Received: 11 November 2021 | Revised: 2 March 2022 | Accepted: 8 April 2022

DOI: 10.1111/ahg.12470

Annals of human genetics WILEY

ORIGINAL ARTICLE

Using potential variable to study gene–gene and gene–environment interaction effects with genetic model uncertainty

Xiaonan Hu[1] | Zhen Meng[2]

[1]NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

[2]School of Statistics, Capital University of Economics and Business, Beijing, China

Correspondence
Zhen Meng, School of Statistics, Capital University of Economics and Business, Xincun street, Fengtai District, 100070, Beijing Municipality, China
Email: mengz@cueb.edu.cn

Funding information
Capital University of Economics and Business Newly Recruited Young Teachers Scientific Research Start-up Foundation Project

Abstract: One of the critical issues in genetic association studies is to evaluate the risk of a disease associated with gene–gene or gene–environment interactions. The commonly employed procedures are derived by assigning a particular set of scores to genotypes. However, the underlying genetic models of inheritance are rarely known in practice. Misspecifying a genetic model may result in power loss. By using some potential genetic variables to separate the genotype coding and genetic model parameter, we construct a model-embedded score test (MEST). Our test is free of assumption of gene–environment independence and allows for covariates in the model. An effective sequential optimization algorithm is developed. Extensive simulations show the proposed MEST is robust and powerful in most of scenarios. Finally, we apply the proposed method to rheumatoid arthritis data from the Genetic Analysis Workshop 16 to further investigate the potential interaction effects.

KEYWORDS
gene–environment interaction, gene–gene interaction, genetic model, potential variable, power

1 | INTRODUCTION

Large-scale genome-wide association studies (GWAS) have successfully discovered hundreds of thousands of single nucleotide polymorphisms (SNPs), which are significantly associated with human complex diseases. However, these loci can only explain a small portion of heritability (Manolio et al., 2009). Since there is a multifactorial etiology of human complex diseases, testing for association incorporating gene–gene or gene–environment interactions may provide another way to identify the missing deleterious genetic variations and more insights into biological mechanisms of disease (Carlborg & Haley, 2004; Marchini et al., 2005).

Population-based case–control studies have been widely used to study the association of gene–gene or gene–environment interactions with the complex human diseases. Several procedures are presented to evaluate gene–gene or gene–environment interactions in the literature. Standard prospective logistic regression analysis for case–control data often yields poor power for detecting the interaction effects due to small sample size of cases or controls. Under gene–environment independence and rare disease assumption, the gene–environment interaction odds ratio can be estimated in a much more precise fashion with case data alone than traditional case–control analysis (Piegorsch et al., 1994), which is extended to a general setting by fitting a suitably constrained log-linear model to case–control data (Umbach & Weinberg, 1997). Further, Chatterjee and Carroll (2005) developed a semiparametric framework for retrospective maximum-likelihood estimate of case–control study under gene–environment independence assumption that it may involve continuous exposures, not require rare disease and account for the gene–environment dependence due to population stratification. Obviously, gene–environment independence is a critical condition for practical application based on above methods, which can lead to serious bias if violated (Albert et al., 2001; Chatterjee & Carroll, 2005; Mukherjee & Chatterjee, 2008; Mukherjee et al., 2008). To trade-off between bias and efficiency, a two-stage procedure is naturally adopted, where one tests for the gene–environment independence based on the data at the first stage and then uses the results to decide whether to adopt the efficient retrospective or the robust prospective method for inference. However, the test in the first stage typically has low power with a modest sample size, and consequently the two-stage procedure still remains significantly biased (Albert et al., 2001). Some Bayesian procedures, such as empirical Bayes-type shrinkage estimator (Mukherjee & Chatterjee, 2008), Bayesian model averaging (Li & Conti, 2009) and full Bayes analysis (Mukherjee et al., 2010), have also been investigated for interaction analysis. These procedures incorporate uncertainty of the independence assumption in a data-adaptive way. In addition, Han, Rosenberg, and Chatterjee (2012) developed multiple gene–environment or gene–gene tests under some biologically plausible constraints through bivariate isotonic regressions for case–control data, and further followed Chatterjee and Carroll (2005) to exploit the independence condition. A likelihood ratio test for detecting additive interactions of case–control studies incorporating the gene–environment independence has been further investigated (Han, Rosenberg, Garcia-Closas et al., 2012). For longitudinal study, He et al. (2017) proposed a set-based test for gene–environment interaction also considering the independence assumption. In recent years, there are still some related studies on this topic for large-scale biobank data, for example, the case-only method for interaction analysis using polygenic risk scores (Meisner et al., 2019), an efficient mixed-model association test (Wang et al., 2020) and the tree-ensemble method (Johnsen et al., 2021).

Existing methods always have some limitations as follows. First, since the underlying gene–gene or gene–environment independence is usually a major assumption for efficiency of the methods, it still need to be validated in practice; second, some of the above approaches have been constructed under a simple setting that only involving genetic and environmental factors with categorical exposures; last, all the methods are derived under a specified genetic model that is a functional relationship of a risk measure given genotypes (Hu et al., 2019; Zheng et al., 2016). As we all know, the commonly used genetic models are recessive (REC), additive (ADD), dominate (DOM) models, among which the test derived under the additive model is often adopted in the literature. Due to that the underlying genetic model is rarely known in practice, test departure from the true genetic models will lead to power loss. Similar to single-marker analysis in GWAS, there is no optimal test for all genetic models, and then a robust test free of model assumption is a better choice.

To tackle these limitations, we first suggest using a potential genetic variable to separate the genetic coding and genetic model parameter. Then we can construct a family of score test statistics varying a suitable range of model parameters, and propose to use the maximum one as our final test statistic, which is termed as a model-embedded score test (MEST). Different from previous research, we contribute to borrow the robustness of traditional logistic regression analysis of case–control study, and then construct a flexible model setting allowing for potential confounders and considering the genetic model uncertainty. Besides, our procedure is also practicable for testing both joint and interaction effects. Simulation studies and real application are conducted to compare the performance between the MEST and some existing procedures.

2 | MATERIALS AND METHODS

2.1 | Model-embedded score test

Consider a biallelic SNP locus with two alleles A and a. Without loss of generality, we assume A is the risk allele. Denote the allele frequency of A by p. Under the Hardy–Weinberg Equilibrium principal, the genotype frequencies of AA, Aa and aa are p^2, $2p(1-p)$ and $(1-p)^2$, respectively. We mainly use the following two logistic regression models for testing gene–gene and gene–environment interactions, respectively.

