) approach of Marchini et al. [5] under simulation settings considered by them. In the second study, we compare the TPLR approach with LASSO-patternsearch using a data structure mimicking the CGEMS prostate cancer data.

4.1. Simulation Study 1

The comparison of TPLR and PMT is based on four models. Each model involves two etiological SNPs. In the first model, the effects of the two SNPs are multiplicative both within and between loci; in the second model, the effects of the two SNPs are multiplicative within but not between loci; in the third model, the two SNPs have threshold interaction effects; in the fourth model, the two SNPs have an interaction effect but no marginal effects. The first three models are taken from Marchini et al. [5]. The details of these models are provided in the supplementary document.

Marchini et al. [5] considered two strategies of PMT. In the first strategy, a logistic model with 9 parameters is fitted for each pair of SNPs, and the Bonferroni corrected significance level $\alpha / \binom{P}{2}$ is used to declare the significant pairs. In the second strategy, the SNPs that are significant in single-SNP tests at a liberal level α_1 are identified, then the significances of all the pairs formed by these SNPs are tested using the Bonferroni corrected level $\alpha / \binom{P\alpha_1}{2}$.

In the first three models, the marginal effects of both loci are nonnegligible and can be picked up by the single-SNP tests at the relaxed significance level. In this situation, the second strategy has an advantage over the first strategy in terms of detection power and false discovery rate. In this study, we compare our approach with the second strategy of PMT under the first three models. In the fourth model, since there are no marginal effects at both loci, the second strategy of PMT cannot be applied since it will fail to pick up any loci at the first step. Hence, we compare our approach with the first strategy of PMT. However, the first strategy involves a stupendous amount of computation which exceeds our computing capacity. To circumvent this dilemma, we consider an artificial version of the first strategy; that is, we only consider the pairs which involve at least one of the etiological SNPs. This artificial version has the same detection power but lower false discovery rate than the full version. The artificial version cannot be implemented with real data since it requires the knowledge of the etiological SNPs. However, it can be implemented with simulated data and serves the purpose of comparison.

Each simulated dataset contains $n = 800$ individuals (400 cases and 400 controls) with genotypes of P SNPs. Two values of P , 1000 and 5000, are considered. The genotypes of disease loci, which are not among the P SNPs, and the disease status of the individuals are generated first. Then, the genotypes of the SNPs which are in linkage disequilibrium with the

disease loci are generated using a square correlation coefficient $r^2 = 0.5$. The genotypes of the remaining SNPs are generated independently assuming Hardy-Weinberg equilibrium. For the first three models, the effects of the disease loci are specified by the prevalence, disease allele frequencies, denoted by q , and marginal effect parameters, denoted by λ_1 and λ_2 . The prevalence is set at 0.01 throughout. The two marginal effects are set equal, that is, $\lambda_1 = \lambda_2 = \lambda$. For the fourth model, the effect is specified through the coefficient in the logistic model. The coefficients are determined by first specifying ξ_{12} and then determining β_1 and β_2 through the constraints of the model while β_0 is set to -5 . The definition of these parameters and the details of the data generation are given in the supplementary document.

The α_1 and α in the PMT approach are taken to be 0.1 and 0.05, respectively, the same as in Marchini et al. [5]. The γ in EBIC is fixed as 1 since it is infeasible to incorporate the consideration on the choice of γ into the simulation study. The average PDR and FDR over 200 simulation replicates under Model 1–4 are given in Tables 2–5, respectively. In Table 5, the entries of the FDR for the PMT approach are lower bounds rather than the actual FDRs, since, as mentioned earlier, only the pairs of SNPs involving at least one etiological SNP are considered in the artificial version of the first strategy of PMT, which results in less false discoveries than the full version while retaining the same positive detections.

The results presented in Tables 2–5 are summarized as follows. Under Model 1, TPLR has much lower FDR and comparable PDR compared with PMT. Under Models 2–4, the PDR of TPLR is significantly higher than PMT in all cases except Model 2 when $\lambda = 0.7, q = 0.2, P = 1000$ (0.95 versus 1) and Model 3 when $\lambda = 1, q = 0.1, P = 1000$ (0.81 versus 0.84). The overall averaged FDRs of TPLR is 0.0487 while that of PMT is 0.7604. It is seen that the FDR of TPLR is always kept at reasonably low levels but that of PMT is intolerably high, and at the same time TPLR is still more powerful than PMT for detecting etiological SNPs. From the simulation results, we can also see the impact of P on PDR and FDR. In general, the increase of P reduces PDR and increases FDR of both approaches. However, the impact on TPLR is less than that on PMT.

