Constrained maximum entropy models to select genotype interactions associated with censored failure times

Aotian Yang*,‡, David Miller†§ and Qing Pan*,¶

*Department of Statistics, George Washington University
Washington, DC 20052, USA

†Department of Electrical Engineering, Pennsylvania State University
State College, PA 16801, USA

‡atyang@gwu.edu
§djmiller@engr.psu.edu
¶qpan@gwu.edu

Received 24 September 2018
Accepted 29 October 2018
Published 19 December 2018

We propose a novel screening method targeting genotype interactions associated with disease risks. The proposed method extends the maximum entropy conditional probability model to address disease occurrences over time. Continuous occurrence times are grouped into intervals. The model estimates the conditional distribution over the disease occurrence intervals given individual genotypes by maximizing the corresponding entropy subject to constraints linking genotype interactions to time intervals. The EM algorithm is employed to handle observations with uncertainty, for which the disease occurrence is censored. Stepwise greedy search is proposed to screen a large number of candidate constraints. The minimum description length is employed to select the optimal set of constraints. Extensive simulations show that five or so quantile-dependent intervals are sufficient to categorize disease outcomes into different risk groups. Performance depends on sample size, number of genotypes, and minor allele frequencies. The proposed method outperforms the likelihood ratio test, Lasso, and a previous maximum entropy method with only binary (disease occurrence, non-occurrence) outcomes. Finally, a GWAS study for type 1 diabetes patients is used to illustrate our method. Novel one-genotype and two-genotype interactions associated with neuropathy are identified.

Keywords: Maximum entropy; lagrange multiplier; censoring; EM algorithm; stepwise greedy search; GWAS.

1. Instruction

There are approximately eleven million Single Nucleotide Polymorphism (SNPs) in the human genome. They are widely studied for associations with diseases. While a specific genotype at a single locus may increase or decrease disease risk,
the majority of SNPs have very minor impacts on biological systems. Most of the time the manifestation of a disease is affected by genetic variants at multiple loci in addition to environmental factors. There have been many examples of gene-gene interactions playing important roles in the genetic basis of complex diseases. The joint effect from two or more genes to affect a phenotype such as disease susceptibility is called epistasis or interaction. Statistically, interactions refer to synergy between multiple genotypes such that their joint effect on the phenotype does not equal the sum of the marginal effects from individual genotypes. Specifically, we are interested in cases where the impacts on disease risk require the presence of all genotypes involved in the interaction. Traditional variable selection methods are not designed to screen for the combination of multiple genotypes. Therefore, statistical computational methods targeting interactions, that is, the combinations of multiple genotypes, are highly desirable.

Both parametric and nonparametric methods have been developed to screen for interactions. Parametric methods use regressions to describe the relationship between the disease outcome and genotype interactions. In contrast, nonparametric methods propose statistics measuring the association between candidate interactions and the phenotypes of interest with few assumptions imposed on the data. Wu et al. modified lasso to fit logistic models in Genome-Wide Association Studies (GWAS), allowing for an automatic selection of prognostic SNPs. Silver et al. extended the idea to group lasso for groups of SNPs. Ayers et al. showed that the penalized methods outperform single marker testing procedures and forward stepwise variable selection. In the presence of a large number of candidate SNPs, multiway interactions raise search complexity and the associated computational load substantially; thus, more sophisticated and efficient searching algorithms are needed. Screening algorithms for genotype interactions can be classified into four categories: exhaustive search, stepwise search, stochastic search and heuristic search.

Miller et al. proposed Maximum Entropy Conditional Probability Models (MECPM), a nonparametric method with a stepwise search algorithm for detecting genotype interactions. It estimates the conditional distribution of disease outcome (disease or not) given individual-level genotype information by maximizing the entropy of the conditional distributions subject to chosen constraints that link genotype interactions to the binary disease status. Multiple constraints are allowed in the model. These equality constraints force the model to agree, on the given population, with joint probabilities of events involving disease status and a genotypic interaction. Models are selected using a minimum description length (MDL) framework. However, the MECPM algorithm of Miller et al. is limited to analyses of binary outcomes. Clinical trials and epidemiology studies often follow subjects over time, which results in disease occurrences over time. Furthermore, the exact occurrence time may be censored, that is, the disease occurrence may only be known to fall into a time range containing several intervals. We extend MECPM to the analysis of disease occurrence times with left, right, interval, and mixed censoring. The proposed
method handles ordered multi-class outcomes specifying the time interval containing disease occurrence.

2. Methods

First, we set up the notation. There are \( n \) subjects \( (i = 1, \ldots, n) \) and \( P \) SNPs \( (p = 1, \ldots, P) \). A genotype is denoted by \( G_{ip} \) for the \( i \)th subject and \( p \)th SNP, with \( G_{ip} \in \{0,1,2\} \) where 0, 1 and 2 are the number of minor alleles, i.e. \{AA, Aa, aa\} respectively. Furthermore, we use \( G_i \) to denote the complete set of genotypes for all \( P \) SNPs of subject \( i \). Although the underlying event times are continuous, the full time axis \((0, +\infty)\) is divided into \( K \) non-overlapping intervals. The disease occurrence for subject \( i \) falling in interval \( k \) is denoted by \( T_i = k, k \in \{1, \ldots, K\} \). Sometimes, \( T_i \) is censored and \( C_{iL} \leq T_i \leq C_{iR} \) is observed where \( C_{iL}, C_{iR} \in \{1, \ldots, K\} \). That is, \( T_i \) may be any interval from \( C_{iL} \) to \( C_{iR} \).

2.1. Conditional probability model

Use \( T \) and \( G \) to denote the random variables of the disease occurrence interval and the random vector of genotypes of one or more SNPs, respectively. Also, \( R \) and \( g \) are some chosen candidate time interval and genotype interaction, respectively, constituting a “constraint”. Specifically, we consider \( R \) in the form of left-open \(((1), (1,2), (1,2,3), \ldots, (1, \ldots, K-1))\) or right-open \(((2, \ldots, K), \ldots, (K-1,K), (K))\) intervals. Let \( N_c \) be the total number of constraints, and \( m = 1, 2, \ldots, N_c \) be the index of the constraints in a model. For the \( m \)th constraint with genotype interaction \( g_m \) and time interval \( R_m \), the joint probability satisfying both is \( P(T \in R_m, G_{(m)} = g_m) \). We use the 0–1 indicator function \( I(T_i \in R_m, G_{i(m)} = g_m) \) to indicate whether or not subject \( i \) contains all genotypes in \( g_m \) and its event time belongs to \( R_m \). The probability satisfying the \( m \)th constraint can be consistently estimated by the corresponding percentage of subjects in the sample:

\[
\hat{P}_g(T \in R_m, G_{(m)} = g_m) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \in R_m, G_{i(m)} = g_m).
\]

The estimate of this joint probability satisfying both \( R_m \) and \( g_m \) from the conditional probability model is

\[
\hat{P}_{ME}(T \in R_m, G_{(m)} = g_m) = \frac{1}{n} \sum_{i=1}^{n} P(T_i \in R_m | G_{i(m)} = g_m) I(G_{i(m)} = g_m).
\]

That is, it is the normalized sum of the conditional probabilities of falling into the time interval \( R_m \) among subjects containing the genotype \( g_m \). To satisfy the constraint, we require

\[
\hat{P}_{ME}(T \in R_m, G_{(m)} = g_m) = \hat{P}_g(T \in R_m, G_{(m)}).
\]
However, the conditional distribution of the $i$th subject’s event time interval, $T_i$, given its genomic information $G_i$, is unidentifiable solely under criterion (1). Shannon’s entropy, $H$, for the conditional distribution is defined as

$$H \equiv -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} P(k | G_i) \log P(k | G_i),$$

where $P(k | G_i)$ denotes the probability of disease onset in interval $k$, i.e. $T_i = k$, given the complete set of genotypes of subject $i$, $G_i$. The conditional distribution $P(k | G_i)$ is estimated by maximizing the entropy $H$ while satisfying $\hat{P}_{ME}(T \in R_m, G_{i(m)} = g_m) = \hat{P}_{g}(T \in R_m, G_{i(m)} = g_m)$, $m = 1, \ldots, N_c$.

