– Supplementary Material – High-Dimensional Structured Feature Screening Using Binary Markov Random Fields

Jie Liu
Department of Computer Sciences
Univ. of Wisconsin-Madison

Chunming Zhang
Department of Statistics
Univ. of Wisconsin-Madison

Catherine McCarty
Essentia Institute of
Rural Health

Peggy Peissig
Biomedical Informatics Research Center
Marshfield Clinic Research Foundation

Elizabeth Burnside
Department of Radiology
Univ. of Wisconsin-Madison

David Page
Biostat. & Medical Informatics Dept.
Univ. of Wisconsin-Madison

1 Connection with Main Text

In Section 2.2 of the main text, we introduce how to calculate p_i from hypothesis testing. We feel this may not interest the common AISTATS community, and therefore we did not give many details. Here we give more details by using two hypothesis testing examples. One example is the two-proportion z-test for binary features (e.g. the simulations in Section 3 of the main text), and the other is logistic regression with likelihood ratio test for GWAS data (e.g. the application in Section 4 of the main text).

In Section 3 of the main text, we compare our algorithm with elastic net. For elastic net, we set the α parameter (the tradeoff parameter between l_1 penalty and l_2 penalty) to be 0.5. Readers may wonder whether the α parameter will make a difference. Here, we also show the results for the other choices of α parameter in the elastic net penalty.

2 Two-proportion z-test

Suppose that we are trying to identify whether a binary feature$_i$ is relevant to the binary response variable $Y \in \{0, 1\}$ with the empirical counts from data shown in Table 1.

Table 1: Empirical counts at feature$_i$ with a binary response variable Y.

	feature$_i = 0$	feature$_i = 1$	Total
$Y = 1$	u_0	u_1	u
$Y = 0$	v_0	v_1	v
Total	n_0	n_1	n

$\mathcal{F}_i^+$ denotes the random variable of the feature$_i$ in the positive samples. $\mathcal{F}_i^-$ denotes the random variable of the feature$_i$ in the negative samples.

$$\mathcal{F}_i^+ \sim Bernoulli(\mathcal{P}_i^+),\ \mathcal{F}_i^- \sim Bernoulli(\mathcal{P}_i^-). \quad (1)$$

$\mathcal{P}_i^+$ and $\mathcal{P}_i^-$ are the population probability that feature$_i$ is 1 in the positive and negative population, respectively. Accordingly, $\hat{\mathcal{P}}_i^+$ and $\hat{\mathcal{P}}_i^-$ are sample-based version of $\mathcal{P}_i^+$ and $\mathcal{P}_i^-$. We can calculate $\hat{\mathcal{P}}_i^+$ and $\hat{\mathcal{P}}_i^-$ from Table 1 as

$$\hat{\mathcal{P}}_i^+ = \frac{u_1}{u},\ \hat{\mathcal{P}}_i^- = \frac{v_1}{v}. \quad (2)$$

The test statistic for feature$_i$ is

$$S_i = \frac{\hat{\mathcal{P}}_i^+ - \hat{\mathcal{P}}_i^-}{\sqrt{Var(\hat{\mathcal{P}}_i^+ - \hat{\mathcal{P}}_i^-)}}. \quad (3)$$

S_i is approximately normally distributed with variance 1 and mean δ_i , where

$$\delta_i = \frac{\mathcal{P}_i^+ - \mathcal{P}_i^-}{\sqrt{\frac{\mathcal{P}_i^+(1-\mathcal{P}_i^+)}{u} + \frac{\mathcal{P}_i^-(1-\mathcal{P}_i^-)}{v}}}. \quad (4)$$

δ_i is termed the *non-centrality parameter*. Under the null hypothesis $\mathcal{H}_0$ of no association, S_i is approximately standard normally distributed. Under alternative hypothesis $\mathcal{H}_1$, S_i is approximately normally distributed with variance 1 and some nonzero mean δ_i. For any given significance level, the power of the test is entirely determined by the absolute value of the

non-centrality parameter. For a given sample set, the larger $|\delta_i|$ we have, the larger the power of the test is.

With hypothesis testing, we usually set p_i to be 1 if the absolute value of the test statistic is greater than or equal to some threshold ξ (for example, the critical value at a certain level) and 0 if otherwise. We term the p_i (from such a "hard" method using some threshold) p_i^H,

$$p_i^H = \begin{cases} 1, & if\ |S_i| \geq \xi, \\ 0, & otherwise. \end{cases}$$

If we know δ_i, we will also know the probability density function of S_i under $\mathcal{H}_1$ (denoted as $f_{S_i|\mathcal{H}_1}$) as well as the probability density function of S_i under $\mathcal{H}_0$ (denoted as $f_{S_i|\mathcal{H}_0}$). By Bayes' rule, we can set

$$p_i^B = \frac{1}{\alpha f_{S_i|\mathcal{H}_0}(s_i) + 1}, \tag{5}$$

and

$$\alpha = \frac{P(\mathcal{H}_0)}{f_{S_i|\mathcal{H}_1}(s_i) P(\mathcal{H}_1)}. \tag{6}$$