Model I (Gene–gene interaction model)

$$\begin{aligned}\operatorname{logit}\Pr(Y_k = 1) &= \gamma_0 + \mathbf{X_k}\boldsymbol{\gamma}_1 + G_{1k}\beta_1 + G_{2k}\beta_2\\ &\quad + G_{1k}G_{2k}\,\beta_{12} = \mathbf{D_{1k}}\,\boldsymbol{\Theta}\end{aligned} \tag{1}$$

Model II (Gene–environment interaction model)

$$\begin{aligned}\operatorname{logit}\Pr(Y_k = 1) &= \gamma_0 + \mathbf{X_k}\boldsymbol{\gamma}_1 + E_k\beta_1 + G_{1k}\beta_2\\ &\quad + G_{1k}E_k\,\beta_{12} = \mathbf{D_{2k}}\,\boldsymbol{\Theta}\end{aligned} \tag{2}$$

where $\operatorname{logit}(v) = \log(v/(1-v)), v \in (0,1)$. In both models, Y_k is a binary outcome taking value 0 or 1, where $Y_k = 1(Y_k = 0)$ represents a case (control), X_k denote the covariates, $\mathrm{G}_{1k}, \mathrm{G}_{2k}$ are genotypes of SNPs, and E_k is an environmental exposure, for $k = 1, 2, \ldots, n$. Let $\boldsymbol{\Theta} = (\gamma_0, \boldsymbol{\gamma}_1^{\tau}, \beta_1, \beta_2, \beta_{12})^{\tau}$, $\mathbf{D_{1k}} = (1, \mathbf{X_k}, G_{1k}, G_{2k}, G_{1k}G_{2k})$ and $\mathbf{D_{2k}} = (1, \mathbf{X_k}, E_k, G_{1k}, G_{1k}E_k)$.

Notice that the additional covariates X_k included in the model are some potential confounders, such as age, gender, and ethnicity, etc. (Chatterjee & Carroll, 2005), while the environmental factor is the covariate of main concern in our study. As an example, the interaction between smoking and N-acetyltransferase 2 (NAT2) has been well-established in the literature, where smoking status (never/former/current) is environmental exposure of interest and the confounders include age, gender, and study (Han, Rosenberg, & Chatterjee, 2012). Assigning a set of scores $s = (s_0, s_1, s_2)$ to genotypes (aa, Aa, AA), one can construct the trend test (Hu et al., 2017; Sasieni, 1997). Usually, $s = (0, 0.5, 1)$ is assigned for the additive model, $s = (0, 1, 1)$ is for the dominant model, and $s = (0, 0, 1)$ is for the recessive model. Because of the uncertainty of the genetic model, we can code the genotypes as $(0, \xi, 1)$, $\xi \in [0, 1]$, where ξ can be regarded as a model parameter. Therefore, we further develop to use some potential variables Z to extract model parameters from genotype coding, and then (1) and (2) are transformed into the following forms.

Model I* (Transformed gene–gene interaction model)

$$\begin{aligned}\operatorname{logit}\Pr(Y_k = 1) = \gamma_0 + \mathbf{X}_\mathrm{k}\boldsymbol{\gamma}_1 + Z_{11k}\beta_1 + Z_{12k}\xi_1\beta_1 \\ + Z_{21k}\beta_2 + Z_{22k}\xi_2\beta_2 + Z_{11k}Z_{21k}\beta_{12} \\ + Z_{11k}Z_{22k}\xi_2\beta_{12} + Z_{12k}Z_{21k}\xi_1\beta_{12} \\ + Z_{12k}Z_{22k}\xi_1\xi_2\beta_{12} = \mathbf{D}^*_{\mathbf{1k}}\boldsymbol{\Theta}\end{aligned} \tag{3}$$

Model II* (Transformed gene–environment interaction model)

$$\begin{aligned}\operatorname{logit}\Pr(Y_k = 1) = \gamma_0 + \mathbf{X}_\mathrm{k}\boldsymbol{\gamma}_1 + E_k\beta_1 + Z_{11k}\beta_2 \\ + Z_{12k}\xi_1\beta_2 + Z_{11k}E_k\beta_{12} \\ + Z_{12k}E_k\xi_1\,\beta_{12} = \mathbf{D}^*_{\mathbf{2k}}\,\boldsymbol{\Theta}\end{aligned} \tag{4}$$

where, for $i = 1, 2$,

$$(Z_{i1k}, Z_{i2k}) = \begin{cases}(0,0) & G_{ik} = 0, \\ (0,1) & G_{ik} = \xi_i, \\ (1,0) & G_{ik} = 1.\end{cases}$$

where ξ_1 and ξ_2 are genetic model parameters for two SNPs. Similarly, denote $\mathbf{D}^*_{\mathbf{1k}}$ and $\mathbf{D}^*_{\mathbf{2k}}$ by $(1, \mathbf{X}_\mathbf{k}, Z_{11k} + Z_{12k}\xi_1, Z_{21k} + Z_{22k}\xi_2, Z_{11k}Z_{21k} + Z_{11k}Z_{22k}\xi_2 + Z_{12k}Z_{21k}\xi_1 + Z_{12k}Z_{22k}\xi_1\xi_2)$ and $(1, \mathbf{X}_\mathbf{k}, E_k, Z_{11k} + Z_{12k}\xi_1, Z_{11k}E_k + Z_{12k}E_k\xi_1)$

In this work, we focus on testing two different types of effects: joint effect and interaction effect. The test for joint effect is an omnibus test that simultaneously evaluates the main effects and interaction effect. The corresponding null hypothesis for Model I* or Model II* is that $H_{01} : \beta_1 = \beta_2 = \beta_{12} = 0$. When the goal of the application is to discover a new genetic or environmental risk factor, and not to identify certain forms of interaction, then testing for joint effect is often an appropriate choice. For interaction effect, the test only evaluates the interaction between two variables, with the null hypothesis for Model I* or Model II* being that $H_{02} : \beta_{12} = 0$. It can be adopted to screen for the discovery of additional susceptibility loci following standard main effects scanned by a GWAS.

Let $\Theta = (\gamma_0, \gamma_1^\tau, \beta_1, \beta_2, \beta_{12})^\tau = (\boldsymbol{\eta}, \boldsymbol{\phi})^\tau$, where $\boldsymbol{\phi}$ denote the parameters of interest and treat others in $\boldsymbol{\eta}$ as the nuisance parameters. For example, when testing joint effect, $\boldsymbol{\phi} = (\beta_1, \beta_2, \beta_{12})$ and $\boldsymbol{\eta} = (\gamma_0, \gamma_1^\tau)$. We point out that in Model I* and Model II*, under the null hypothesis of $H_{0u}, u = 1, 2$, the corresponding parameter set is $\Theta_\mathrm{u} = (\boldsymbol{\eta_u}, \boldsymbol{\phi_u})^\tau$, and the model parameters $\boldsymbol{\xi} = \{\xi_1, \xi_2\}$ may disappear from the model, and thus are unestimable from the data. Therefore, the traditional tests, such as score test or likelihood-ratio test, which need to estimate all the nuisance parameters under the null hypothesis, are improper. Once we fix the values of $\boldsymbol{\xi}$, a score statistic can be formed for testing H_{0u}. Varying the values of $\boldsymbol{\xi}$ from an appropriate range, we can obtain a family of score statistics. Finally, we propose to use the maximum value of these statistics as our test statistics.

Denote $p_k = \Pr(Y_k = 1)$, for fixed value of $\boldsymbol{\xi}$, the log-likelihood function for testing H_{0u} is

$$\mathrm{l}(\boldsymbol{\Theta}_\mathrm{u}; \boldsymbol{\xi}) = \sum_{k=1}^{n} [Y_k p_k + (1 - Y_k)(1 - p_k)], \tag{5}$$

where p_k for Model I* and Model II* are expressed as $\exp(\mathbf{D}^*_{\mathbf{1k}}\Theta_\mathrm{u})/(1 + \exp(\mathbf{D}^*_{\mathbf{1k}}\Theta_\mathrm{u}))$ and $\exp(\mathbf{D}^*_{\mathbf{2k}}\Theta_\mathrm{u})/(1 + \exp(\mathbf{D}^*_{\mathbf{2k}}\Theta_\mathrm{u}))$ according to (3) and (4), respectively.