4.2. Simulation Study 2

The data for this simulation study is generated mimicking the structure of the CGEMS prostate cancer data. The cases and controls are generated using a logistic model with the following linear predictor:

$$\eta = \beta_0 + \sum_{j=1}^5 \beta_j x_j + \xi_1 x_6 x_7 + \xi_2 x_8 x_9 + \xi_3 x_{10} x_{11} + \xi_4 x_{12} x_{13} + \xi_5 x_{13} x_{14}, \quad (4.1)$$

where x_j 's are feature values of 14 SNPs. The parameter values are taken as

$$\begin{aligned} \beta &= (-8.65, 0.89, 1.1, 0.74, 1.18, 1.25), \\ \xi &= (1.95, 1.62, 1.9, 1.8, 1.1). \end{aligned} \quad (4.2)$$

Table 2: The simulated average PDR and FDR under Model 1: multiplicative effects both within and between loci.

(n, P)	λ	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	0.8	0.1	0.610	0.780	0.358	0.996
	0.9	0.1	0.850	0.900	0.320	0.998
	1.0	0.1	0.960	1.000	0.219	0.999
(800,5000)	0.8	0.1	0.470	0.660	0.405	0.999
	0.9	0.1	0.750	0.870	0.380	0.999
	1.0	0.1	0.890	0.930	0.233	0.999

Table 3: The simulated average PDR and FDR under Model 2: multiplicative effects within loci but not between loci.

(n, P)	λ	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	0.5	0.1	0.265	0.175	0.086	0.352
	0.5	0.2	0.650	0.550	0.071	0.763
	0.7	0.1	0.790	0.710	0.048	0.758
	0.7	0.2	0.950	1.000	0.050	0.954
(800,5000)	0.5	0.1	0.175	0.085	0.079	0.595
	0.5	0.2	0.610	0.405	0.077	0.928
	0.7	0.1	0.720	0.480	0.062	0.776
	0.7	0.2	0.940	0.930	0.051	0.980

The SNPs in the above model mimic the 14 SNPs involved in the top 5 main-effect features and top 5 interaction features of the CGEMS prostate cancer data. The minor allele frequencies (MAF) of the SNPs, which are estimated from the prostate cancer data, are given as follows:

$$\text{MAF} = (0.31, 0.12, 0.29, 0.12, 0.13, 0.13, 0.47, 0.18, 0.29, 0.16, 0.04, 0.12, 0.36, 0.40). \quad (4.3)$$

The genotypes of these 14 SNPs are generated by using the MAF, assuming Hardy-Weinberg Equilibrium. In addition to these 14 SNPs, 20,000 noncausal SNPs are randomly selected (without replacement) from the 294,179 SNPs of the prostate cancer data in each simulation replicate. For each simulation replicate, 1,000 cases and 1,000 controls are generated. They are matched by randomly selected (without replacement) individuals from the prostate cancer data. Their genotypes at the 20,000 noncausal SNPs are taken the same as those in the prostate cancer data.

In the TPLR approach, 50 main effect features and 50 interaction features are selected in the screening stage using the tournament screening strategy. In the selection stage, EBIC(γ) values are calculated for the nested models with γ in the range $0(0.1)2$, that is, from 0 to 2 in space of 0.1.

In the LASSO-patternsearch approach, at the screening stage, 0.05 and 0.002 are used as thresholds for the P -values of the main-effect features and interaction features, respectively. At the LASSO selection step, a 5-fold cross-validation is used for the choice of penalty parameter. At the hypothesis testing step, 9 α levels are considered, that is,

Table 4: The simulated average PDR and FDR under Model 3: two-locus threshold interaction effects.