However, in most of the cases δ_i is unknown to us, but we can use its data-driven version δ_i^* by replacing $\mathcal{P}_i^+$ and $\mathcal{P}_i^-$ in (4) with the sample probabilities $\hat{\mathcal{P}}_i^+$ and $\hat{\mathcal{P}}_i^-$ (as (2)). This step has a flattening effect on calculating p_i because it assumes the values of the test statistic for relevant features are uniformly distributed. Therefore, we introduce an adaptive procedure for calculating p_i by

$$p_i = \gamma p_i^H + (1-\gamma) p_i^B, \tag{7}$$

where $0 \leq \gamma \leq 1$. We choose ξ in p_i^H to be the test statistic that makes p_i^B be 0.5 in (5). In addition, we also need to specify $P(\mathcal{H}_0)/P(\mathcal{H}_1)$ which can be given from prior knowledge.

3 Logistic Regression with Likelihood Ratio Test

Many GWAS applications employ logistic regression followed by a hypothesis test to identify associated SNPs. A first step builds a logistic regression model (in (8)) to predict disease from each SNP individually; in such a model the SNP is coded by two indicator variables, one for heterozygous carrier of the minor allele (X_1) and one for homozygous carrier of the minor allele (X_2). In other words, we convert AA into "X_1=0, X_2=0", AB into "X_1=1, X_2=0", and BB into "X_1=0, X_2=1" where A stands for the common allele at this locus and B stands for the minor allele. The dichotomous response variable Y is coded as 1 for cases and 0 for controls.

$$\log \frac{P(Y=1|X_1, X_2)}{1 - P(Y=1|X_1, X_2)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{8}$$

In the second step, a hypothesis test is performed to test the fit of each logistic model and to return a P-value for each SNP. In the test, the null hypothesis $\mathcal{H}_0$ is that the SNP is not associated, namely β_1 and β_2 are zeros. The alternative hypothesis $\mathcal{H}_1$ is that the feature is associated, namely either β_1 or β_2 are nonzero. Finally, SNPs are ranked by the P-values. The likelihood ratio test is the most commonly used method, and the test statistic is

$$S = 2(\log L_1 - \log L_0), \tag{9}$$

where $\log L_1$ and $\log L_0$ are the log-likelihood under $\mathcal{H}_1$ and $\mathcal{H}_0$ respectively. Under $\mathcal{H}_0$, the test statistic has an asymptotic χ^2 distribution with 2 degrees of freedom. Under $\mathcal{H}_1$, the test statistic has an asymptotic non-central χ^2 distribution with 2 degrees of freedom. The rest of the calculation of p_i is the same as in the binary feature case, namely using formulas (5), (6) and (7).

4 More Simulations

In Section 3 of the main text, we compare our algorithm with elastic net. For elastic net, we set the α parameter (the tradeoff parameter between l_1 penalty and l_2 penalty) to be 0.5. Readers may wonder whether different α parameter will make a difference. Here, we show the results for the other choices of α parameter in the elastic net penalty.

For the first set of experiments, we set $n = 500$, $h = 1000$, $m = 5$, t_i uniformly distributed on the interval $(0.8, 1.0)$, $\pi = \{0.025, 0.05\}$, and $rr = \{1.1, 1.2, 1.3\}$. For elastic net, we try 4 values for α, namely 0.2, 0.4, 0.6, and 0.8. The ROC curves are shown in Figure 1. The precision-recall curves are shown in Figure 2.

For the second set of experiments, we set $n = 500$, $h = 1000$, $\pi = 0.05$, rr uniformly distributed on the interval $(1.1, 1.3)$, $m = \{2, 5, 10\}$, and t_i uniformly distributed on the interval $(\tau, 1.0)$ where $\tau = \{0.5, 0.8, 0.9\}$. For elastic net, we try 4 values for α, namely 0.2, 0.4, 0.6, and 0.8. The ROC curves are shown in Figure 3. The precision-recall curves are shown in Figure 4.

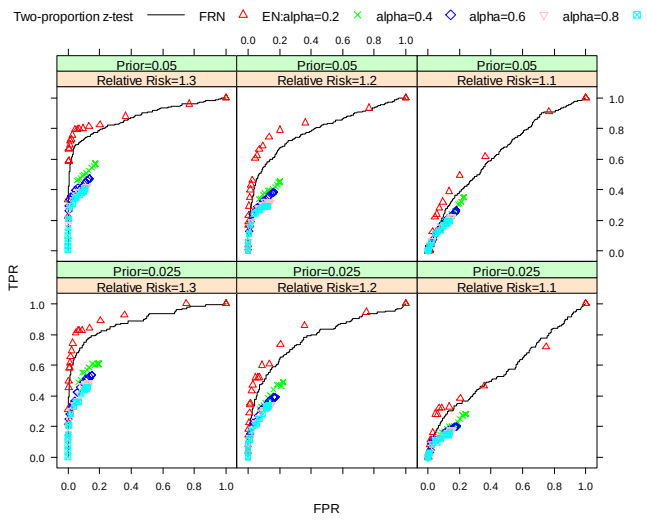

Figure 1: ROC curves of two-proportion z-test (Two-prop z-test), feature relevance network (FRN) and elastic net ($\alpha = 0.2$, $\alpha = 0.4$, $\alpha = 0.6$, and $\alpha = 0.8$) when we choose different prior probabilities and different relative risk levels.