The score function can be written as

$$U(\Theta_\mathrm{u}; \xi) = \begin{bmatrix} U_{\eta_\mathbf{u}}(\Theta_\mathrm{u}; \xi) \\ U_{\phi_\mathbf{u}}(\Theta_\mathrm{u}; \xi) \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathrm{l}(\Theta_\mathrm{u}; \xi)}{\partial \eta_\mathbf{u}} \\ \frac{\partial \mathrm{l}(\Theta_\mathrm{u}; \xi)}{\partial \phi_\mathbf{u}} \end{bmatrix}. \tag{6}$$

The observed Fisher information matrix $\mathrm{I}(\Theta_\mathrm{u}; \boldsymbol{\xi})$ can be decomposed as

$$\mathrm{I}(\Theta_\mathrm{u}; \xi) = -\frac{1}{n} \cdot \frac{\partial^2 \mathrm{l}(\Theta_\mathrm{u}; \xi)}{\partial \Theta_\mathrm{u} \partial \Theta_\mathrm{u}^\tau} = \begin{bmatrix} \mathrm{I}_{\eta_\mathbf{u}\eta_\mathbf{u}}(\Theta_\mathrm{u}; \xi) & \mathrm{I}_{\eta_\mathbf{u}\phi_\mathbf{u}}(\Theta_\mathrm{u}; \xi) \\ \mathrm{I}_{\phi_\mathbf{u}\eta_\mathbf{u}}(\Theta_\mathrm{u}; \xi) & \mathrm{I}_{\phi_\mathbf{u}\phi_\mathbf{u}}(\Theta_\mathrm{u}; \xi) \end{bmatrix}. \tag{7}$$

The inverse of $\mathrm{I}(\boldsymbol{\Theta}_\mathrm{u}; \boldsymbol{\xi})$ is

$$\mathrm{I}^{-1}(\Theta_\mathrm{u}; \xi) = \begin{bmatrix} \mathrm{I}^{\eta_\mathbf{u}\eta_\mathbf{u}}(\Theta_\mathrm{u}; \xi) & \mathrm{I}^{\eta_\mathbf{u}\phi_\mathbf{u}}(\Theta_\mathrm{u}; \xi) \\ \mathrm{I}^{\phi_\mathbf{u}\eta_\mathbf{u}}(\Theta_\mathrm{u}; \xi) & \mathrm{I}^{\phi_\mathbf{u}\phi_\mathbf{u}}(\Theta_\mathrm{u}; \xi) \end{bmatrix}. \tag{8}$$

Then the proposed test statistics can be computed as

$$\mathrm{MEST} = \max_{\xi} T\left(\widehat{\Theta}_{\mathrm{u}};\xi\right) = \max_{\xi}\frac{1}{\mathrm{n}}U^{\tau}_{\boldsymbol{\phi}_{\mathbf{u}}} \qquad (9)$$

$$\left(\widehat{\Theta}_{\mathrm{u}};\xi\right)\mathrm{I}^{\boldsymbol{\phi}_{\mathbf{u}}\boldsymbol{\phi}_{\mathbf{u}}}\left(\widehat{\Theta}_{\mathrm{u}};\xi\right)U_{\boldsymbol{\phi}_{\mathbf{u}}}\left(\widehat{\Theta}_{\mathrm{u}};\xi\right). \qquad (9)$$

Denote the maximum likelihood estimation (MLE) as $\widehat{\Theta}_{\mathrm{u}} = (\boldsymbol{\eta}_{\mathbf{u}}, 0)^{\tau}$, which maximizes $\mathrm{l}(\boldsymbol{\Theta}_{\mathrm{u}};\boldsymbol{\xi})$ under the null hypothesis $\mathrm{H}_{0\mathrm{u}}$. In practice, the estimation is efficiently implemented by the R function 'glm'. Specifically, the corresponding MEST for Model I* and Model II* are $\max_{0\le\xi_1\le1, 0\le\xi_2\le1} T(\hat{\Theta}_{\mathrm{u}};\xi_1,\xi_2)$ and $\max_{0\le\xi_1\le1} T(\hat{\Theta}_{\mathrm{u}};\xi_1)$, respectively.

2.2 | Sequential algorithm for optimization

How to vary the values of $\boldsymbol{\xi}$? For Model II*, an intuitive method is to scatter many points equidistantly in the interval $[0, 1]$, which results in massive computations under Model I*. To reduce the computational cost, we follow Fang and Wang's (1993) suggestion that generating points by number-theoretic net (NT-nets) and using a sequential algorithm for optimization with NT-nets (abbreviated by SNTO) to compute our test statistics.

Number-theoretic method (NTM) is a special method combining number theory and numerical analysis, whose critical process is to find a uniformly scattered set of points instead of random number in Monte Carlo method. It has been intensively studied with a wide range of applications in statistics, especially for optimization problems. Different from other commonly used procedures, such as the Newton–Gauss method, the simplex method and the conjugate direction methods, fewer dependent conditions for the objective function and initial values are required for the NTM to achieve global optimum in application. Note that the SNTO procedure is an extension to the NTM, which not only inherits the merits of the NTM but also requires smaller sample size to achieve the same accuracy. The main reason to choose SNTO is due to that it provides an efficient and simple algorithm for our problem setting of continuous objective on rectangle region with a theoretical guarantee of global optimum.

Let D be a rectangle $[\mathbf{a}, \mathbf{b}]$, where $\mathbf{a} = (a_1, a_2), \mathbf{b} = (b_1, b_2)$, and $\boldsymbol{\xi} = (\xi_1, \xi_2)$. The procedure is illustrated in Algorithm 1. Since the null distribution of the MEST is intensively complicated to drive, so we consider a permutation or bootstrap procedure to calculate its p-value.

Algorithm 1: A Sequential Algorithm for Optimization with NT-nets (SNTO)

```
**Input** : The binary outcome $Y$, the SNP variable $G_1, G_2$, additional covariates $\mathbf{X}$, and the environment variable $E$;
**Output**: The proposed test statistics, MEST;
1: **Initialization**. Set iteration number $t = 0$, $D^{(0)} = D$, $\mathbf{a}^{(0)} = (a_1^{(0)}, a_2^{(0)}) = (0, 0)$, and $\mathbf{b}^{(0)} = (b_1^{(0)}, b_2^{(0)}) = (1, 1)$;
2: **Generate an NT-net**. Use a number-theoretic method (Fang & Wang, 1993) to generate $n_t$ points $P^{(t)}$ uniformly scattered on $D^{(t)} = [\mathbf{a}^{(t)}, \mathbf{b}^{(t)}]$;
3: **Compute a new approximation**. Find $\boldsymbol{\xi}^{(t)} \in P^{(t)} \cup \{\boldsymbol{\xi}^{(t-1)}\}$ and $M^{(\mathrm{t})}$ such that
   $M^{(t)} = \mathrm{MEST}\left(\boldsymbol{\xi}^{(t)}\right) \ge \mathrm{MEST}(\boldsymbol{\psi}), \forall \boldsymbol{\psi} \in P^{(t)} \cup \left\{\boldsymbol{\xi}^{(t-1)}\right\}$,
   where $\boldsymbol{\xi}^{(-1)}$ is the empty set;
4: **Termination criterion**. Let $\mathbf{c}^{(t)} = (c_1^{(t)}, c_2^{(t)}) = (\mathbf{b}^{(t)} - \mathbf{a}^{(t)})/2$. If $\max\{\mathbf{c}^{(t)}\} < \delta$, a pre-assigned small number, then $\boldsymbol{\xi}^{(\mathrm{t})}, M^{(\mathrm{t})}$ are acceptable; terminate algorithm. Otherwise, proceed to the next step;
5: **Contract domain**. Form new domain $D^{(t+1)} = [\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}]$ as follows, for $i = 1, 2$:
   $a_i^{(t+1)} = \max\left(\xi_i^{(t)} - \omega c_i^{(t)}, a_i^{(t)}\right)$
   and
   $b_i^{(t+1)} = \min\left(\xi_i^{(t)} + \omega c_i^{(t)}, b_i^{(t)}\right)$
   where $\omega$ is a pre-defined contraction ratio. Set $t = t + 1$. Go to Step 2;
6: **return**MEST($\boldsymbol{\xi}^{(\mathrm{t})}$);
```