**Proposition 1.** The solution that maximizes $H$ subject to $\hat{P}_{ME}(T \in R_m, G_{i(m)} = g_m) = \hat{P}_{g}(T \in R_m, G_{i(m)} = g_m)$, $m = 1, \ldots, N_c$, exists and is unique.

The proof is outlined in Appendix A.

The unconstrained Lagrangian dual objective which satisfies the constraints and maximizes $H$ is

$$L = H + \sum_{m=1}^{N_c} \gamma_m \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ P(T_i \in R_m | G_{i(m)} = g_m) \right] I(G_{i(m)} = g_m) - I(T_i \in R_m, G_{i(m)} = g_m) \right\} + \frac{1}{n} \sum_{i=1}^{n} \lambda_i \left\{ \sum_{k=1}^{K} P(k | G_i) - 1 \right\}.$$  

Here $\gamma_m$ denotes the Lagrange multiplier for the $m$th constraint, and $\lambda_i$ is the Lagrange multiplier corresponding to the constraint that the conditional probabilities for each subject sum up to one. For a given $i$, setting $\frac{\partial L}{\partial P(k | G_i)} |_{P(k | G_i) = P_{ME}(k | G_i)} = 0$, we get

$$\hat{P}_{ME}(k | G_i) = \exp \left( \sum_{m=1}^{N_c} \gamma_m I(k \in R_m, G_{i(m)} = g_m) \right) \exp (\lambda_i - 1).$$

Furthermore, $\sum_{k=1}^{K} \hat{P}_{ME}(k | G_i) = 1$. Then $\exp (\lambda_i - 1) = \frac{1}{\sum_{k=1}^{K} \{ \exp (\sum_{m=1}^{N_c} \gamma_m I(k \in R_m, G_{i(m)} = g_m)) \}}$. Therefore, the conditional probability estimator maximizing $L$ is

$$\hat{P}_{ME}(k | G_i) = \frac{\exp \left( \sum_{m=1}^{N_c} \gamma_m I(k \in R_m, G_{i(m)} = g_m) \right)}{\sum_{k' = 1}^{K} \{ \exp (\sum_{m=1}^{N_c} \gamma_m I(k' \in R_m, G_{i(m')} = g_{m'})) \}}.$$  

Plugging $\hat{P}_{ME}(k | G_i)$ into the Lagrangian Dual objective $L$, we get a simpler form.

**Proposition 2.** When $P(k | G_i) = \hat{P}_{ME}(k | G_i)$,

$$\hat{L} = L |_{P(k | G_i) = \hat{P}_{ME}(k | G_i)} = -\log \prod_{i=1}^{n} \hat{P}_{ME}(t_i | G_i).$$

Here $\hat{P}_{ME}(t_i | G_i)$ is the maximum entropy estimator for the conditional probability of $T_i = t_i$ given $G_i$, where $t_i$ is the observed value of $T_i$. The values of $\gamma_m$, $m = 1, \ldots, N_c$, that maximize the likelihood $\prod_{i=1}^{n} \hat{P}_{ME}(t_i | G_i)$ also minimize the constrained entropy $L$. The proof of Proposition 2 is outlined in Appendix B.
2.2. **EM algorithm**

In Proposition 2, we establish the equivalency between the constrained entropy function and the negative log-likelihood of the conditional distribution of $t_i$ given $G_i$. However, $t_i$ is not always observed; instead we may observe $C_{iL} \leq T_i \leq C_{iR}$. It is assumed that $C_{iL}$ and $C_{iR}$ are independent of $T_i$ conditional on $G_i$. Then we treat $t_i$ as missing data and employ the EM algorithm. Taking the conditional expectation of the complete data joint log-likelihood of observed $t_i$ and missing $t_i$, we get the following $Q$ function.

### 2.2.1. E-step

Let $\Delta_i$ be the indicator that $t_i$ is observed. Define

$$Q(\gamma) = \sum_{i=1}^{n} \sum_{m=1}^{N} \gamma_m I(t_i = R_m, G_{i(m)} = g_m) \Delta_i$$

$$+ \sum_{i=1}^{n} \mathbb{E}_{T_i | C_{iL} \leq T_i \leq C_{iR}, G_i} \left\{ \sum_{m=1}^{N} \gamma_m I(T_i = R_m, G_{i(m)} = g_m) \right\} (1 - \Delta_i)$$

$$- \sum_{i=1}^{n} \log \left( \sum_{k' = 1}^{K} \exp \left\{ \sum_{m=1}^{N} \gamma_m I(k' = R_m, G_{i(m)} = g_m) \right\} \right).$$

When $\gamma$ is updated by $\gamma + \Delta \gamma$,

$$Q(\gamma + \Delta \gamma) = \sum_{i=1}^{n} \sum_{m=1}^{N} (\gamma_m + \Delta \gamma_m) I(t_i = R_m, G_{i(m)} = g_m) \Delta_i$$

$$+ \sum_{i=1}^{n} \mathbb{E}_{T_i | C_{iL} \leq T_i \leq C_{iR}, G_i} \left\{ \sum_{m=1}^{N} (\gamma_m + \Delta \gamma_m) I(T_i = R_m, G_{i(m)} = g_m) \right\} (1 - \Delta_i)$$

$$- \sum_{i=1}^{n} \log \left( \sum_{k' = 1}^{K} \exp \left\{ \sum_{m=1}^{N} (\gamma_m + \Delta \gamma_m) I(k' = R_m, G_{i(m)} = g_m) \right\} \right).$$

Therefore,

$$Q(\gamma + \Delta \gamma) - Q(\gamma) = \sum_{i=1}^{n} \sum_{m=1}^{N} \Delta \gamma_m I(t_i = R_m, G_{i(m)} = g_m) \Delta_i$$

$$+ \sum_{i=1}^{n} \mathbb{E}_{T_i | C_{iL} \leq T_i \leq C_{iR}, G_i} \left\{ \sum_{m=1}^{N} \Delta \gamma_m I(T_i = R_m, G_{i(m)} = g_m) \right\} (1 - \Delta_i)$$

$$- \sum_{i=1}^{n} \log \left( \sum_{k' = 1}^{K} \exp \left\{ \sum_{m=1}^{N} (\gamma_m + \Delta \gamma_m) I(k' = R_m, G_{i(m)} = g_m) \right\} \right).$$
Proposition 3. It is proved in Appendix C that

\[
Q(\gamma + \Delta \gamma) - Q(\gamma) \geq \sum_{i=1}^{n} \left( \sum_{m=1}^{N_r} \Delta \gamma_m I(t_i \in R_m, G_{i(m)} = g_m) + 1 \right) \Delta_i \\
+ \sum_{i=1}^{n} \left( \mathbb{E}_{T_i|C_i \leq T_i \leq C_i \gamma_i G} \frac{1}{C_0} \right) \times \left\{ \sum_{m=1}^{N_r} \Delta \gamma_m I(T_i = R_m, G_{i(m)} = g_m) + 1 \right\} (1 - \Delta_i) \\
- \frac{1}{N_c} \sum_{i=1}^{n} \sum_{k'=1}^{K} \sum_{m=1}^{N_r} \hat{P}_{ME}(k'|G_{i(m)}) \exp \left\{ N_c \Delta \gamma_m I(k' = R_m, G_{i(m)} = g_m) \right\}.
\]

2.2.2. M-step

Motivated by the lower bound in Proposition 3, an auxiliary function \( A(\Delta \gamma|\gamma) \) is constructed as follows:

\[
A(\Delta \gamma|\gamma) = \sum_{i=1}^{n} \left( \sum_{m=1}^{N_r} \Delta \gamma_m I(t_i = R_m, G_{i(m)} = g_m) + 1 \right) \Delta_i \\
+ \sum_{i=1}^{n} \left( \mathbb{E}_{T_i|C_i \leq T_i \leq C_i \gamma_i G} \frac{1}{C_0} \right) \times \left\{ \sum_{m=1}^{N_r} \Delta \gamma_m I(T_i = R_m, G_{i(m)} = g_m) + 1 \right\} (1 - \Delta_i) \\
- \frac{1}{N_c} \sum_{i=1}^{n} \sum_{k'=1}^{K} \sum_{m=1}^{N_r} \hat{P}_{ME}(k'|G_{i(m)}) \exp \left\{ N_c \Delta \gamma_m I(k' = R_m, G_{i(m)} = g_m) \right\}.
\]

