J Syst Sci Complex (2017) 30: 101–110

A Model-Embedded Trend Test with Incorporating Hardy-Weinberg Equilibrium Information*

HU Xiaonan · DUAN Xiaogang · PAN Dongdong · ZHANG Sanguo · LI Qizhai

DOI: 10.1007/s11424-017-6187-4
Received: 25 August 2016 / Revised: 20 September 2016

Abstract The genetic models are greatly important in the analysis of genetic epidemiologic studies and many of the studies are conducted using the trend test under the additive model. However, for many complex diseases and traits, the underlying genetic model for a genetic locus is usually uncertain. So a robust test free of genetic model is appropriate. In this paper, the authors propose a model-embedded trend test by incorporating Hardy-Weinberg equilibrium information and obtain the explicit formula to calculate its statistical significance. Extensive simulation studies show the proposed test is more robust than the existing procedures. Finally, a real application is further analyzed to show the performance of the proposed test.

Keywords Genetic model, hardy-weinberg equilibrium, power, robust, trend test.

HU Xiaonan
School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, *China; Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing* 100190, *China.* Email: huxiaonan14@mails.ucas.ac.cn.
DUAN Xiaogang
School of Statistics, Beijing Normal University, Beijing 100875, *China.* Email: xgduan@bnu.edu.cn.
PAN Dongdong (Corresponding author)
Department of Statistics, Yunnan University, Kunming 650091, *China.* Email: ddpan@ynu.edu.cn.
ZHANG Sanguo
School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, *China; Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing* 100190, *China.* Email: sgzhang@ucas.ac.cn.
LI Qizhai
Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, *China.* Email: liqz@amss.ac.cn.

*This research was partial supported by Special National Key Research and Development Plan under Grant No. 2016YFD0400206, the Breakthrough Project of Strategic Priority Program of Chinese Academy of Sciences under Grant No. XDB13040600 and Youth Innovation Promotion Association of Chinese Academy of Sciences, the National Science Foundation of China under Grant Nos. 11371353, 11661080, 61134013, and Special Fund of the University of Chinese Academy of Sciences for Scientific Research Cooperation.
◇ *This paper was recommended for publication by Editor XU Jin.*

1 Introduction

In genetic epidemiologic studies, one of the goals is to identify the disease-associated susceptibilities, mostly single nucleotide polymorphism (SNP). Single-marker analysis is commonly adopted in many genetic association studies such as genome-wide association study[1, 2] or multiple-SNPs association studies[3, 4], where the trend test derived under the additive genetic model (TTA) is employed. The genetic model is a functional relationship of a risk measure given genotypes, which describes the impact on phenotypes by the genotypes[5]. The commonly used genetic models are the recessive, the additive, and the dominate models. Using trend test based on the additive model may lose power if the genetic mode of inherence is not additive. In practice, the models of the true disease loci are usually uncertain for many complex diseases.

To handle the situation in which the genetic model is unknown, a robust test free of the model assumption is preferable. There are some robust procedures used in genetic studies, such as the allele-based test[6], which is less powerful than the genotype-based test, the CHI2, a 2 degrees of freedom Chi-Squared test[7], the MAX3, the maximum value of trend tests derived under three genetic models[8], the MIN2, the minimum p-value of the TTA and the CHI2. Beside these tests, there are some other procedures such as the restricted likelihood ratio test[9], which has been shown to be less powerful than the MAX3 if the genetic model is restricted to be from recessive to dominant.

The Hardy-Weinberg equilibrium (HWE) law plays an important role in population genetics. It needs to check whether the observed genotypes satisfy the HWE law in control population routinely before conducting an association analysis, since the deviations from HWE can indicate many problems such as population stratification, genotyping error, and so on[10–12]. Consider a biallelic SNP locus with two alleles A and a. Denote the allele frequency of A by p. Under the HWE principal, the genotype frequencies of AA, Aa and aa are p^2, $2p(1-p)$ and $(1-p)^2$, respectively.