2.3 | Simulation design

We conduct simulation studies to compare the performance of the proposed test with some existing procedures, such as the standard score test under the additive model (denote it by AST), the constrained maximum-likelihood test developed by Chatterjee and Carroll (2005) (denote it by CML), and the empirical-Bayes shrinkage estimator method (denote if by EBM). For Model I*, to test H_{01}, the parameter values are set as $\gamma_0 = -2.94$, $\gamma_1 = (0.8, 1)^{\tau}$, $\beta_1 = \log(1.2)$, $\beta_2 = \log(1.2)$, $\beta_{12} = \log(1.4)$, and the corresponding values for H_{02} are $\gamma_0 = -2.94$, $\gamma_1 = (0.8, 1)^{\tau}$, $\beta_1 = \log(1.6)$, $\beta_2 = \log(1.6)$, $\beta_{12} = \log(1.8)$.
For Model II*, to test H_{01}, the parameter values are set as $\gamma_0 = -2.94$, $\gamma_1 = (0.8, 1)^{\tau}$, $\beta_1 = \log(1.2)$, $\beta_2 = \log(1.2)$, $\beta_{12} = \log(1.4)$, and the corresponding values for H_{02} are $\gamma_0 = -2.94$, $\gamma_1 = (0.8, 1)^{\tau}$, $\beta_1 = \log(1.4)$, $\beta_2 = \log(1.4)$, $\beta_{12} = \log(1.6)$.

TABLE 1 The empirical type I error rates of MEST, AST, CML, and EBM for Model I*, based on 1400 cases and 1400 controls for joint effect, and 2500 cases and 2500 controls for interaction effect

	Joint effect					Interaction effect				
MAF	0.05	0.15	0.25	0.35	0.45	0.05	0.15	0.25	0.35	0.45
MEST	0.048	0.035	0.047	0.042	0.054	0.041	0.044	0.045	0.043	0.036
AST	0.044	0.040	0.048	0.041	0.048	0.034	0.051	0.063	0.052	0.057
CML	0.043	0.036	0.061	0.043	0.049	0.054	0.057	0.060	0.067	0.047
EBM	0.035	0.027	0.043	0.032	0.030	0.031	0.031	0.027	0.043	0.026

TABLE 2 The empirical type I error rates of MEST, AST, CML, and EBM for Model II*, based on 800 cases and 800 controls for joint effect, and 2400 cases and 2400 controls for interaction effect

	Joint effect					Interaction effect				
MAF	0.05	0.15	0.25	0.35	0.45	0.05	0.15	0.25	0.35	0.45
MEST	0.041	0.049	0.048	0.057	0.054	0.047	0.041	0.037	0.059	0.048
AST	0.051	0.055	0.041	0.052	0.049	0.058	0.053	0.039	0.067	0.049
CML	0.044	0.052	0.042	0.050	0.049	0.053	0.048	0.046	0.061	0.058
EBM	0.038	0.042	0.028	0.040	0.042	0.036	0.035	0.027	0.040	0.037

Note: The environment samples are generated from normal distribution with mean of 0 and variance of 0.5.

TABLE 3 The empirical type I error rates of MEST, AST, CML, and EBM for Model II*, based on 800 cases and 800 controls for joint effect, and 2400 cases and 2400 controls for interaction effect

	Joint effect					Interaction effect				
MAF	0.05	0.15	0.25	0.35	0.45	0.05	0.15	0.25	0.35	0.45
MEST	0.046	0.056	0.043	0.064	0.058	0.063	0.057	0.052	0.050	0.047
AST	0.055	0.055	0.045	0.055	0.048	0.059	0.067	0.043	0.051	0.046
CML	0.054	0.053	0.047	0.046	0.058	0.051	0.064	0.057	0.048	0.044
EBM	0.031	0.038	0.032	0.028	0.040	0.047	0.047	0.045	0.040	0.037

Note: The environment samples are generated from Bernoulli distribution with parameter of 0.3.

To make easy comparison on power, for Model I*, 1400 cases and 1400 controls are independently sampled to test for joint effect, 2500 cases and 2500 controls are for interaction effect. For Model II*, 800 cases and 800 controls are generated to test for joint effect, 2400 cases and 2400 controls are for interaction effect. We consider the minor allele frequencies (MAF) as {0.05,0.15,0.25,0.35,0.45}. The significance level α is set as 0.05.

The covariates $\mathbf{X}_k = (X_{1k}, X_{2k})$ are generated from a bivariate normal distribution $N(\boldsymbol{\mu}, \mathbf{I}_2)$, where $\boldsymbol{\mu} = (1.2, 1.4)^\tau$, $\mathbf{I}_2$ is the identity matrix. G_{1k}, G_{2k} are sampled from trinomial distributions under the HWE, that is, the corresponding frequencies of coded genotypes $(0, 1, 2)$ are $(1-p)^2, 2p(1-p), p^2$. We consider to separately generate two types of environmental factors E_k, that the categorical samples are from a Bernoulli distribution

TABLE 4 The empirical type I error rates of MEST, AST, CML, and EBM for Model II* under gene–environment dependence, based on 800 cases and 800 controls for joint effect, and 2400 cases and 2400 controls for interaction effect

	Joint effect					Interaction effect				
MAF	0.05	0.15	0.25	0.35	0.45	0.05	0.15	0.25	0.35	0.45
MEST	0.045	0.043	0.042	0.062	0.053	0.043	0.029	0.035	0.034	0.072
AST	0.049	0.052	0.046	0.058	0.048	0.050	0.049	0.046	0.045	0.053
CML	0.259	0.398	0.490	0.534	0.562	0.807	0.960	0.983	0.985	0.980
EBM	0.085	0.095	0.098	0.109	0.086	0.114	0.100	0.072	0.064	0.077