(n, P)	λ	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	0.8	0.1	0.530	0.455	0.086	0.884
	0.9	0.1	0.730	0.695	0.052	0.965
	1.0	0.1	0.810	0.840	0.047	0.970
(800,5000)	0.8	0.1	0.350	0.270	0.028	0.800
	0.9	0.1	0.620	0.490	0.101	0.999
	1.0	0.1	0.712	0.657	0.060	0.982

Table 5: The simulated average PDR and FDR under Model 4: significant interaction effect but zero marginal effects.

(n, P)	ξ_{12}	q	PDR		FDR	
			TPLR	MT	TPLR	MT
(800,1000)	1.9	0.1	0.828	0.702	0.012	≥ 0.550
	2.0	0.1	0.945	0.860	0.026	≥ 0.641
	2.1	0.1	0.965	0.915	0.015	≥ 0.915
(800,5000)	1.9	0.1	0.555	0.460	0.009	≥ 0.406
	2.0	0.1	0.730	0.710	0.014	≥ 0.427
	2.1	0.1	0.885	0.795	0.006	≥ 0.562

$\alpha = 10^{-k}, k = 0, 1, \dots, 8$. The case $\alpha = 1$ amounts to stopping the procedure at the LASSO selection step.

Since in the TPLR approach there is not a definite choice of γ , to facilitate the comparison, we calculate PDR and FDR for each fixed γ value in the TPLR approach, and for each fixed α level in LASSO-patternsearch. The PDR and FDR are calculated separately for the detection of true main-effect and interaction features. They are also calculated for the detection of causal SNPs. A causal SNP is considered positively discovered if it is selected either as a main-effect feature or a constituent in an interaction feature. The simulated FDR and PDR over 100 replicates of TPLR with $n_M = n_I = 50$ and $\gamma = 0(0.2)2$ and those of LASSO-patternsearch with $\alpha = 10^{-k}, k = 0, 1, \dots, 8$ are reported in Table 6. It is actually the γ values in the higher end and α levels in the lower end that will be involved in the final selection. The comparison of the results with those values is more relevant. As shown by the bold digits in Table 6, TPLR has higher PDR and lower FDR than LASSO-patternsearch across-the-board. For the main-effect features, the lowest FDR of TPLR is 0.006 while it achieves PDR around 0.65, but the lowest FDR of LASSO-patternsearch is around 0.2 while it only achieves PDR around 0.6. The FDR and PDR on interaction features and causal SNPs have the same pattern. When the two approaches have about the same PDR, the LASSO-patternsearch has a much larger undesirable FDR than TPLR. For example, on the SNPs, when the PDR is 0.608 for TPLR and 0.609 for LASSO-patternsearch, the FDRs are, respectively, 0.041 and 0.654; on the main-effect features, when the PDR is 0.646 for both TPLR and LASSO-patternsearch, the FDRs are, respectively, 0.006 and 0.220. The ROC curves of the two approaches in identifying etiological SNPs are plotted in Figure 1. Figure 1 shows clearly that the PDR of TPLR is much higher than the PDR of LASSO-patternsearch when FDR is the same, which is true uniformly over FDR.

Table 6: Comparison of TPLR approach and LASSO-patternsearch (the PDR and FDR with subscript M , I , and S indicate the rates calculated for main-effect features, interaction features, and SNPs resp.)

γ	TPLR Approach					
	PDR _M	FDR _M	PDR _I	FDR _I	PDR _S	FDR _S
0.0	0.964	0.902	0.884	0.907	0.947	0.855
0.2	0.768	0.926	0.884	0.900	0.947	0.852
0.4	0.714	0.505	0.748	0.677	0.803	0.494
0.6	0.692	0.199	0.680	0.413	0.730	0.200
0.8	0.684	0.112	0.642	0.331	0.691	0.126
1.0	0.672	0.023	0.610	0.267	0.654	0.073
1.2	0.668	0.021	0.594	0.237	0.638	0.057
1.4	0.658	0.009	0.566	0.201	0.608	0.041
1.6	0.654	0.006	0.536	0.165	0.578	0.026
1.8	0.646	0.006	0.524	0.144	0.563	0.021
2.0	0.630	0.006	0.492	0.109	0.526	0.015
α	LASSO-patternsearch					
	PDR _M	FDR _M	PDR _I	FDR _I	PDR _S	FDR _S
10 ⁻⁰	0.882	0.445	0.710	0.967	0.847	0.940
10 ⁻¹	0.816	0.332	0.696	0.957	0.827	0.926
10 ⁻²	0.786	0.283	0.664	0.929	0.774	0.885
10 ⁻³	0.718	0.241	0.618	0.869	0.694	0.802
10 ⁻⁴	0.646	0.220	0.556	0.752	0.609	0.654
10 ⁻⁵	0.578	0.193	0.486	0.563	0.531	0.444
10 ⁻⁶	0.504	0.184	0.414	0.355	0.453	0.254
10 ⁻⁷	0.400	0.190	0.360	0.196	0.383	0.130
10 ⁻⁸	0.332	0.202	0.292	0.076	0.316	0.047