When \( \Delta \gamma = 0 \), \( A(0|\gamma) = n \cdot \Delta_i - n \cdot \Delta_i + n \cdot (1 - \Delta_i) - n \cdot (1 - \Delta_i) = 0 \). When \( \Delta \gamma^* = \arg \max_{\Delta \gamma} A(\Delta \gamma|\gamma) \),

\[
Q(\gamma + \Delta \gamma^*|\gamma^{(j)}) - Q(\gamma|\gamma^{(j)}) \geq A(\Delta \gamma^*|\gamma) \geq 0.
\]

The optimal \( \Delta \gamma \) that guarantees \( Q(\gamma^{(j)} + \Delta \gamma) \geq Q(\gamma^{(j)}) \) and maximizes \( A(\Delta \gamma|\gamma) \) after each update is found through the improved iterative scaling (IIS) method.
Constrained maximum entropy models to select genotype interactions

Differentiating $A(\Delta \gamma | \gamma)$ with respect to $\Delta \gamma_m$, 
\[
\frac{\partial A(\Delta \gamma | \gamma)}{\partial \Delta \gamma_m} = \sum_{i=1}^{n} I(t_i \in R_m, G_{i(m)} = g_m) \Delta_i \\
+ \sum_{i=1}^{n} \mathbb{E}_{T_i | C_i \leq T_i \leq C_{iR} \gamma_i G} \{ I(T_i = R_m, G_{i(m)} = g_m) \} (1 - \Delta_i) \\
- \frac{1}{N_c} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{P}_{ME}(k' | G_i) \exp\{ N_c \Delta \gamma_m I(k' = R_m, G_{i(m)} = g_m) \} \\
\times N_c I(k' = R_m, G_{i(m)} = g_m).
\]

Let $\frac{\partial A(\Delta \gamma | \gamma)}{\partial \Delta \gamma_m} = 0$. First, consider the case without censoring, i.e. $\Delta_i = 1$. We have 
\[
\sum_{i=1}^{n} I(t_i \in R_m, G_{i(m)} = g_m) \\
= \frac{1}{N_c} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{P}_{ME}(k' | G_i) \exp\{ N_c \Delta \gamma_m I(k' = R_m, G_{i(m)} = g_m) \} \\
\times N_c I(k' = R_m, G_{i(m)} = g_m) \\
= \sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) \exp\{ N_c \Delta \gamma_m I(G_{i(m)} = g_m) \}.
\]

Therefore, 
\[
\Delta \gamma^*_m = \frac{1}{N_c} \log \left\{ \frac{\sum_{i=1}^{n} I(t_i = R_m, G_{i(m)} = g_m)}{\sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) I(G_{i(m)} = g_m)} \right\}.
\]

Next, consider the case of censoring where $\Delta_i = 0$, $i = 1, \ldots, n$, 
\[
\sum_{i=1}^{n} \mathbb{E}_{T_i | C_i \leq T_i \leq C_{iR} \gamma_i G} \{ I(t_i = R_m, G_{i(m)} = g_m) \} \\
= \sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) \exp\{ N_c \Delta \gamma_m I(G_{i(m)} = g_m) \}.
\]

\[
\Delta \gamma^*_m = \frac{1}{N_c} \log \left\{ \frac{\sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) I(G_{i(m)} = g_m)}{\sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) I(G_{i(m)} = g_m)} \right\} \\
= \frac{1}{N_c} \log \left\{ \frac{\sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) I(G_{i(m)} = g_m)}{\sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) I(G_{i(m)} = g_m)} \right\} \\
= \frac{1}{N_c} \log \left\{ \frac{\sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) I(G_{i(m)} = g_m)}{\sum_{i=1}^{n} \hat{P}_{ME}(R_m | G_i) I(G_{i(m)} = g_m)} \right\}.
\]

1840024-7
Combining both scenarios where $\Delta_i = 1$ or $\Delta_i = 0$, the Lagrange multipliers $\gamma^{(j)}$ at the $j$th iteration are updated by $\gamma^{(j+1)} = \gamma^{(j)} + \Delta \gamma^*$, where

$$\Delta \gamma^*_m = \frac{1}{N_c} \log \frac{\hat{P}_g(T \in R_m, G_{(m)} = g_m)}{\hat{P}_{ME}(T \in R_m, G_{(m)} = g_m)},$$

$$\hat{P}_g(T \in R_m, G_{(m)} = g_m) = \frac{1}{n} \sum_{i=1}^{n} I(t_i \in R_m, G_{(i(m))} = g_m) \Delta_i$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{k=C_{il}}^{C_{il}} I(k \in R_m) \exp \left\{ \sum_{m'=1}^{N_c} \gamma^{(j)}_{m'} I(k = R_{m'}, G_{(m')} = g_{m'}) \right\} \right\}$$

$$\times I(G_{(i(m))} = g_m) (1 - \Delta_i),$$

$$\hat{P}_{ME}(R_m, G_{(m)} = g_m) = \frac{1}{n} \sum_{k=1}^{K} I(k \in R_m) \exp \left\{ \sum_{m'=1}^{N_c} \gamma^{(j)}_{m'} I(k = R_{m'}, G_{(m')} = g_{m'}) \right\}.$$  

When $\Delta \gamma^* = 0$, $\hat{P}_{ME}(R_m, G_{(m)} = g_m) = \hat{P}_g(T \in R_m, G_{(m)} = g_m)$ and $A(\Delta \gamma^* | \gamma) = 0$.

### 2.3. Greedy search method

Exhaustive search is unrealistic given the large number of candidate constraints. Maximum entropy methods, which generally require heuristic search or other heuristic techniques for determining which constraints to encode in the model, have been widely used in natural language, speech, as well as general scientific and social science domains. Instead of exhausting all possible genotype combinations, we decompose the search of a genotype combination into several steps; one genotype is added at each step, and the interaction considered in a later step includes genotypes selected in the previous steps.

**Step 1:** Exhaustive search of all the order-1 and order-2 constraints. The number of all order-1 constraints is $P \times 3 \times 2 \times (K - 1)$. Here, $P$ is the number of candidate SNPs. There are 3 types of genotypes from each SNP and $2 \times (K - 1)$ candidate time intervals. The number of order-2 constraints is $C_2^P \times 3 \times 2 \times (K - 1)$. We sort all the constraints with respect to their MDL, choose the 20 models with the smallest MDL and denote them as $F_1, \ldots, F_{20}$. This is the initial candidate constraint set.

**Step 2:** Take an initial model, $F_j$, with an order-1 constraint as an example. Fit all MECPM models with order-2 constraints including the genotype already selected in $F_j$. Again, we choose the 20 models with the smallest MDL among all such order-2 constraints. For each of the 20 order-2 constraints initiated from $F_j$, we continue adding one more genotype into the
constraint on top of the two selected genotypes, keeping the 20 order-3 constraints producing the smallest MDL. Then we pool the $20 \times 20 = 400$ order-3 candidate constraints and keep the top 20 with the smallest MDL among the 400 models.

**Step 3:** For each candidate $F_j$, we search all the way up to order 5 by keeping the top 20 candidate constraints at each order. We save the 5 (or 4 if we start from order 2) models $F_j^{(1)}, F_j^{(2)}, F_j^{(3)}, F_j^{(4)}, F_j^{(5)}$ with minimum MDL at each order, respectively. We compare these 5 models and choose the $F_j^*$ with the smallest MDL among $F_j^{(1)}, \ldots, F_j^{(5)}$.

**Step 4:** Compare the 20 constraints $F_1^*, \ldots, F_{20}^*$ and choose the one with the smallest MDL as our final pick.

**Step 5:** Repeat steps 1 to 4 to add another constraint to the current model, until the MDL stops decreasing when an additional constraint is added. We then report the final model and its associated constraints.