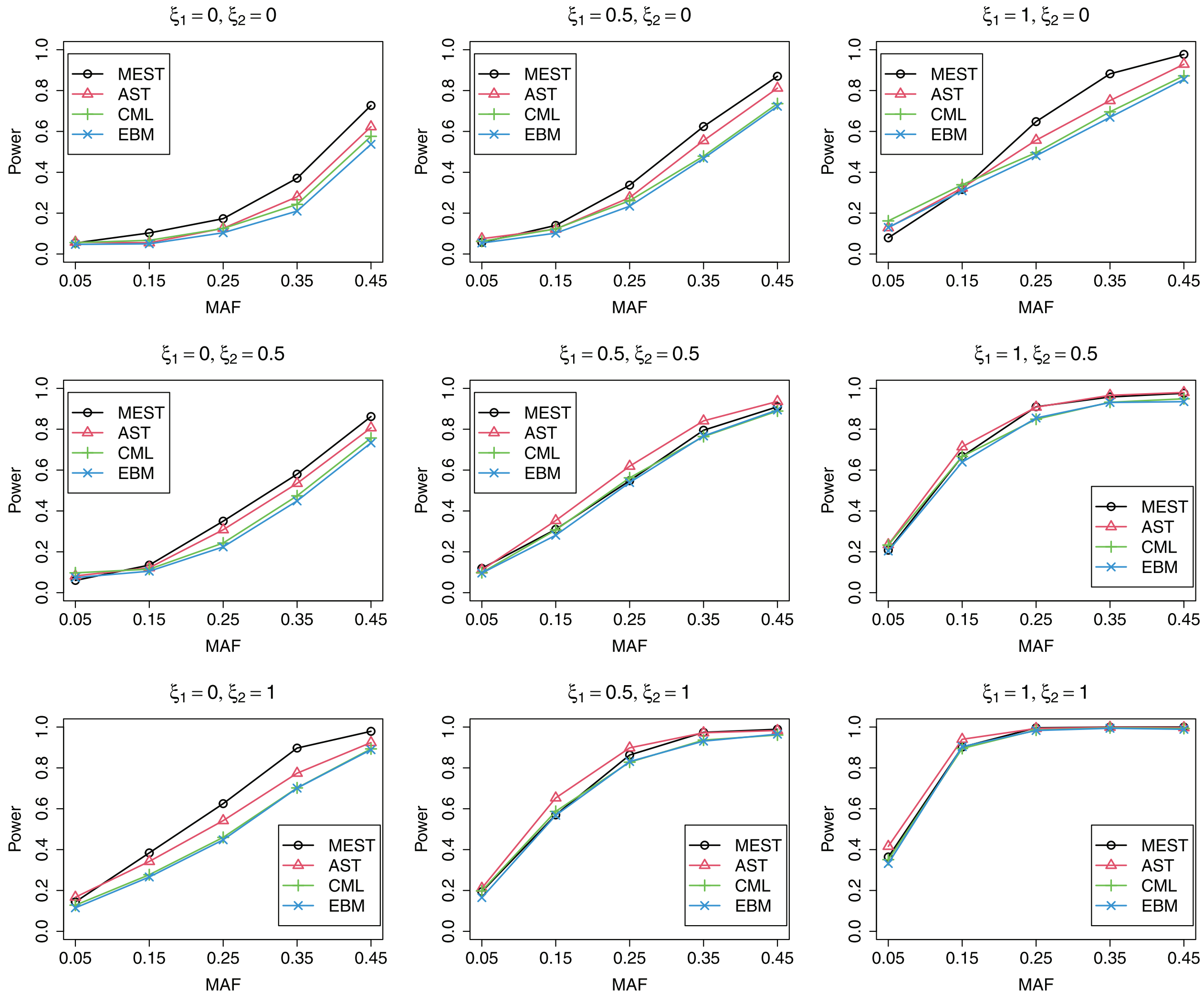

FIGURE 1 The empirical power of MEST, AST, CML, and EBM for Model I* to test joint effect under various genetic models based on 1400 cases and 1400 controls

Bernoulli(0.3) followed by Han, Rosenberg, and Chatterjee (2012), and the continuous ones are from a normal distribution $N(0, 0.5)$. Then cases and controls are generated from the logistic regression model under the prespecific genetic model, and we consider the REC ($\xi = 0$), the ADD ($\xi = 0.5$), and the DOM ($\xi = 1$) as the true models.

It is worth noting that the proposed test is constructed by standard logistic regression settings of case–control data, previous evidence has illustrated the robustness for the gene–gene or gene–environment independence assumption (Mukherjee & Chatterjee, 2008; Mukherjee et al., 2008). To further verify this advantage of our method, we also conduct additional setting with gene–environment dependence. Specifically, we first generate samples from a bivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ with $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where set $\rho = 0.1$. Then the one is treated as genotype data that are discretized to {0,1,2} by the corresponding quantiles based on HWE for a given MAF, and another one is directly taken as continuous environment data.

When using the SNTO algorithm, we take $n_0 > n_1 = \cdots$, where $n_0 = 1000$, $n_1 = n_2 = \cdots = 120$ for testing joint effect and $n_0 = 300$, $n_1 = n_2 = \cdots = 100$ for testing interaction effect. Set $\delta = 10^{-2}$, and define $\omega = \kappa^{t+1}$ at tth iteration followed by Niederreiter and Peart (1986), where $\kappa = 0.5$.

3 | RESULTS

3.1 | The empirical type I error rates

Tables 1–4 illustrate the empirical type I error rates of MEST, AST, CML, and EBM for testing both effects of Model I* and Model II*. It can be seen that all the values are close to the nominal significance level of 0.05, which means all the considered procedures can control the type I error rates properly. For example, when MAF $= 0.25$ for joint effect in Table 1, the empirical type I error rates of MEST, AST, CML, and EBM are 0.047, 0.048, 0.061, and 0.043, respectively. When MAF $= 0.35$ for interaction