To investigate the effect of the choice of n_M and n_I , we considered $n_M = n_I = 15, 25$, and 50 which are 3, 5, and 10 times of the actual number of causal features, respectively. The simulation results show that, though there is a slight difference between the choice of 15 and the other two choices, there is no substantial difference between the choice of 25 and 50. This justifies the strategy given at the end of Section 2. The detailed results on the comparison of the choices are given in the supplementary document.

We also investigated whether the ranking step in the TPLR approach really reflects the actual importance of the features. The average ranks of the ten causal features over the 100 simulation replicates are given in Table 7.

On the average, the causal features are all among the top ten ranks. This gives a justification for the ranking step in the selection stage of the TPLR approach.

5. Some Remarks

It is a common understanding that individual SNPs are unlikely to play an important role in the development of complex diseases, and, instead, it is the interactions of many SNPs that are behind disease developments, see Garte [28]. The finding that only interaction features are selected (since they are more significant than main-effect features) in our analysis provides some evidence to this understanding. Perhaps, even higher-order interactions should be

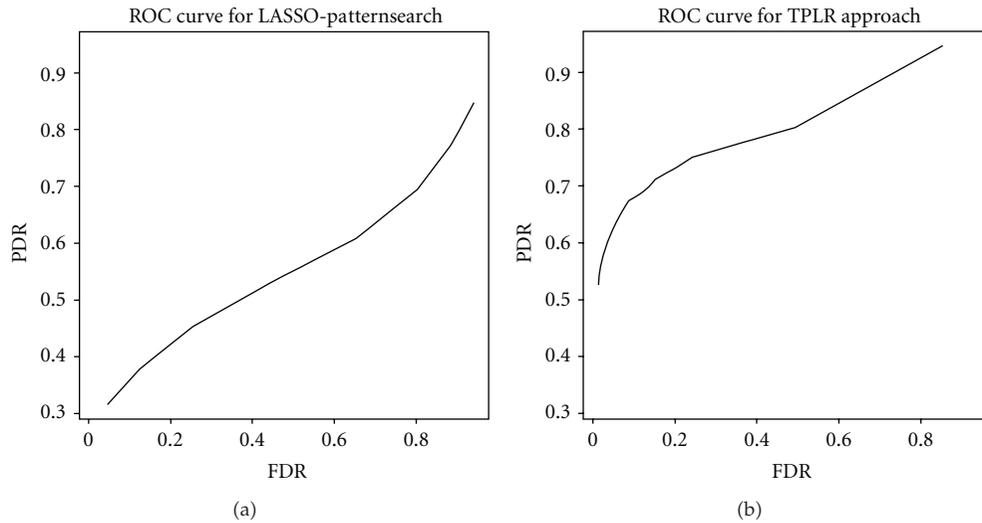


Figure 1: The ROC curves of the LASSO-patternsearch and the TPLR approach for identifying etiological SNPs.

Table 7

Features	1	2	3	4	5	(6,7)	(8,9)	(10,11)	(12,13)	(14,15)
Avg. ranks	4.7	2.0	7.2	6.1	5.4	7.6	6.8	9.2	3.0	1.1

investigated. This makes methods such as the penalized logistic regression which can deal with interactions even more desirable.

The analysis of the CGEMS prostate cancer data can be refined by replacing the binary logistic model with a polytomous logistic regression model taking into account that the genetic mechanisms behind aggressive and nonaggressive prostate cancers might be different. Accordingly, the penalty in the penalized likelihood can be replaced by some variants of the group LASSO penalty considered by Huang et al. [29]. A polytomous logistic regression model with an appropriate penalty function is of general interest in feature selection with multinomial responses, which will be pursued elsewhere.