### 2.4. Model selection with MDL

Minimum Description Length (MDL), which is equivalent to Bayesian Information Criterion,

10 is used to select a set of predictive yet simple constraints. Following Miller et al.,

20 MDL is defined as

$$MDL = B_\Theta - \sum_{i=1}^{n} \log_2 \hat{P}_{ME}(t_i|G_j),$$

where $B_\Theta$ is the number of bits used to describe the model. In our model,

$$B_\Theta = \sum_{m=1}^{N_c} \left[ \frac{1}{2} \log_2(n) + \log_2(5) + \log_2(C_{n_m}^P \cdot 3^{n_m} \cdot 2(K-1)) \right],$$

where $5$ is the assumed maximum number of genotypes in a constraint and $n_m$ is the number of genotypes in the $m$th constraint, $C_{n_m}^P$ is the number of ways choosing $n_m$ SNPs from a pool of size $P$, $3^{n_m}$ is the number of genotype combinations from $n_m$ SNPs when we choose one genotype per SNP, and $2(K-1)$ is the number of candidate time intervals.

We illustrate MDL using a simulated case with eight genotype interactions truly associated with disease risk. A large sample size of 50,000 is generated. The MDL values are plotted against the number of selected constraints via the greedy search in Fig. 1. At each iteration of MECPM, the selected genotype interaction is marked on the curve, where the number is the ID of the selected SNP. The selected genotypes are marked in the parentheses, where “A” and “a” represent the major and minor alleles, respectively. MDL decreases when more constraints are selected and reaches its minimum after all eight true constraints are identified. No false positive constraints are selected by the 8th iteration, where MDL reaches the minimum.
3. Simulation Studies

We examine the kind of epistasis where the effects on disease exist if and only if all genotypes in an interaction are present. The genotypes of all candidate SNPs are generated as independent multinomial random variables with probabilities $(p^2, 2p(1-p), (1-p)^2)$, where $p$ denotes the minor allele frequency. Three values of $p$ (0.1, 0.2, 0.3) are tried. Then genotype interactions are treated as binary covariates. That is, if a subject carries all the genotypes in a specific interaction, the covariate indicating the interaction is 1 for the subject; and 0 otherwise. The Cox proportional hazards model is employed to generate the subject-specific hazard of the disease over time and the corresponding disease occurrence time,

$$
\lambda(T_i | G_j) = \lambda_0 \cdot \exp \left( \sum_{m=1}^{N_c} \beta_m I(G_i(m) = g_{m}) \right).
$$

Here $\beta_m$ is the coefficient for constraint $m$, representing the size and direction of the genotype interaction effect, and $I(G_i(m) = g_{m})$ is the indicator of whether the $i$th subject satisfies all the genotypes in the $m$th constraint. The baseline hazard rate is set as $\lambda_0 = 0.02$. Censoring time $C_i$ is generated under a uniform distribution on $(0, 160)$, independent of event time $T_i$. Any event time $T_i$ larger than the corresponding censoring time $C_i$ will be set to missing. The number of candidate SNPs is 50, and the number of subjects is 2000.
Finally, each experimental setup is tested on 2000 datasets. For all the comparisons, we report the performance (sensitivity, specificity, and FDR) on three levels—SNP, genotype, and interaction. For the first two levels, as long as the associated SNPs or genotypes are correctly selected, even if they form interactions different from the true ones, we count them as correct.

For binary outcomes, Miller et al. (2009) showed through simulations that MECPM out-performs seven other popular methods, including SVMs, Pearsons $\chi^2$ test, logistic regression, information gain, and a full interaction model. For censored failure time data, the likelihood ratio test (LRT) and the least absolute shrinkage and selection operator (lasso) are commonly used variable selection methods and employed here as references for the proposed method. Greedy search method is employed again except that LRT and lasso keep 20 SNPs, that is, 60 genotypes, in the initial screen. Therefore, the search complexity of the reference methods is no less than that for MECPM, if not more. The definition of MDL is modified as
\[
\text{MDL} = -\log L_{\text{cox}} + \frac{1}{2} \log_2(n) + \log_2(5) + \log_2(C_{nm}^P \times 3^n \times 2)
\]
where $L_{\text{cox}}$ is the profile likelihood from the Cox model using the selected genotype interactions as covariates.

The reference methods and the proposed MECPM method are compared in Table 1. LRT performs well for models with single constraints or multiple constraints whose effects are in the same direction. However, for models with two genotype interactions whose effects are in opposite directions, LRT selects the genotype interaction with larger effect and misses the other one or chooses its complementary genotypes. Lasso tends to select more genotypes than the true set, resulting in low specificity and high FDR. Furthermore, Lasso cannot differentiate two order-1 constraints from one order-2 constraint. For example, there are cases where lasso combines the genotype in an order-1 constraint which increases risk and the complementary genotype in another order-1 constraint which decreases risk as a single order-2 constraint increasing the risk.

The robustness of the proposed method under various settings is compared in Table 2. The performances of the proposed method using different time intervals are compared. Two sets of time intervals are tried - evenly-distributed intervals $(0, 32)$, $(32, 64)$, $(64, 96)$, $(96, 128)$, $(128, 160)$ and $(160, +\infty)$ versus data-dependent intervals using the empirical 20th, 40th, 60th and 80th quantiles. Second, the proposed method categorizing the time axis into five quantile-dependent intervals is compared to MECPM in Miller et al., which uses binary disease outcomes. In the simulated data, subjects with event time larger than the right censoring time are considered to have no disease. Third, a larger number of intervals potentially differentiate risk of the outcome better. At the same time, finer intervals will contain less observations in each interval. The optimal number of intervals is probably somewhere in the middle, where a balance is reached. Three, five and twelve quantile-dependent intervals are tried on the same datasets and their performances are compared side by side. Fourth, the more constraints or the more complicated the constraints get, the larger the required sample size to achieve the same specificity, sensitivity and FDR. We examine the performance to identify two to four one-way constraints using sample sizes ranging from 500 to 20,000. Fifth, genotype interactions including a minor allele
with low frequency are harder to detect because less subjects satisfy such genotype interactions. Therefore, performance of the proposed method in the presence of minor allele frequencies (MAF) of 0.1, 0.2 and 0.3 is examined.

In Comparison one, the empirical quantiles outperform the evenly spaced nodes because absent of any genotype effects, the proposed method based on maximum entropy would give estimates of equal probability falling into all intervals and the quantile-based nodes distribute approximately equal number of events into each interval. In Comparison two, outcomes in the form of five intervals outperform the binary outcomes because we consider cases and controls with different lengths of followup (e.g. a control with a very short follow-up may have higher risk than a case with very long follow-up) and more intervals differentiate the levels of risk better.

Table 1. Comparison of Sensitivity(×100), Specificity(×100) and FDR(×100) of LRT, Lasso versus MECPM.