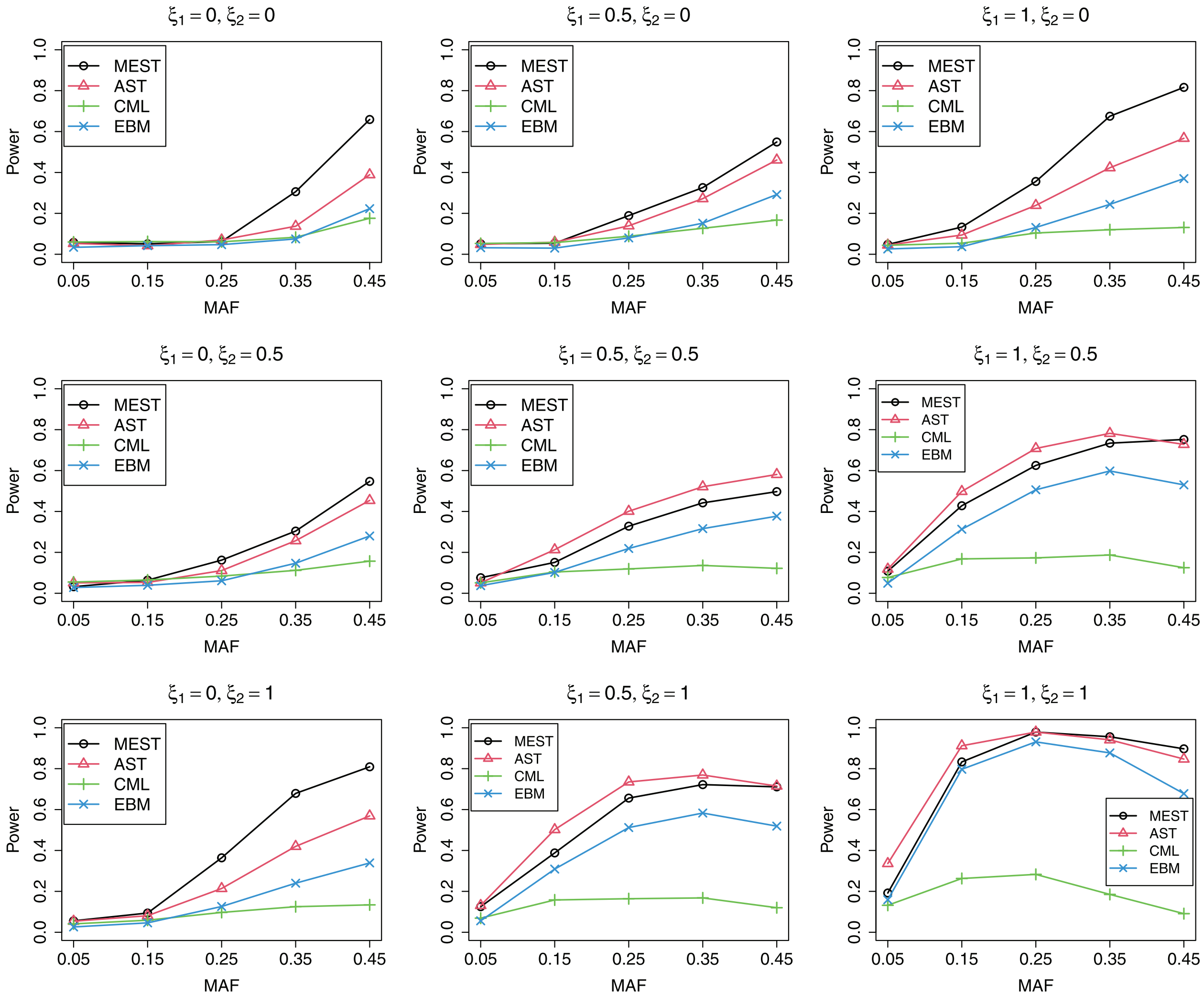

FIGURE 2 The empirical power of MEST, AST, CML, and EBM for Model I* to test interaction effect under various genetic models based on 2500 cases and 2500 controls

effect in Table 3, the corresponding values are 0.050, 0.051, 0.048, and 0.040, respectively. In addition, the AST and our method can still well control the type I error rates under gene–environment dependence setting, whereas the corresponding rates of the CML and the EBM fail even if the correlation is small. Especially from the results of the CML, it can be seen that there are serious deviation from the nominal significance level, which is not surprising because the CML is constructed incorporating the gene–environment independence assumption.

3.2 | The empirical power

Figures 1 and 2 show the empirical power of MEST, AST, CML, and EBM for testing two effects of Model I* under various combinations of true genetic models. For Model II*, the corresponding results are shown in Tables 5–7.

It can be obviously seen that, to test for gene–gene interaction, the MEST is more powerful than CML and EBM for the combination of the recessive model and all the three models, but slightly less effective than the AST for other cases. For example, when MAF = 0.45, $\xi_1 = 0$, $\xi_2 = 0$ for testing joint effect in Figure 1, the empirical power of MEST, AST, CML, and EBM are 0.727, 0.623, 0.576, and 0.537, and the corresponding values are 0.795, 0.841, 0.764, and 0.767 for MAF = 0.35, $\xi_1 = 0.5$, $\xi_2 = 0.5$, respectively. For the gene–environment interaction, similar results can be found. The MEST can obtain larger power than other three tests under the recessive model, but slightly lower than the AST under the additive and dominant models. For example, when MAF = 0.45, $\xi_1 = 0$ for testing interaction effect in Table 5, the empirical power of MEST, AST, CML, and EBM are 0.724, 0.589, 0.297, and 0.415, and the corresponding values for testing joint effect when MAF = 0.35, $\xi_1 = 0.5$ are 0.685, 0.721, 0.673, and 0.669, respectively. This is sensible because the AST is optimal for the

TABLE 5 The empirical power of MEST, AST, CML, and EBM for Model II* under three genetic models based on 800 cases and 800 controls for joint effect, and 2400 cases and 2400 controls for interaction effect

MAF	0.05	0.15	0.25	0.35	0.45	0.05	0.15	0.25	0.35	0.45
	$\xi_1 = 0$, interaction effect						$\xi_1 = 0$, joint effect			
MEST	0.065	0.120	0.296	0.575	0.724	0.233	0.234	0.322	0.453	0.608
AST	0.055	0.095	0.214	0.387	0.589	0.229	0.253	0.323	0.437	0.572
CML	0.055	0.076	0.118	0.199	0.297	0.231	0.237	0.305	0.413	0.534
EBM	0.037	0.060	0.115	0.257	0.415	0.222	0.225	0.288	0.396	0.543
	$\xi_1 = 0.5$, interaction effect					$\xi_1 = 0.5$, joint effect				
MEST	0.126	0.374	0.536	0.609	0.658	0.295	0.413	0.564	0.685	0.783
AST	0.162	0.447	0.585	0.645	0.697	0.302	0.479	0.612	0.721	0.788
CML	0.099	0.216	0.276	0.293	0.336	0.288	0.416	0.553	0.673	0.758
EBM	0.093	0.302	0.397	0.464	0.522	0.275	0.418	0.554	0.669	0.772
	$\xi_1 = 1$, interaction effect					$\xi_1 = 1$, joint effect				
MEST	0.432	0.847	0.915	0.898	0.851	0.458	0.721	0.822	0.902	0.912
AST	0.521	0.853	0.902	0.866	0.789	0.431	0.700	0.846	0.892	0.930
CML	0.246	0.449	0.490	0.434	0.353	0.397	0.629	0.797	0.865	0.908
EBM	0.327	0.684	0.779	0.713	0.620	0.393	0.663	0.814	0.872	0.924

Note: The environment samples are generated from normal distribution with mean of 0 and variance of 0.5.

TABLE 6 The empirical power of MEST, AST, CML, and EBM for Model II* under three genetic models based on 800 cases and 800 controls for joint effect, and 2400 cases and 2400 controls for interaction effect

MAF	0.05	0.15	0.25	0.35	0.45	0.05	0.15	0.25	0.35	0.45
	$\xi_1 = 0$, interaction effect					$\xi_1 = 0$, joint effect				
MEST	0.054	0.096	0.232	0.390	0.690	0.211	0.226	0.320	0.488	0.592
AST	0.049	0.075	0.160	0.311	0.549	0.214	0.240	0.313	0.444	0.597
CML	0.049	0.068	0.095	0.141	0.228	0.225	0.237	0.302	0.403	0.520
EBM	0.030	0.043	0.089	0.182	0.369	0.209	0.218	0.290	0.380	0.518
	$\xi_1 = 0.5$, interaction effect					$\xi_1 = 0.5$, joint effect				
MEST	0.113	0.344	0.463	0.582	0.556	0.295	0.501	0.637	0.645	0.787
AST	0.160	0.377	0.522	0.598	0.634	0.317	0.518	0.645	0.715	0.807
CML	0.100	0.173	0.188	0.231	0.217	0.299	0.482	0.568	0.672	0.748
EBM	0.092	0.222	0.321	0.391	0.433	0.266	0.460	0.563	0.676	0.756
	$\xi_1 = 1$, interaction effect					$\xi_1 = 1$, joint effect				
MEST	0.455	0.797	0.857	0.855	0.786	0.476	0.754	0.879	0.932	0.938
AST	0.480	0.814	0.855	0.809	0.725	0.520	0.791	0.896	0.921	0.940
CML	0.189	0.307	0.296	0.274	0.202	0.444	0.705	0.823	0.872	0.910
EBM	0.281	0.621	0.688	0.620	0.523	0.433	0.689	0.824	0.880	0.921

Note: The environment samples are generated from Bernoulli distribution with parameter of 0.3.

additive model. Therefore, the proposed MEST is a procedure with greater efficiency robustness than others.

