Web-based Supplementary Materials for

A two-stage penalized logistic regression approach to case-control genome-wide association studies

by Jingyuan Zhao and Zehua Chen

1. A pseudo-algorithm

The following is a pseudo-algorithm for the implementation of the two-stage penalized logistic regression procedure:

Initialization:

Input $\boldsymbol{y}$ — the vector of disease status of individuals and $\boldsymbol{X}$ — the matrix of coded SNP genotypes. Specify procedure parameters: n_G — group size in tournament screening, n_M — number of main-effect features to be retained, n_I — number of interaction features to be retained, γ — a vector of values for the γ parameter in EBIC. Generate the index vector S for the SNPs.

Screening Stage:

Main-effect Screening: Partition S into $S = \cup_k s_k$ with $|s_k| \approx n_G$. For each k, minimize

$$-2 \log L(X(s_k)\boldsymbol{\beta}(s_k)) + \lambda \sum_{k \in s_j} |\beta_j|$$

by tuning the value of λ to retain $n_j (\approx n_M)$ features. Here $X(s_j)$ denotes the columns of X with column numbers in s_j. Similarly for the notation $\boldsymbol{\beta}(s_j)$. If $\sum_j n_j > n_G$, repeat the above process with all retained features;

otherwise, using L_1-penalized logistic model again, reduce the retained features to n_M. Let s_M denote the indices of these n_M features.

Interaction Screening: Form interaction features in groups of size about n_G. Let $Y(s_j, s_l)$ denote the products of the columns of X with indices in s_j and s_l . For each (j, l) minimize

$$-2\log L(X(s_M)\boldsymbol{\beta}(s_M), Y(s_j, s_l)\boldsymbol{\xi}(s_j, s_l)) + \lambda \sum_{k\in s_j, m\in s_l} |\xi_{km}|$$

by tuning the value of λ to retain $n_{jl}(\approx n_I)$ features. If $\sum_{j,l} n_{jl} > n_G$, repeat the above process with all retained features; otherwise, using L_1-penalized logistic model again, reduce the retained interaction features to n_I.

Selection Stage:

Ranking: Using a logistic model penalized by SCAD and Jeffrey's prior penalty with the n_M main-effect features and n_I interaction features, delete the features one at a time by gradually tuning the penalty parameters. Rank the features according to their inverse order of being deleted.

Model Selection: Form a sequence of nested models according to the ranks of the features. Compute EBIC with values in γ for each model. For each value of γ, select the model with the smallest EBIC.

Output:

Report the selected models and their corresponding γ values.

2. **Genetic models**

In the following, the details of the genetic models considered in simulation study 1 are provided.

Model 1: Multiplicative effects both within and between loci.

	aa	Aa	AA
bb	α	$\alpha(1+\theta_A)$	$\alpha(1+\theta_A)^2$
Bb	$\alpha(1+\theta_B)$	$\alpha(1+\theta_A)(1+\theta_B)$	$\alpha(1+\theta_A)^2(1+\theta_B)$
BB	$\alpha(1+\theta_B)^2$	$\alpha(1+\theta_A)(1+\theta_B)^2$	$\alpha(1+\theta_A)^2(1+\theta_B)^2$

The entries of the table are conditional disease odds given the genotypes of the two loci, e.g., $p(D|\text{aa, bb})/p(\bar{D}|\text{aa, bb}) = \alpha$, etc. The log odds of this model can be expressed as $\eta(x_A, x_B) = \log(\alpha) + \log(1+\theta_A)x_A + \log(1+\theta_B)x_B$, where x_A and x_B are the number of disease allele at locus A and B respectively. At the log scale, both the allele effects and the locus effects are additive in this model.

Model 2: Multiplicative effects within loci but not between loci.

	aa	Aa	AA
bb	α	α	α
Bb	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
BB	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

The log odds of this model can be expressed as $\eta(x_A, x_B) = \log(\alpha) + \log(1+\theta)x_A x_B$. In this model, the allele effects at both loci are additive at the log scale, but it is not the case for the locus effects; that is, there is an interaction between the two loci at the log scale.

Model 3: Two-locus threshold interaction effects.

	aa	Aa	AA
bb	α	α	α
Bb	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
BB	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$

In this model, as long as a disease allele either at locus A or locus B is present, the disease odds will change. However, unlike in the previous models, the increase of the number of disease alleles at both loci will not increase the risk of disease. The log odds cannot be expressed in terms of the number of disease alleles. Instead, it can be expressed as $\eta(\delta_{A1}, \delta_{A2}, \delta_{B1}, \delta_{B2}) = \log(\alpha) + \log(1+\theta)(\delta_{A1} + \delta_{A2} + \delta_{B1} + \delta_{B2} - \delta_{A1}\delta_{B1} - \delta_{A1}\delta_{B2} - \delta_{A2}\delta_{B1} - \delta_{A2}\delta_{B2})$, where δ_{A1}, δ_{A2}, δ_{B1}, δ_{B2} are the indicators of the genotypes Aa, AA, Bb, BB respectively. In this model, neither the allele effects nor the locus effects are additive at the log scale.