<table>
<thead>
<tr>
<th>Method</th>
<th>β</th>
<th>SNP</th>
<th>GENOTYPE</th>
<th>INTERACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SEN SPE FDR</td>
<td>SEN SPE FDR</td>
<td>SEN SPE FDR</td>
</tr>
<tr>
<td>one order-1 constraint</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRT</td>
<td>1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>LRT</td>
<td>0.5</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>LRT</td>
<td>−0.5</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>lasso</td>
<td>1</td>
<td>100 99.9 33.1</td>
<td>100 99.7 33.1</td>
<td>50.5 99.9 49.5</td>
</tr>
<tr>
<td>lasso</td>
<td>0.5</td>
<td>100 99.9 33.8</td>
<td>100 99.7 33.8</td>
<td>49.9 99.9 50.1</td>
</tr>
<tr>
<td>lasso</td>
<td>−0.5</td>
<td>100 98.7 39.8</td>
<td>100 99.6 39.8</td>
<td>34.0 99.9 66.0</td>
</tr>
<tr>
<td>MECPM</td>
<td>1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>MECPM</td>
<td>0.5</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>MECPM</td>
<td>−0.5</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>one order-2 constraint</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRT</td>
<td>1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>LRT</td>
<td>0.5</td>
<td>95.9 99.9 0.9</td>
<td>95.9 99.9 0.9</td>
<td>93.3 99.9 6.7</td>
</tr>
<tr>
<td>LRT</td>
<td>−0.5</td>
<td>93.6 99.9 1.3</td>
<td>93.2 99.9 1.8</td>
<td>89.3 99.9 10.7</td>
</tr>
<tr>
<td>lasso</td>
<td>1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>lasso</td>
<td>0.5</td>
<td>93.8 99.9 1.4</td>
<td>93.8 99.9 1.4</td>
<td>90.0 99.9 10.1</td>
</tr>
<tr>
<td>lasso</td>
<td>−0.5</td>
<td>91.0 99.9 1.7</td>
<td>90.4 99.9 2.4</td>
<td>84.7 99.9 15.4</td>
</tr>
<tr>
<td>MECPM</td>
<td>1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>MECPM</td>
<td>0.5</td>
<td>93.0 99.8 3.9</td>
<td>92.9 99.9 4.0</td>
<td>92.6 99.9 7.5</td>
</tr>
<tr>
<td>MECPM</td>
<td>−0.5</td>
<td>92.8 99.8 4.0</td>
<td>92.7 99.9 4.1</td>
<td>92.0 99.9 8.0</td>
</tr>
<tr>
<td>two order-1 constraint</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRT</td>
<td>2, −1</td>
<td>51.2 99.0 48.8</td>
<td>50.0 99.7 50.0</td>
<td>50.0 99.9 50.0</td>
</tr>
<tr>
<td>LRT</td>
<td>−2, 1</td>
<td>50.0 99.0 50.0</td>
<td>50.0 99.7 50.0</td>
<td>50.0 99.9 50.0</td>
</tr>
<tr>
<td>LRT</td>
<td>0.5, 1</td>
<td>98.8 99.9 1.2</td>
<td>98.8 99.9 1.2</td>
<td>98.8 99.9 1.2</td>
</tr>
<tr>
<td>LRT</td>
<td>−0.5, −1</td>
<td>99.3 99.9 0.8</td>
<td>99.3 99.9 0.8</td>
<td>99.3 99.9 0.8</td>
</tr>
<tr>
<td>lasso</td>
<td>2, −1</td>
<td>30.1 98.8 52.6</td>
<td>28.5 99.5 59.9</td>
<td>10.6 99.9 89.4</td>
</tr>
<tr>
<td>lasso</td>
<td>−2, 1</td>
<td>29.6 98.5 55.4</td>
<td>25.0 99.4 62.3</td>
<td>8.6 99.9 91.4</td>
</tr>
<tr>
<td>lasso</td>
<td>0.5, 1</td>
<td>50.1 99.4 23.2</td>
<td>50.1 99.8 23.2</td>
<td>29.1 99.9 70.9</td>
</tr>
<tr>
<td>lasso</td>
<td>−0.5, −1</td>
<td>48.6 99.3 25.1</td>
<td>48.6 99.8 25.1</td>
<td>26.5 99.9 73.5</td>
</tr>
<tr>
<td>MECPM</td>
<td>2, −1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>MECPM</td>
<td>−2, 1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>MECPM</td>
<td>0.5, 1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
<tr>
<td>MECPM</td>
<td>−0.5, −1</td>
<td>100 100 0</td>
<td>100 100 0</td>
<td>100 100 0</td>
</tr>
</tbody>
</table>
In the third Comparison, five intervals perform best, which coincides with the recommendation of five to six pieces in fitting piecewise exponential models\textsuperscript{11} as well as the statement in Cochran\textsuperscript{6} that the choice of five or six subclasses/strata would effectively remove 90\% or more of the bias from using adjustment of subclassification or stratification for a continuous underlying variable. In Comparison four, the performance of the proposed method depends on the number of genotypes and sample size, regardless of whether the genotypes form one high order constraint or multiple order-1 constraints. For constraints composed of a couple of genotypes, 500 is a sufficient sample size. However, for constraints containing more genotypes, sample

Table 2. Comparisons for optimal intervals and impacts of constraint complexity and MAF.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>( n_m )</th>
<th>( \beta )</th>
<th>SEN (SNP)</th>
<th>SPE (SNP)</th>
<th>FDR (SNP)</th>
<th>SEN (Genotype)</th>
<th>SPE (Genotype)</th>
<th>FDR (Genotype)</th>
<th>SEN (Interaction)</th>
<th>SPE (Interaction)</th>
<th>FDR (Interaction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Even</td>
<td>2</td>
<td>2</td>
<td>85.6</td>
<td>99.1</td>
<td>4.1</td>
<td>85.6</td>
<td>99.7</td>
<td>4.1</td>
<td>56.2</td>
<td>99.9</td>
<td>46.9</td>
</tr>
<tr>
<td>1 Even</td>
<td>2</td>
<td>1</td>
<td>46.1</td>
<td>98.8</td>
<td>6.7</td>
<td>46.1</td>
<td>99.4</td>
<td>6.7</td>
<td>7.3</td>
<td>99.9</td>
<td>93.7</td>
</tr>
<tr>
<td>1 Even</td>
<td>2</td>
<td>0.5</td>
<td>31.2</td>
<td>93.5</td>
<td>47</td>
<td>31.2</td>
<td>98.1</td>
<td>47</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>1 Quantile</td>
<td>2</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>1 Quantile</td>
<td>2</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>1 Quantile</td>
<td>2</td>
<td>0.5</td>
<td>92.7</td>
<td>98.3</td>
<td>1.8</td>
<td>92.7</td>
<td>99.8</td>
<td>1.8</td>
<td>87.6</td>
<td>99.9</td>
<td>12.3</td>
</tr>
<tr>
<td>2</td>
<td>1/0</td>
<td>2</td>
<td>83.7</td>
<td>97.9</td>
<td>17.7</td>
<td>83.7</td>
<td>99.4</td>
<td>17.7</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>1/0</td>
<td>2</td>
<td>34.1</td>
<td>96.4</td>
<td>30.3</td>
<td>34.1</td>
<td>98.9</td>
<td>30.3</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>1/0</td>
<td>2</td>
<td>15.9</td>
<td>91.2</td>
<td>74.4</td>
<td>15.9</td>
<td>97.5</td>
<td>74.4</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>2</td>
<td>92.7</td>
<td>98.3</td>
<td>1.8</td>
<td>92.7</td>
<td>99.8</td>
<td>1.8</td>
<td>87.6</td>
<td>99.9</td>
<td>12.3</td>
</tr>
<tr>
<td>2</td>
<td>1/0</td>
<td>3</td>
<td>10.7</td>
<td>89.2</td>
<td>64.2</td>
<td>5.4</td>
<td>97.2</td>
<td>72.6</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>1/0</td>
<td>3</td>
<td>8.3</td>
<td>82.6</td>
<td>83.3</td>
<td>2.6</td>
<td>94.8</td>
<td>89.7</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>1/0</td>
<td>3</td>
<td>5.7</td>
<td>81.7</td>
<td>87.9</td>
<td>0.4</td>
<td>91.9</td>
<td>94.2</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>3</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>3</td>
<td>78.3</td>
<td>97.1</td>
<td>7.8</td>
<td>71.6</td>
<td>98.5</td>
<td>13.7</td>
<td>67.9</td>
<td>99.9</td>
<td>27.4</td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>3</td>
<td>18.7</td>
<td>87.1</td>
<td>78.2</td>
<td>1.2</td>
<td>93.6</td>
<td>85.8</td>
<td>0</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>K = 3</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>K = 3</td>
<td>0.5</td>
<td>55.7</td>
<td>96.9</td>
<td>17.5</td>
<td>55.7</td>
<td>99.1</td>
<td>17.5</td>
<td>32.7</td>
<td>99.9</td>
<td>59.4</td>
</tr>
<tr>
<td>3</td>
<td>K = 5</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>K = 5</td>
<td>0.5</td>
<td>92.7</td>
<td>98.3</td>
<td>1.8</td>
<td>92.7</td>
<td>99.8</td>
<td>1.8</td>
<td>87.6</td>
<td>99.9</td>
<td>12.3</td>
</tr>
<tr>
<td>3</td>
<td>K = 12</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>K = 12</td>
<td>0.5</td>
<td>36.1</td>
<td>95.7</td>
<td>23</td>
<td>19.8</td>
<td>96.9</td>
<td>51.6</td>
<td>26.4</td>
<td>99.9</td>
<td>73.5</td>
</tr>
<tr>
<td>4</td>
<td>N = 500</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>N = 1000</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>N = 2000</td>
<td>3</td>
<td>90.9</td>
<td>89.4</td>
<td>36.3</td>
<td>84.3</td>
<td>98.4</td>
<td>31.7</td>
<td>42.6</td>
<td>99.9</td>
<td>56.8</td>
</tr>
<tr>
<td>4</td>
<td>N = 20000</td>
<td>3</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>N = 20000</td>
<td>4</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>MAF = 0.3</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>MAF = 0.3</td>
<td>0.5</td>
<td>92.4</td>
<td>98.1</td>
<td>4.2</td>
<td>92.4</td>
<td>99.6</td>
<td>4.2</td>
<td>86.9</td>
<td>99.9</td>
<td>13.4</td>
</tr>
<tr>
<td>5</td>
<td>MAF = 0.2</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>MAF = 0.2</td>
<td>0.5</td>
<td>87.9</td>
<td>97.1</td>
<td>2.8</td>
<td>87.9</td>
<td>99.4</td>
<td>2.8</td>
<td>85.1</td>
<td>99.9</td>
<td>15.4</td>
</tr>
<tr>
<td>5</td>
<td>MAF = 0.1</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>MAF = 0.1</td>
<td>0.5</td>
<td>85.1</td>
<td>95.7</td>
<td>4.7</td>
<td>85.1</td>
<td>98.5</td>
<td>4.7</td>
<td>82.9</td>
<td>99.9</td>
<td>16.1</td>
</tr>
</tbody>
</table>