3.3 | Application to rheumatoid arthritis data

Rheumatoid arthritis (RA) is a common chronic systemic disease that is attributed to a complex interplay of genetic and environmental factors (MacGregor et al., 2000). Previous GWAS have identified many risk loci located on chromosome 6 and 9 (Chanda et al., 2009; Chen et al., 2009; Plenge et al., 2007). Our primary goal is to further investigate the potential RA associated gene–gene interaction effects, and hence we carry out our analysis under the reported candidate genome regions.

TABLE 7 The empirical power of MEST, AST, CML, and EBM for Model II* under three genetic models with gene–environment dependence based on 800 cases and 800 controls for joint effect, and 2400 cases and 2400 controls for interaction effect

MAF	0.05	0.15	0.25	0.35	0.45	0.05	0.15	0.25	0.35	0.45
	$\xi_1 = 0$, interaction effect					$\xi_1 = 0$, joint effect				
MEST	0.059	0.346	0.772	0.993	0.999	0.763	0.785	0.893	0.966	0.998
AST	0.050	0.200	0.552	0.907	0.992	0.777	0.814	0.893	0.964	0.996
CML	0.868	0.993	1.000	1.000	1.000	0.876	0.958	0.987	0.998	1.000
EBM	0.156	0.337	0.670	0.956	0.997	0.725	0.751	0.862	0.960	0.993
	$\xi_1 = 0.5$, interaction effect					$\xi_1 = 0.5$, joint effect				
MEST	0.496	0.901	0.974	0.989	0.994	0.839	0.958	0.993	0.998	1.000
AST	0.541	0.919	0.976	0.992	0.995	0.872	0.970	0.996	1.000	1.000
CML	0.993	1.000	1.000	1.000	1.000	0.95	0.997	1.000	1.000	1.000
EBM	0.767	0.970	0.994	0.998	0.997	0.872	0.964	0.995	0.998	1.000
	$\xi_1 = 1$, interaction effect					$\xi_1 = 1$, joint effect				
MEST	0.949	1.000	1.000	1.000	1.000	0.946	0.997	1.000	1.000	1.000
AST	0.980	1.000	0.999	1.000	0.997	0.962	0.997	1.000	1.000	1.000
CML	1.000	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	1.000
EBM	0.998	1.000	0.999	1.000	0.997	0.973	0.997	1.000	1.000	1.000

We apply our method to the case–control data collected from the Genetic Analysis Workshop 16 (GAW16). After quality control, the dataset includes 868 cases and 1194 controls (Amos et al., 2009). In this section, we just analysis the regions of HLA-DRB1 and 6q21.3 in chromosomes 6, which contains six SNPs and 45 SNPs in our datasets. To account for the possible confounding of population stratification, we match 12898 structure inference SNPs with low local background LD, and then 12749 SNPs are selected to construct four principal coordinates by the multidimensional scaling method (Li & Yu, 2008). We put all principal coordinates into our model as covariates. The nominal significance level is set to be 0.05. Note that 1000 permutations are set to obtain the significance of the proposed test.

Some pairs of SNPs interactions are successfully discovered using our method. Here we only demonstrate a few examples to verify the utility of the MEST. For the interaction effects test, a strong interaction between the SNP rs660895 in region HLA-DRB1 and the SNP rs2246986 in region 6q21.33 is identified by the MEST with the p-value of 0.015, and the corresponding p-value results of the UML, CML, EBM are 0.039, 0.775, and 0.128. For the joint effects test, a significant result between rs3749946 and rs3132454 in region 6q21.33 is observed using our method with p-value of 0.028, whereas the alternatives miss this finding with p-value larger than 0.05. Therefore, the proposed method is easier to detect the new signals in the interested region.

4 | DISCUSSION

In large-scale genetic association studies, evaluating gene–gene or gene–environment interaction associated with human diseases is an important issue. The existing methods, such as case–control method, case-only method, the empirical-Bayes method, have some limitations including that low power with small sample size, requiring the independence assumption of gene–environment and ignoring potential confounders in the model setting. On the other hand, since the genetic model of the true disease loci is usually uncertain in practice, analysis based on additive model may not always be a proper option. For example, rs420259 at chromosome 16p12 has been investigated to be significantly associated with bipolar disorder under recessive model (Wellcome Trust Case Control Consortium, 2007). Robust procedures in single marker analysis accommodating genetic model uncertainty have been intensively investigated in the literature (Freidlin et al., 2002; Hu et al., 2017; Li et al., 2008; Zheng & Ng, 2008). In this work, we contribute to develop a procedure called MEST considering this type of robustness properly in analysis of gene–gene or gene–environment interaction to obtain additional benefits of power. Specifically, we use some potential variables to separate the unknown model parameters and construct a novel score test incorporating the model information to test genetic association with interaction effects in a case–control design. Extensive simulations and real data analysis show that the proposed test achieves better performance than comparative approaches.

To obtain the p-value of MEST, a bootstrap or permutation procedure needs to be conducted. However, it is a complicated task to solve a constrained optimization and compute the values of test statistics, that standard method may be computationally intensive. Number-theoretic method (NTM) is an effective alternative suiting for this problem. To further reduce the computational burden, we consider using the extended sequential optimization algorithm based on NTM (Fang & Wang, 1993). Specifically, the algorithm shrinks the searching space of the parameters via finding a uniformly scattered set of points in each iteration. This algorithm is very useful for multidimensional optimization problem to search the global solution with theoretical guarantee. On the other hand, although we assume the HWE holds in the source population in the simulation studies, the MEST is actually free of the assumption of HWE. Obviously, the MEST also does not rely on the gene–environment independence assumption.

A natural extension of our method in the future is to consider multiple genetic variants or environment exposures and their potential interactions simultaneously, which is usually under high- or ultrahigh-dimensional settings and the model screening approach can be properly adopted (Li et al., 2014). As the simulation results shown, the power of MEST is lower than the AST when the true genetic model is additive model. Despite it is reasonable to see, further exploiting the gene–environment independence assumption (Chatterjee & Carroll, 2005) is still a promising direction to enhance the efficiency.

ACKNOWLEDGEMENTS

The authors thank the editor and two anonymous reviewers for their constructive suggestions and comments that have greatly improve this article. This study was supported by Capital University of Economics and Business Newly Recruited Young Teachers Scientific Research Start-up Foundation Project (Grant No. XRZ2022066).

AUTHOR CONTRIBUTIONS

Xiaonan Hu and Zhen Meng conducted the project, developed the method and algorithm, performed the simulation study, analyzed the RA data and wrote the manuscript. All the authors discussed and reviewed the final manuscript.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data for genome-wide association analysis of rheumatoid arthritis that support this study of gene–gene interaction analysis was approved by Genetic Analysis Workshop 16.