In addition to the above three models, we also consider a model where there is an interaction effect between two loci but the marginal effects of both loci are zero. The model is specified in terms of log odds as follows.

Model 4:

$$\eta(x_A, x_B) = \beta_0 + \beta_1 x_A + \beta_2 x_B + \xi_{12} x_A x_B$$

subject to that both $\sum_{x_B} \eta(x_A, x_B) p(x_B)$ and $\sum_{x_A} \eta(x_A, x_B) p(x_A)$ are constants.

3. Data generation details for simulation study 1

In all settings of the simulation study, the disease loci are assumed unlinked. The genotypes of the disease loci are generated at random with specified disease allele frequencies assuming HWE. The effects of the disease loci under Model 1 — 3 are specified by three parameters: the prevalence p, the marginal effects λ_A and λ_B. The definition of these parameters are given below.

Let g_A and g_B denote the genotypes at locus A and B. The prevalence is defined as

$$p = \sum_{g_A, g_B} p(D|g_A, g_B)p(g_A, g_B). \tag{1}$$

The marginal effect parameter λ_A is defined as

$$\lambda_A = \frac{p(D|Aa)}{p(\bar{D}|Aa)} / \frac{p(D|aa)}{p(\bar{D}|aa)} - 1. \tag{2}$$

Similarly, the marginal effect parameter λ_B is defined as

$$\lambda_B = \frac{p(D|Bb)}{p(\bar{D}|Bb)} / \frac{p(D|bb)}{p(\bar{D}|bb)} - 1. \tag{3}$$

In the above definition the marginal penetrances $p(D|g_A)$and $p(D|g_B)$ are given by

$$p(D|g_A) = \sum_{g_B} p(D|g_A, g_B)p(g_B),\ p(D|g_B) = \sum_{g_A} p(D|g_A, g_B)p(g_A).$$

With the specified forms of the odds ratios in Models 1 — 3, the parameters p ,λ_A and λ_B are functions of α and θ ((θ_A, θ_B) under Model 1). Therefore, these parameters can be obtained by solving equations (1) — (3) with given p, λ_A and λ_B. Once the values of α and θ are determined, the penetrance of a particular genotype at loci A and B can be derived through the odds ratios as

$$p(D|g_A, g_B) = \frac{p(D|g_A, g_B)/p(\bar{D}|g_A, g_B)}{1 + p(D|g_A, g_B)/p(\bar{D}|g_A, g_B)}$$

The disease status of an individual is then generated by the Bernoulli distribution with probability of success $p(D|g_A, g_B)$.

In the simulation studies, the disease allele frequencies of the two disease loci are taken the same, so are the marginal effects of the two disease loci. The prevalence is set as 0.01 in all settings. The values of α and θ corresponding to different specification of disease allele frequency and marginal effect for Models 1 — 3 are given as follows. Note, by taking the same disease allele frequencies and the same marginal effects at locus A and B, $\theta_A = \theta_B$ in Model 1.

	λ	q	α	θ
Model 1:	0.8	0.1	0.0073	0.80
	0.9	0.1	0.0072	0.90
	1.0	0.1	0.0068	1.00
Model 2:	0.5	0.1	0.0089	2.30
	0.5	0.1	0.0081	1.13
	0.7	0.2	0.0084	3.15
	0.7	0.2	0.0073	1.54
Model 3:	0.8	0.1	0.0088	4.34
	0.9	0.1	0.0087	4.90
	1.0	0.1	0.0085	5.50

For the simulation under Model 4, the β coefficients corresponding to the specified ξ_{12} values are given below:

ξ_{12}	β_0	$\beta_1 = \beta_2$
1.9	-5	-0.38
2.0	-5	-0.40
2.1	-5	-0.42

In the simulation study, the disease loci are not included in the collection of SNPs. Instead, for each disease locus, a SNP in linkage disequilibrium (LD) with the locus is generated. Let A, a denote the alleles of the disease locus and X and x the alleles of the SNP in LD with the disease locus. Given the alleles of the disease locus, the alleles of the marker SNP are generated by the conditional probabilities $P(X|A)$ and $P(X|a)$ with constraint $p(X|A) = 1 - p(X|a)$. The conditional probabilities are determined by setting $r^2_{AX} = 0.5$ where r^2_{AX} is a parameter to measure the magnitude of LD (see Pritchard and Przeworski, 2001). Let q_A and q_X be the allele frequencies of A and X respectively. r^2_{AX} is defined by

$$r^2_{AX} = [p(X|A) - p(X|a)]^2 \frac{q_A(1-q_A)}{q_X(1-q_X)}.$$

4. **More details of the results in simulation study 2.**

In simulation study 2, for the TPLR approach, we considered the cases $n_M = n_I = 15, 30$ and 50 in order to investigate the impact of the choice of n_M and n_I on the final selection results. Only a partial result with $n_M = n_I = 50$ is reported in the paper. The full results are given here. The PDR and FDR are plotted against γ in the following figure. The details are reported in Tables 1 - 3 that follow.