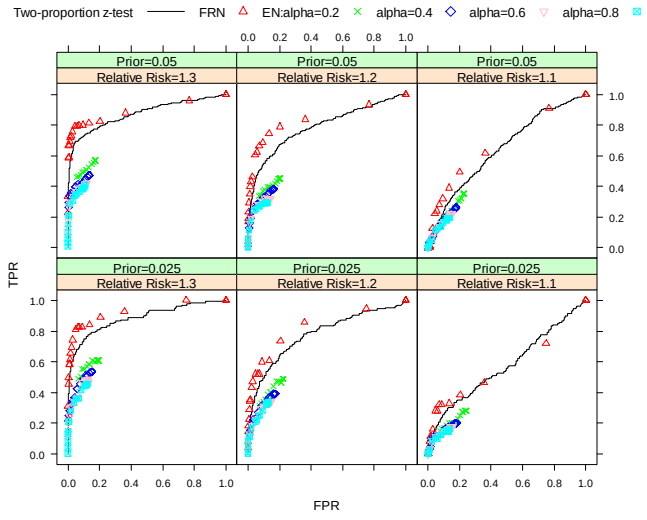

Figure 2: Precision-recall curves of two-proportion z-test (Two-prop z-test), feature relevance network (FRN) and elastic net ($\alpha = 0.2$, $\alpha = 0.4$, $\alpha = 0.6$, and $\alpha = 0.8$) when we choose different prior probabilities and different relative risk levels.

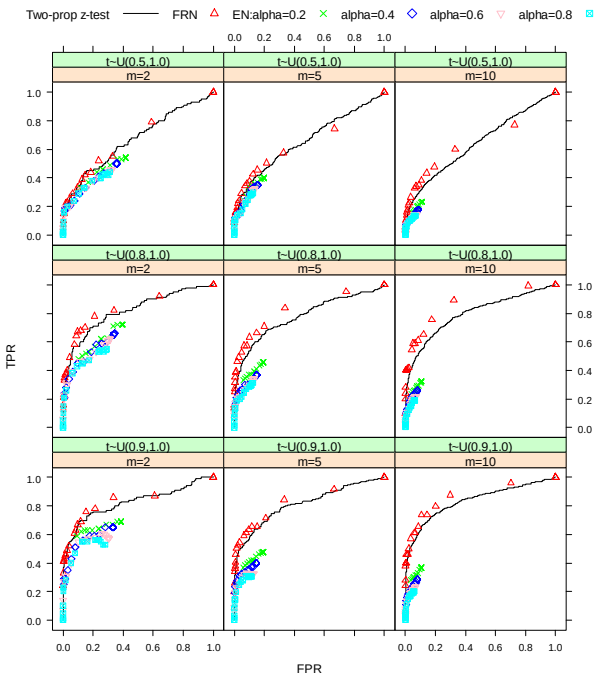

Figure 3: ROC curves of two-proportion z-test, feature relevance network (FRN) and elastic net ($\alpha = 0.2$, $\alpha = 0.4$, $\alpha = 0.6$, and $\alpha = 0.8$) when we choose different correlation structures of covariates.

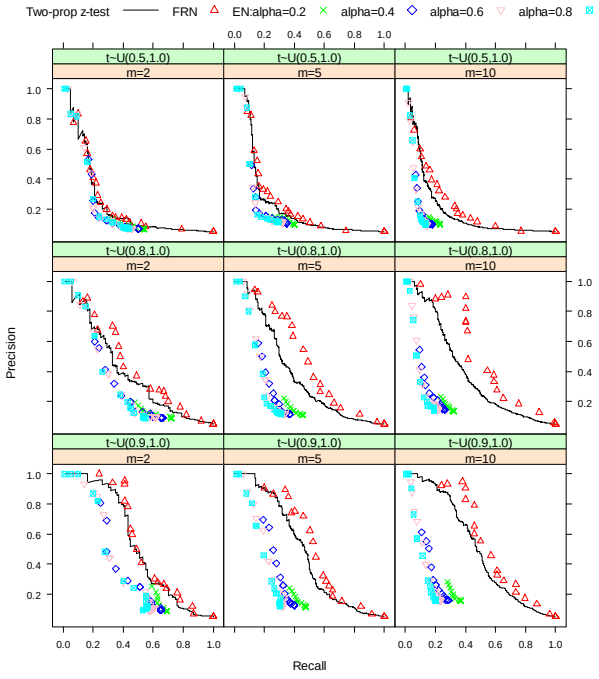

Figure 4: Precision-recall curves of two-proportion z-test, feature relevance network (FRN) and elastic net ($\alpha = 0.2$, $\alpha = 0.4$, $\alpha = 0.6$, and $\alpha = 0.8$) when we choose different correlation structures of covariates.