REFERENCES

Albert, P. S., Duminda, R., Joseph, T., & Sholom, W. (2001). Limitations of the case–only design for identifying gene–environment interactions. *American Journal of Epidemiology*, *154*, 687–693. https://doi.org/10.1093/aje/154.8.687

Amos, C. I., Chen, W. V., Seldin, M. F., Remmers, E. F., Taylor, K. E., Criswell, L. A., Lee, A. T., Plenge, R. M., Kastner, D. L., & Gregersen, P. K. (2009). Data for genetic analysis workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proceedings*, *3*, S2. https://doi.org/10.1186/1753-6561-3-S7-S2

Carlborg, O., & Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies. *Nature Reviews Genetics*, *5*, 618–625. https://doi.org/10.1038/nrg1407

Chanda, P., Zhang, A., Sucheston, L., & Ramanathan, M. (2009). A two-stage search strategy for detecting multiple loci associated with rheumatoid arthritis. *BMC Proceedings*, *3*, S72. https://doi.org/10.1186/1753-6561-3-S7-S72

Chatterjee, N., & Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case–control studies. *Biometrika*, *92*, 399–418. https://doi.org/10.1093/biomet/92.2.399

Chen, L., Zhong, M., Chen, W. V., Amos, C. I., & Fan, R. (2009). A genome-wide association scan for rheumatoid arthritis data by Hotelling's T2 tests. *BMC Proceedings*, *3*, S6. https://doi.org/10.1186/1753-6561-3-S7-S6

Fang, K. T., & Wang, Y. (1993). *Number–Theoretic methods in statistics*. CRC Press.

Freidlin, B., Zheng, G., Li, Z., & Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity*, *53*, 146–152. https://doi.org/10.1159/000064976

Han, S. S., Rosenberg, P. S., & Chatterjee, N. (2012). Testing for gene-environment and gene-gene interactions under monotonicity constraints. *Journal of the American Statistical Association*, *107*, 1441–1452. https://doi.org/10.1080/01621459.2012.726892

Han, S. S., Rosenberg, P. S., Garcia-Closas, M., Figueroa, J. D., Silverman, D., Chanock, S. J., Rothman, N., & Chatterjee, N. (2012). Likelihood ratio test for detecting gene (G)-environment (E) interactions under an additive risk model exploiting G-E independence for case-control data. *American Journal of Epidemiology*, *176*, 1060–1067. https://doi.org/10.1093/aje/kws166

Hu, X., Duan, X., Pan, D., Zhang, S., & Li, Q. (2017). A model-embedded trend test with incorporating Hardy-Weinberg equilibrium information. *Journal of Systems Science and Complexity*, *30*, 101–110. https://doi.org/10.1007/s11424-017-6187-4

He, Z., Zhang, M., Lee, S., Smith, J. A., Kardia, S. L., Roux, V. D., & Mukherjee, B. (2017). Set-based tests for the gene-environment interaction in longitudinal studies. *Journal of the American Statistical Association*, *112*, 966–978. https://doi.org/10.1080/01621459.2016.1252266

Hu, J., Zhang, W., Li, X., Pan, D., & Li, Q. (2019). Efficient estimation of disease odds ratios for the follow-up genetic association studies. *Statistical Methods in Medical Research*, *28*, 1927–1941. https://doi.org/10.1177/0962280217741771

Johnsen, P. V., Riemer-Sørensen, S., DeWan, A. T., Cahill, M. E., & Langaas, M. (2021). A new method for exploring gene-gene and gene-environment interactions in GWAS with tree ensemble methods and SHAP values. *BMC Bioinformatics [Electronic Resource]*, *22*, 1–29. https://doi.org/10.1186/s12859-021-04041-7

Li, Q., & Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology*, *32*, 215–226. https://doi.org/10.1002/gepi.20296

Li, Q., Zheng, G., Li, Z., & Yu, K. (2008). Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. *Annals of Human Genetics*, *72*, 397–406. https://doi.org/10.1111/j.1469-1809.2008.00437.x

Li, D., & Conti, D. V. (2009). Detecting gene-environment interactions using a combined case-only and case-control approach. *American Journal of Epidemiology*, *169*, 497–504. https://doi.org/10.1093/aje/kwn339

Li, J., Zhong, W., Li, R., & Wu, R. (2014). A fast algorithm for detecting gene-gene interactions in genome-wide association studies. *The Annals of Applied Statistics*, *8*, 2292. https://doi.org/10.1214/14-AOAS771

MacGregor, A. J., Snieder, H., Rigby, A. S., Koskenvuo, M., Kaprio, J., Aho, K., & Silman, A. J. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *43*, 30–37. https://doi.org/10.1002/1529-0131(200001)43:1⟨30::AID-ANR5⟩3.0.CO;2-B

Marchini, J., Donnelly, P., & Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, *37*, 413–417. https://doi.org/10.1038/ng1537

Mukherjee, B., & Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, *64*, 685–694. https://doi.org/10.1111/j.1541-0420.2007.00953.x

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753. https://doi.org/10.1038/nature08494

Meisner, A., Kundu, P., & Chatterjee, N. (2019). Case-only analysis of gene-environment interactions using polygenic risk scores. *American Journal of Epidemiology*, *188*, 2013–2020. https://doi.org/10.1093/aje/kwz175

Mukherjee, B., Ahn, J., Gruber, S. B., Rennert, G., Moreno, V., & Chatterjee, N. (2008). Tests for gene-environment interaction from case-control data: A novel study of type I error, power and designs. *Genetic Epidemiology*, *32*, 615–626. https://doi.org/10.1002/gepi.20337

Mukherjee, B., Ahn, J., Gruber, S. B., Ghosh, M., & Chatterjee, N. (2010). Case-control studies of gene-environment interaction: Bayesian design and analysis. *Biometrics*, *66*, 934–948. https://doi.org/10.1111/j.1541-0420.2009.01357.x

Niederreiter, H., & Peart, P. (1986). Localization of search in quasi-Monte Carlo methods for global optimization. *SIAM Journal on Scientific and Statistical Computing*, *7*, 660–664. https://doi.org/10.1137/0907044

Piegorsch, W. W., Weinberg, C. R., & Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, *13*, 153–162. https://doi.org/10.1002/sim.4780130206

Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L. R., Li, W., Tan, A. K., Bonnard, C., Ong, R. T., Thalamuthu, A., Pettersson, S., Liu, C., Tian, C., Chen, W. V., … Gregersen, P. K. (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis-a genomewide study. *New England Journal of Medicine*, *357*, 1199–1209. https://doi.org/10.1056/NEJMoa073491

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics*, *53*, 1253–1261. https://doi.org/10.2307/2533494

Umbach, D. M., & Weinberg, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, *16*, 1731–1743. https://doi.org/10.1002/(SICI)1097-0258(19970815)16:15⟨1731::AID-SIM595⟩3.0.CO;2-S

Wang, X., Lim, E., Liu, C. T., Sung, Y. J., Rao, D. C., Morrison, A. C., Manning, A. K., & Chen, H. (2020). Efficient gene-environment interaction tests for large biobank-scale sequencing studies. *Genetic Epidemiology*, *44*, 908–923. https://doi.org/10.1002/gepi.22351

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661. https://doi.org/10.1038/nature05911

Zheng, G., & Ng, H. K. T. (2008). Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics (Oxford, England)*, *9*, 391–399. https://doi.org/10.1093/biostatistics/kxm039

Zheng, G., Zhang, W., Xu, J., Yuan, A., Li, Q., & Gastwirth, J. L. (2016). Genetic risks and genetic model specification. *Journal of Theoretical Biology*, *403*, 68–74. https://doi.org/10.1016/j.jtbi.2016.05.016

How to cite this article: Hu, X., & Meng, Z. (2022). Using potential variable to study gene–gene and gene–environment interaction effects with genetic model uncertainty. *Annals of Human Genetics*, 1–11. https://doi.org/10.1111/ahg.12470