In the third Comparison, five intervals perform best, which coincides with the recommendation of five to six pieces in fitting piecewise exponential models\textsuperscript{11} as well as the statement in Cochran\textsuperscript{6} that the choice of five or six subclasses/strata would effectively remove 90\% or more of the bias from using adjustment of subclassification or stratification for a continuous underlying variable. In Comparison four, the performance of the proposed method depends on the number of genotypes and sample size, regardless of whether the genotypes form one high order constraint or multiple order-1 constraints. For constraints composed of a couple of genotypes, 500 is a sufficient sample size. However, for constraints containing more genotypes, sample
sizes as large as 2000 are necessary. In Comparison five, low MAF reduces sensitivity in the presence of weaker associations ($\beta = 0.5$).

4. Real Data Analysis

Microvascular complications are one of the main threats to the expectancy and quality of life among diabetic patients. The Diabetes Control and Complications Trial (DCCT) and the subsequent Epidemiology of Diabetes Intervention and Complication (EDIC) study is a randomized clinical trial targeting better monitoring of blood glucose level, hence reducing the risk of complications in Type 1 diabetic patients. The DCCT/EDIC cohort provides a rich resource of clinical and genomic information on Type 1 diabetes. We are specifically interested in detecting genotype interactions associated with microvascular complication symptoms in the neuronal system - peripheral diabetic neuropathy.

In total, 695 males and 609 females from the EDIC/DCCT cohort participated in the ancillary GWAS study, and 1039611 SNPs were mapped throughout their genomes. We started from a restricted set of genes which had shown association with diabetic neuropathy in the literature $^{18,22,24}$ — ACE, MTHFR, GLO1, APOE, VEGF, IL-4, GPX1, cNOS, ADRA2B, MIR146A, MIR128A, GFRA2, GSST1, TCF7L2. In the Single Nucleotide Polymorphism Database (dbSNP) hosted by the National Center for Biotechnology Information (NCBI), the 14 neuropathy-related genes contain 29188 SNPs, among which 305 were mapped in the 1304 DCCT/EDIC participants. A series of quality checks including minor allele frequency (MAF) less than 0.1, missing percentage higher than 0.05 and missing probability significantly associated with the flanking genotypes or outcomes, Hardy–Weinberg equilibrium (HWE), were run to clean the SNP data. After pre-processing, 118 SNPs were removed and 187 remained.

Clinical neuropathy is defined as having at least two positive responses among clinical symptoms, sensory signs or reflex changes consistent with a distal symmetrical polyneuropathy examined by a neurologist. Electro-diagnostic neuropathy is defined as abnormalities involving two or more nerves among the median, peroneal, and sural nerves in the nerve conduction exam. A participant is diagnosed with neuropathy when either clinical neuropathy or electro-diagnostic neuropathy is detected. Due to the lengthy procedure and high costs associated with neuropathy tests, neuropathy symptoms were assessed four times over the 26 years of follow-up at DCCT baseline, DCCT closeout, EDIC year 13 and EDIC year 16, respectively. The number of neuropathy cases in the first four intervals are 416, 241, 126 and 77, respectively. There are 444 patients without any neuropathy symptoms by EDIC year 16, falling in the 5th time interval. We combine the 2nd, 3rd and 4th intervals in the real data so that each interval contains approximately equal number of subjects — 416, 444, 444.

Table 3 shows the first seven constraints selected by the MECPM algorithm. We listed all genotype interactions before the MDL starts increasing rapidly. If a
genotype interaction increases the probability of neuropathy occurrence in the first two intervals, it is equivalent to say that it decreases neuropathy probability in the third interval. The algorithm always picks the constraint satisfied by more subjects; therefore the longer intervals \(\{1,2\}\) or \(\{2,3\}\) are chosen, rather than \(\{3\}\) or \(\{1\}\). Among the identified SNPs, SNP rs10885395 is associated with small intestinal Crohn’s disease\(^{16}\) and rs699947 is associated with renal cell carcinoma\(^{19}\) in other independent studies, both of which are microvascular-related,\(^{14}\) similar to neuropathy. Furthermore, rs699947 genotype TT is associated with the last time interval with a \(p\)-value of 0.0025 using a Pearson’s Chi-square test for the association between genotype (TT versus TA+AA) and time intervals. However, \(p\)-values for the other identified genotype interactions are larger than 0.05, possibly due to small sample sizes satisfying the genotype interactions. The two-way genotype interaction (rs2736655(AT), rs6585196(CC)) will be missed if genotypes are examined one by one. For the 366 subjects that have genotype rs2736655(AT), the percentage falling into the third interval is 35%. For the 88 subjects that satisfy genotype rs6585196(CC), the percentage falling into the third interval is 34%. However, when a subject possesses both genotype rs2736655(AT) and rs6585196(CC), the probability of falling into the third interval increases to 47%. Furthermore, we also tried LRT and Lasso methods on the real data and genotype rs11196171(GA) is identified by both methods, which is also a top candidate in the initial screening of the MECPM method. Genotype rs11196171(GA) is statistically significant in a Pearson’s Chi-Square test of independence between the genotype (GA versus GG+AA) and time intervals (30% in interval 1, 31% in interval 2, 39% in interval 3; \(p\)-value=0.0071) as well as a Cox model (Hazards Ratio=0.906, \(p\)-value=0.086). All three methods fail to report genotype interactions strongly associated with neuropathy in the real data although MECPM provides slightly more plausible candidates. This is not unusual for complex clinical outcomes such as neuropathy. The clinical diagnosis of neuropathy are inaccurate and examinations often give different diagnosis because patients’ status and experiences of the neurologist change. Besides, the real neuropathy occurrence times are never observed and we approximate the neuropathy occurrence times by the fixed

<table>
<thead>
<tr>
<th>Constraint</th>
<th>SNP(Genotype)</th>
<th>Time intervals</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs11775754,T(AA)</td>
<td>{1,2}</td>
<td>GFRA2</td>
</tr>
<tr>
<td>2</td>
<td>rs7081062,G(CC)</td>
<td>{1,2}</td>
<td>TCF7L2</td>
</tr>
<tr>
<td>3</td>
<td>rs2736655,A(AT), rs6585196,C(CC)</td>
<td>{1,2}</td>
<td>GLO1,TCF7L2</td>
</tr>
<tr>
<td>4</td>
<td>rs10885395,T(TT)</td>
<td>{1,2}</td>
<td>TCF7L2</td>
</tr>
<tr>
<td>5</td>
<td>rs4739214,A(AA)</td>
<td>{2,3}</td>
<td>GFRA2</td>
</tr>
<tr>
<td>6</td>
<td>rs4277044,A(TT)</td>
<td>{1,2}</td>
<td>TCF7L2</td>
</tr>
<tr>
<td>7</td>
<td>rs699947,A(TT)</td>
<td>{1,2}</td>
<td>VEGFA</td>
</tr>
</tbody>
</table>
screening times. Finally, the mechanism to develop neuropathy is complex involving environmental and epigenetic factors beyond SNP and the effects of SNP genotype interactions are often trivial.\(^1\)