Acknowledgments

The authors would like to thank the National Cancer Institute of USA for granting the access to the CGEMS prostate cancer data. The research of the authors is supported by Research Grant R-155-000-065-112 of the National University of Singapore, and the research of the first author was done when she was a Ph.D. student at the National University of Singapore.

References

- [1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software, Belmont, Calif, USA, 1984.
- [2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [3] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

- [4] H. Schwender and K. Ickstadt, "Identification of SNP interactions using logic regression," *Biostatistics*, vol. 9, no. 1, pp. 187–198, 2008.
- [5] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genetics*, vol. 37, no. 4, pp. 413–417, 2005.
- [6] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.
- [7] B. Efron and R. Tibshirani, "Empirical Bayes methods and false discovery rates for microarrays," *Genetic Epidemiology*, vol. 23, no. 1, pp. 70–86, 2002.
- [8] J. D. Storey and R. Tibshirani, "Statistical Methods for Identifying Differentially Expressed Genes in DNA Microarrays," *Functional Genomics: Methods in Molecular Biology*, vol. 224, pp. 149–157, 1993.
- [9] J. Hoh and J. Ott, "Mathematical multi-locus approaches to localizing complex human trait genes," *Nature Reviews Genetics*, vol. 4, no. 9, pp. 701–709, 2003.
- [10] Z. Chen and J. Chen, "Tournament screening cum EBIC for feature selection with high-dimensional feature spaces," *Science in China. Series A*, vol. 52, no. 6, pp. 1327–1341, 2009.
- [11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [12] D. Firth, "Bias reduction of maximum likelihood estimates," *Biometrika*, vol. 80, no. 1, pp. 27–38, 1993.
- [13] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.
- [14] J. Chen and Z. Chen, "Extended BIC for small- n -large- P sparse GLM," *Statistica Sinica*. In press.
- [15] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," *The Annals of Statistics*, vol. 36, no. 6, pp. 2605–2637, 2008.
- [16] W. Shi, K. E. Lee, and G. Wahba, "Detecting disease-causing genes by LASSO-Patternsearch algorithm," *BMC Proceedings*, vol. 1, supplement 1, p. S60, 2007.
- [17] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society. Series B*, vol. 70, no. 5, pp. 849–911, 2008.
- [18] K. K. Wu, *Comparison of sure independence screening and tournament screening for feature selection with ultra-high dimensional feature space*, Honor's thesis, Department of Statistics & Applied Probability, National University of Singapore, 2010.
- [19] W. L. H. Koh, *The comparison of two-stage feature selection methods in small- n -large- p problems*, Honor's thesis, Department of Statistics & Applied Probability, National University of Singapore, 2011.
- [20] J. Zhao, *Model selection methods and their applications in genome-wide association studies*, Ph.D. thesis, Department of Statistics and Applied Probability, National University of Singapore, 2008.
- [21] A. Albert and J. A. Anderson, "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, vol. 71, no. 1, pp. 1–10, 1984.
- [22] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, vol. 9, no. 1, pp. 30–50, 2008.
- [23] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.
- [24] P. Armitage, *Statistical Methods in Medical Research*, Blackwell, Oxford, UK, 1971.
- [25] N. Breslow and N. E. Day, *Statistical Methods in Cancer Research*, vol. 1 of *The Analysis of Case-Control Studies*, International Agency for Research on Cancer Scientific Publications, Lyon, France, 1980.
- [26] M. Y. Park and T. Hastie, "An L_1 regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society. Series B*, vol. 69, no. 4, pp. 659–677, 2007.
- [27] M. Yeager, N. Orr, R. B. Hayes et al., "Genome-wide association study of prostate cancer identifies a second risk locus at 8q24," *Nature Genetics*, vol. 39, no. 5, pp. 645–649, 2007.
- [28] S. Garte, "Metabolic susceptibility genes as cancer risk factors: time for a reassessment?" *Cancer Epidemiology Biomarkers and Prevention*, vol. 10, no. 12, pp. 1233–1237, 2001.
- [29] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, "A group bridge approach for variable selection," *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