Figure 1. The PDR and FDR of TPLR approach with different $N(= n_M = n_I)$ (Square — N=15, Circle — N=30, Triangle — N=50).

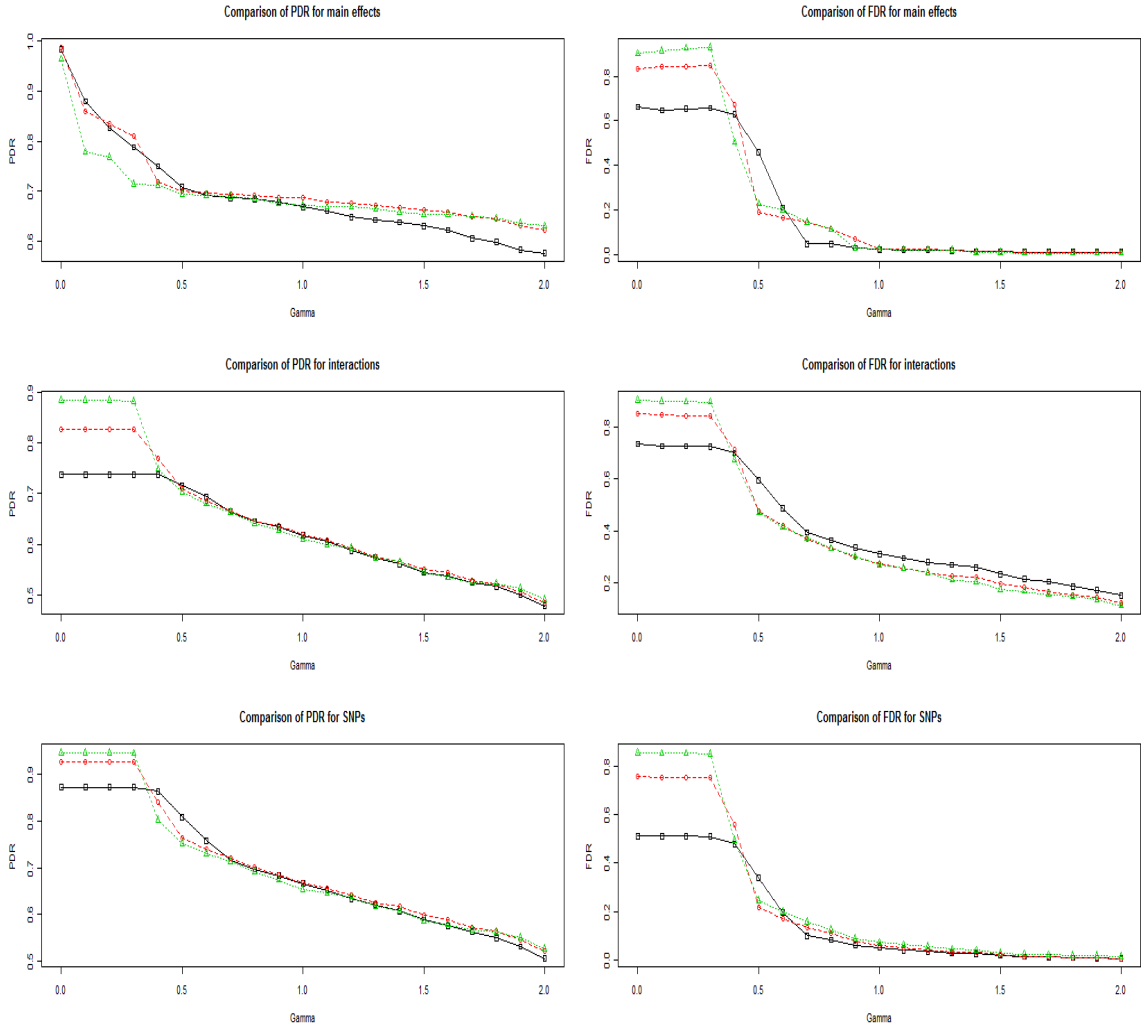

Table 1: The results of TPLR with $n_M = n_I = 15$

γ	PDR_M	FDR_M	PDR_I	FDR_I	PDR_S	FDR_S
0.0	0.984	0.663	0.738	0.737	0.873	0.512
0.1	0.88	0.647	0.738	0.728	0.873	0.512
0.2	0.826	0.654	0.738	0.728	0.873	0.512
0.3	0.788	0.657	0.738	0.726	0.873	0.51
0.4	0.75	0.628	0.738	0.702	0.864	0.479
0.5	0.708	0.457	0.716	0.596	0.809	0.34
0.6	0.692	0.21	0.694	0.486	0.759	0.197
0.7	0.686	0.047	0.664	0.394	0.715	0.101
0.8	0.684	0.047	0.646	0.363	0.697	0.084
0.9	0.678	0.029	0.634	0.334	0.683	0.063
1	0.668	0.023	0.618	0.31	0.666	0.052
1.1	0.66	0.021	0.606	0.294	0.652	0.042
1.2	0.648	0.021	0.588	0.278	0.634	0.036
1.3	0.642	0.018	0.574	0.268	0.621	0.03
1.4	0.638	0.012	0.562	0.257	0.607	0.027
1.5	0.63	0.013	0.546	0.231	0.588	0.02
1.6	0.622	0.01	0.538	0.213	0.576	0.015
1.7	0.606	0.01	0.526	0.203	0.563	0.014
1.8	0.598	0.01	0.518	0.183	0.55	0.01
1.9	0.582	0.01	0.502	0.169	0.531	0.009
2	0.576	0.01	0.48	0.149	0.506	0.008

Table 2: The results of TPLR with $n_M = n_I = 30$

γ	PDR_M	FDR_M	PDR_I	FDR_I	PDR_S	FDR_S
0	0.988	0.832	0.826	0.853	0.928	0.755
0.1	0.86	0.841	0.826	0.848	0.928	0.753
0.2	0.836	0.843	0.826	0.847	0.928	0.753
0.3	0.81	0.847	0.826	0.846	0.928	0.752
0.4	0.718	0.673	0.77	0.716	0.841	0.56
0.5	0.7	0.192	0.71	0.476	0.764	0.218
0.6	0.696	0.163	0.686	0.418	0.741	0.172