5. Discussion

This paper extends the MECPM method originally proposed for the analysis of binary outcomes to disease occurrences over time, i.e. failure times. Failure times are discretized into time intervals. Censored cases are also incorporated using the EM algorithm. The proposed method is compared to the commonly used LRT and lasso methods for the Cox model, with the greedy search algorithm employed in all three methods. Our method has better specificity and FDR, especially in identifying the exact form of the genotype interactions or multiple interactions. Unsurprisingly, disease outcomes in the form of multiple intervals provide higher power than binary outcomes. Five to six-time intervals are recommended. Furthermore, quantile-dependent time intervals perform better than evenly spaced time intervals. The performance depends on the ratio of sample size to number of genotypes. In usual studies with moderate sample size (e.g. 1304 in the DCCT/EDIC example), it is recommended to narrow down the candidate SNP pool with preliminary screening, such as literature review or one-at-a-time regressions. Furthermore, it is better not to keep “redundant” genotypes that possess high linkage disequilibrium with each other. When a genotype is significantly associated with a phenotype, other nearby linked genotypes are likely to be associated as well, though most of these are false positives. Instead of the truly associated genotype interactions, combinations of other linked genotypes may be selected, leading to low specificity and high FDR. Finally, the proposed MECPM method does not infer statistical significance. The final report of the algorithm is a list of possibly influential genotype interactions that require further confirmatory studies.

The MECPM with censored disease occurrence times can be used to screen biomarker interactions other than SNPs. For example, interaction between mRNA and micro RNA in expression studies or interaction between metabolites in metabolomics study. The expression of a continuous biomarker level can be categorized as under-expressed, normal, and over-expressed.

Acknowledgments

Aotian Yang and Qing Pan were supported by Award Numbers UL1TR000075 and KL2TR000076 from the NIH National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.
Appendix A. Proof of Proposition 1

It can be seen that \( \frac{\partial^2 H}{\partial r_i \partial r_j} = -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{P(r_i=k|G)} < 0 \) for \( k = 1, \ldots, K \), and \( \frac{\partial^2 H}{\partial r_i \partial r_l} = 0 \) for any \( k, l = 1, \ldots, K, k \neq l \). The Hessian matrix of \( H \) is negative definite. Thus, \( H \) is a concave function of \( P(k|G) \).

The Lagrangian method is used here to find the best solution. Let \( P(b) \) denote the optimization problem that maximizes \( f(x) \) subject to \( h(x) = b, x \in X \). Let \( x \in X(b) = \{ x \in X : h(x) = b \} \). We say that \( x \) is feasible if \( x \in X(b) \). We define the Lagrangian as

\[
L(x, \lambda) = f(x) + \lambda^\top(h(x) - b).
\]

Typically, \( X \subseteq \mathbb{R}^n, h: \mathbb{R}^n \mapsto \mathbb{R}^m \), with \( b, \lambda \in \mathbb{R}^m \). Here \( \lambda \) is a vector of Lagrange multipliers.

Note that if \( \bar{x} \) is feasible for \( P(b) \) and there exists \( \bar{\lambda} \) such that

\[
\sup_{x \in X} L(x, \bar{\lambda}) = L(\bar{x}, \bar{\lambda}),
\]

then \( \bar{x} \) is optimal for \( P(b) \). We prove it as follows:

\[
\forall x \in X(b) \text{ and } \lambda, \text{ we have } f(x) = f(x) + \lambda^\top(h(x) - b) = L(x, \lambda).
\]

Now \( \bar{x} \in X(b) \subseteq X \) and so by assumption,

\[
f(\bar{x}) = L(\bar{x}, \bar{\lambda}) \geq L(x, \bar{\lambda}) = f(x)
\]

for all \( x \in X(b) \). Thus, \( \bar{x} \) is optimal for the optimization problem \( P(b) \).

Now, we let \( h = h(x) \), so that \( f(x) \) would be a function of \( h \): \( f(x) = \tilde{f}(h) \). Our problem is converted into maximizing \( \tilde{f}(h) \) under the constraint that \( h = b \).

Now, fixing \( \lambda \), we need to find the \( \tilde{h} \) that maximizes \( L(\tilde{h}, \lambda) \). Therefore, \( \tilde{h} \) is a function of \( \lambda \) and \( L_{\max} = g(\lambda) \). To find such \( \tilde{h} \), we take the derivative of \( L \) with respect to \( h \):

\[
\frac{dL}{dh} = \tilde{f}'(\tilde{h}) - \lambda.
\]

And by making \( \frac{dL}{dh} = 0 \) we have \( \tilde{f}'(\tilde{h}) = \lambda \). Therefore, \( L_{\max} = g(\lambda) = g(\tilde{f}'(\tilde{h})) = \varphi(\tilde{h}) \). The maximum of \( \varphi(\tilde{h}) \) is obtained when \( \tilde{h} = b \). Since

\[
\varphi(\tilde{h}) = \tilde{f}(\tilde{h}) - \tilde{f}'(\tilde{h})(\tilde{h} - b),
\]

we have

\[
\frac{d\varphi(\tilde{h})}{dh} = \tilde{f}'(\tilde{h}) - \tilde{f}'(\tilde{h}) - \tilde{f}''(\tilde{h})(\tilde{h} - b) = -\tilde{f}''(\tilde{h})(\tilde{h} - b).
\]
When $f(h)$ is a concave function, $\hat{f}''(\hat{h}) < 0$,

\[
\hat{h} = b, \quad \frac{d\varphi(\hat{h})}{dh} = 0,
\]

\[
\hat{h} > b, \quad \frac{d\varphi(\hat{h})}{dh} > 0,
\]

\[
\hat{h} < b, \quad \frac{d\varphi(\hat{h})}{dh} < 0.
\]

In our case $f(p) = -p \log p$ and thus $f''(p) = -\frac{1}{p} < 0$. $f(p)$ therefore is a concave function, $\hat{p} = b$ is the maximum point of $\varphi(\hat{p})$, also the maximum point of $g(\lambda)$.

**Appendix B. Proof of Proposition 2**

We prove proposition 2 for $H_i, L_i, \tilde{L}_i$ from the $i$th item and then sum across all subjects $i = 1, \ldots, n$. And we use $t_i$ to denote the observed value of random variable $T_i$, the event occurrence interval. Plugging the posterior probability back into Shannon’s entropy $H$, for each subject $i$, we can rewrite the Lagrangian expression. First take a look at $H$:

\[
H_i = - \sum_{k=1}^{K} P_{ME}[T_i = k|G_i] \log \sum_{k'=1}^{K} \exp(\sum_{m'=1}^{N} \gamma_{m'} I[k' \in R_{m'}, G_{i(m')} = g_{m'}])
\]

\[
= - \sum_{k=1}^{K} P_{ME}[T_i = k|G_i] \left( \sum_{m'=1}^{N} \gamma_{m'} I[k' \in R_{m'}, G_{i(m')} = g_{m'}] \right)
\]

\[
+ \log \left( \sum_{k'=1}^{K} \exp(\sum_{m'=1}^{N} \gamma_{m'} I[k' \in R_{m'}, G_{i(m')} = g_{m'}]) \right).
\]

For simplicity, we write

\[
H_i = - A_i + B_i
\]

\[
A_i = \sum_{k=1}^{K} P_{ME}[T_i = k|G_i] \left( \sum_{m'=1}^{N} \gamma_{m'} I[k' \in R_{m'}, G_{i(m')} = g_{m'}] \right)
\]

\[
B_i = \log \left( \sum_{k'=1}^{K} \exp(\sum_{m'=1}^{N} \gamma_{m'} I[k' \in R_{m'}, G_{i(m')} = g_{m'}]) \right).
\]