0.7	0.694	0.145	0.666	0.368	0.72	0.137
0.8	0.692	0.115	0.646	0.331	0.701	0.11
0.9	0.688	0.07	0.638	0.296	0.686	0.081
1	0.686	0.026	0.62	0.272	0.669	0.059
1.1	0.678	0.023	0.608	0.255	0.656	0.05
1.2	0.676	0.023	0.594	0.238	0.642	0.044
1.3	0.672	0.018	0.576	0.224	0.626	0.035
1.4	0.666	0.012	0.568	0.22	0.617	0.032
1.5	0.662	0.012	0.552	0.193	0.598	0.023
1.6	0.658	0.009	0.546	0.18	0.59	0.017
1.7	0.648	0.009	0.53	0.164	0.573	0.014
1.8	0.644	0.009	0.524	0.152	0.564	0.011
1.9	0.632	0.009	0.508	0.142	0.546	0.01
2	0.622	0.01	0.486	0.12	0.522	0.008

Table 3: The results of TPLR with $n_M = n_I = 50$

γ	PDR_M	FDR_M	PDR_I	FDR_I	PDR_S	FDR_S
0	0.964	0.902	0.884	0.907	0.947	0.855
0.1	0.778	0.915	0.884	0.902	0.947	0.853
0.2	0.768	0.926	0.884	0.9	0.947	0.852
0.3	0.714	0.93	0.882	0.899	0.946	0.85
0.4	0.712	0.505	0.748	0.677	0.803	0.494
0.5	0.694	0.227	0.704	0.47	0.752	0.246
0.6	0.692	0.199	0.68	0.413	0.73	0.2
0.7	0.69	0.144	0.664	0.371	0.714	0.158
0.8	0.684	0.112	0.642	0.331	0.691	0.126
0.9	0.676	0.029	0.628	0.301	0.674	0.089
1	0.672	0.023	0.61	0.267	0.654	0.073
1.1	0.668	0.021	0.6	0.254	0.646	0.064
1.2	0.668	0.021	0.594	0.237	0.638	0.057
1.3	0.664	0.018	0.574	0.209	0.618	0.045
1.4	0.658	0.009	0.566	0.201	0.608	0.041
1.5	0.654	0.009	0.546	0.17	0.586	0.03
1.6	0.654	0.006	0.536	0.165	0.578	0.026
1.7	0.65	0.006	0.526	0.152	0.567	0.023
1.8	0.646	0.006	0.524	0.144	0.563	0.021
1.9	0.636	0.006	0.514	0.132	0.55	0.019
2	0.63	0.006	0.492	0.109	0.526	0.015