Let

\[
C_i = \sum_{m=1}^{N} \gamma_m ((P_{ME}[T_i \in R_m|G_{i(m)} = g_m] I[G_{i(m)} = g_m]),
\]

\[
D_i = \sum_{m=1}^{N} \gamma_m I[t_i \in R_m, G_{i(m)} = g_m].
\]
We have \( L_i = -A_i + B_i + C_i - D_i \). Note for \( m \) such that \( \mathcal{G}_{i(m)} \neq \mathcal{g}_m \)
\[ A_i = B_i = C_i = D_i = 0. \]
So we only need to consider the case that \( \mathcal{G}_{i(m)} = \mathcal{g}_m \).
\[
A_i = \sum_{m=1}^{N_c} \gamma_m \sum_{k=1}^{K} P_{ME}[k | \mathcal{G}_i] I[k \in R_m, \mathcal{G}_{i(m)} = \mathcal{g}_m];
\]
\[
C_i = \sum_{m=1}^{N_c} \gamma_m \sum_{k=1}^{K} P_{ME}[k | \mathcal{G}_{i(m)} = \mathcal{g}_m] I[\mathcal{G}_{i(m)} = \mathcal{g}_m].
\]
Hence \( A_i = C_i \). Therefore,
\[
\hat{L}_i = -A_i + B_i + C_i - D_i
= -D_i + B_i
= -\sum_{m=1}^{N_c} \gamma_m I[t_i \in R_m, \mathcal{G}_{i(m)} = \mathcal{g}_m] + B_i
= -\log \left( \exp \left( \sum_{m=1}^{N_c} \gamma_m I[t_i \in R_m, \mathcal{G}_{i(m)} = \mathcal{g}_m] \right) \right) + B_i
= -\log \left( \sum_{k=1}^{K} I[k = t_i] \exp \left( \sum_{m=1}^{N_c} \gamma_m I[k \in R_m, \mathcal{G}_{i(m)} = \mathcal{g}_m] \right) \right)
+ \log \left( \sum_{k=1}^{K} \exp \left( \sum_{m=1}^{N_c} \gamma_m I[k' \in R_{m'}, \mathcal{G}_{i(m')} = \mathcal{g}_{m'}] \right) \right)
= -\log \left( \frac{\exp \left( \sum_{m=1}^{N_c} \gamma_m I[t_i \in R_m, \mathcal{G}_{i(m)} = \mathcal{g}_m] \right)}{\sum_{k=1}^{K} \exp \left( \sum_{m=1}^{N_c} \gamma_m I[k' \in R_{m'}, \mathcal{G}_{i(m')} = \mathcal{g}_{m'}] \right)} \right).
\]
Summing over \( i = 1, \ldots, n \),
\[
L = -\sum_{i=1}^{n} \log \left\{ \frac{\exp \left( \sum_{m=1}^{N_c} \gamma_m I[t_i \in R_m, \mathcal{G}_{i(m)} = \mathcal{g}_m] \right)}{\sum_{k=1}^{K} \exp \left( \sum_{m=1}^{N_c} \gamma_m I[k' \in R_{m'}, \mathcal{G}_{i(m')} = \mathcal{g}_{m'}] \right)} \right\}.
\]

**Appendix C. Proof for Proposition 3**

Using the inequality \( -\log(\alpha) \geq 1 - \alpha \),
\[
Q(\gamma + \Delta \gamma) - Q(\gamma)
\geq \sum_{i=1}^{n} \sum_{m=1}^{N_c} \Delta \gamma_m I(t_i \in R_{m,m(m)} = \mathcal{g}_m) \Delta_i
\]
A. Yang, D. Miller & Q. Pan

\[ + \sum_{i=1}^{n} \mathbb{E}_{T_i|C_i \leq T_i \leq C_i \gamma G} \left\{ \sum_{m=1}^{N_c} \Delta \gamma_m I(T_i \in R_m, G_i(m) = g_m) \right\} (1 - \Delta_i) \]

\[ + \sum_{i=1}^{n} \left( 1 - \frac{\sum_{k'=1}^{K} \exp\left( \sum_{m=1}^{N_c} (\gamma_m + \Delta \gamma_m) I(k' \in R_m, G_i(m) = g_m) \right)}{\sum_{k'=1}^{K} \exp\left( \sum_{m=1}^{N_c} \gamma_m I(k' \in R_m, G_i(m) = g_m) \right)} \right) \]

\[ = \sum_{i=1}^{n} \left( \sum_{m=1}^{N_c} \Delta \gamma_m I(t_i \in R_m, G_i(m) = g_m) + 1 \right) \Delta_i \]

\[ + \sum_{i=1}^{n} \left( \mathbb{E}_{T_i|C_i \leq T_i \leq C_i \gamma G} \left\{ \sum_{m=1}^{N_c} \Delta \gamma_m I(T_i \in R_m, G_i(m) = g_m) \right\} + 1 \right) (1 - \Delta_i) \]

\[ - \sum_{i=1}^{n} \sum_{k'=1}^{K} \left( \hat{P}_{ME}(k'|G_i) \exp\left( \sum_{m=1}^{N_c} \frac{1}{N_c} \Delta \gamma_m I(k' \in R_m, G_i(m) = g_m) \right) \right) \]

By Jensen’s inequality, \( e \sum x a(x)b(x) \leq \sum x a(x)e^{b(x)} \),

\[ Q(\gamma + \Delta \gamma) - Q(\gamma) \]

\[ \geq \sum_{i=1}^{n} \left( \sum_{m=1}^{N_c} \Delta \gamma_m I(t_i \in R_m, G_i(m) = g_m) + 1 \right) \Delta_i \]

\[ + \sum_{i=1}^{n} \left( \mathbb{E}_{T_i|C_i \leq T_i \leq C_i \gamma G} \left\{ \sum_{m=1}^{N_c} \Delta \gamma_m I(T_i \in R_m, G_i(m) = g_m) \right\} + 1 \right) (1 - \Delta_i) \]

\[ - \frac{1}{N_c} \sum_{i=1}^{n} \sum_{k'=1}^{K} \sum_{m=1}^{N_c} \hat{P}_{ME}(k'|G_i) \exp\{N_c \Delta \gamma_m I(k' \in R_m, G_i(m) = g_m)\} \]

References


**Aotian Yang** is a quantitative Research Associate working at JPMorgan & Chase Company. He received a B.S. degree in Information and Computing Science from Dalian University of Technology in 2010, and an M.S. degree in Statistics from University of California, San Diego in 2012. He received his Ph.D. degree in Statistics from the George Washington University in 2017. His research interests focus on machine learning methods in bioinformatical applications.

**David Miller** received a B.S.E. degree from Princeton University in 1987, an M.S.E. degree from University of Pennsylvania in 1990, and a Ph.D. degree from the University of California, Santa Barbara (UCSB) in 1995, all in electrical engineering. He was also employed as a communications engineer by General Atronics Corp., Wyndmoor PA, from 1988–1990. Dr. Miller joined Penn State’s Electrical Engineering Department in 1995, was promoted to tenured Associate Professor in 2001 and to Full Professor in 2007. Dr. Miller is an active researcher in the areas of machine learning, pattern recognition, neural networks, data compression, image processing and image-based pattern recognition, bioinformatics, data mining and intrusion detection for computer networks, source and channel coding, and statistical estimation. Dr. Miller did early influential work on semisupervised learning (more than 2500 combined citations and secondary citations) and semisupervised class discovery. Dr. Miller received an NSF CAREER Award in 1996. He was Chair of the Machine Learning for Signal Processing Technical Committee, within the IEEE Signal Processing Society from 2007–2009. Dr. Miller was an Associate Editor for IEEE Transactions on Signal Processing from 2004–2007.
Qing Pan is tenured Associate Professor of Statistics at George Washington University. She received a B.S. degree in Biotechnology from Beijing University in 2000, an M.S. degree in Statistics from University of Georgia in 2003 and a Ph.D. degree in Biostatistics from University of Michigan in 2007. Dr. Pan's research focuses on novel statistical and machine learning methods with applications in biostatistics and bioinformatics including survival analysis, electronic health records, network analysis and “omics” data. She was/is an important investigator in the Prospective Payment System for End Stage Renal Disease, Scientific Registry of Transplant Recipients, Diabetes Prevention Program, Antibacterial Resistance Leadership Group.