Recently, Zheng and Ng[13] proposed a two-phase procedure (TPP) to do single-marker analysis, where they suggested to select the genetic model in the first phase and then the trend test derived under the chosen model was used in the second phase. They used the Hardy-Weinberg equilibrium test derived in both cases and controls for the genetic model selection. As pointed out previously, the Hardy-Weinberg equilibrium holds in controls naturally. So, using control samples to check the derivation from the Hardy-Weinberg equilibrium might bring some notices, thus the selection rates were reduced. On the other hand, we would like to point out that Zheng and Ng[13] used the adjusted significance level to judge whether there existed the associations for a nominal size. The adjusted significance level was a function of the minor allele frequency (MAF), the disease prevalence and the relative risks. So it changed when other parameters changed and then one needed to calculate 1×10^6 adjusted significance levels for a genome-wide association study with 1×10^6 SNPs genotyped. To overcome both disadvantages, we first use the Hardy-Weinberg equilibrium test derived in cases only for model selection and then combine it with the corresponding trend test. We call this procedure the model-embedded trend test (MET). We give the explicit expression to calculate the significance level of the MET. The extensive computer simulations and a real data application are conducted to compare the performances between the MET and some existing procedures.

2 Methods

2.1 The Model-embedded Trend Test

Consider a disease locus with 2 alleles A and a, without loss of generality, let A be the risk allele. Then the corresponding genotypes are denoted by aa, Aa, and AA. Denote the minor allele frequency (the frequency of allele A) by p, that is, $p = \Pr(A)$, and the disease prevalence by k. Using the Bayes formula and that the HWE holds in the source population, the genotype frequencies of (aa, Aa, AA) in cases are (p_0, p_1, p_2) and those in controls are (q_0, q_1, q_2), where

$$p_0 = \frac{(1-p)^2 f_0}{k}, \quad p_1 = \frac{2p(1-p)\lambda_1 f_0}{k}, \quad p_2 = \frac{p^2\lambda_2 f_0}{k}, \tag{1}$$

and

$$q_0 = \frac{(1-p)^2(1-f_0)}{1-k}, \quad q_1 = \frac{2p(1-p)(1-\lambda_1 f_0)}{1-k}, \quad q_2 = \frac{p^2(1-\lambda_2 f_0)}{1-k}, \tag{2}$$

with $f_0 = \frac{k}{(1-p)^2+2p(1-p)\lambda_1+p^2\lambda_2}$ and (λ_1, λ_2) being the genotype relative risks, $\lambda_1 \geq 1$ and $\lambda_2 > 1$. The recessive, additive, and dominant models are corresponding to $(\lambda_1 = 1, \lambda_2 > 1)$, $(\lambda_1 > 1, \lambda_2 = 2\lambda_1 - 1)$, and $(\lambda_1 = \lambda_2 > 1)$, respectively. The null hypothesis that the locus is not associated with the disease is $p_i = q_i, i = 0, 1, 2$ or equivalently $\lambda_1 = \lambda_2 = 1$.

Assume that r cases and s controls are independently sampled in a case-control study, and the observed counts for (aa, Aa, AA) in cases and controls are (r_0, r_1, r_2) and (s_0, s_1, s_2). Denote $n_i = r_i + s_i, i = 0, 1, 2$ and $n = r + s$, and let $U_\theta = (\widehat{p}_2 + \theta\widehat{p}_1) - (\widehat{q}_2 + \theta\widehat{q}_1)$, where $\widehat{p}_i = r_i/r$ and $\widehat{q}_i = s_i/s$ are the estimators of p_i and q_i $(i = 0, 1, 2)$. Then the trend test[14] can be constructed by $Z_\theta = U_\theta/\{\widehat{\text{var}}_0(U_\theta)\}^{1/2}$, where

$$\widehat{\text{var}}_0\,(U_\theta) = \frac{(\widetilde{p}_2 + \theta^2\widetilde{p}_1) - (\widetilde{p}_2 + \theta\widetilde{p}_1)^2}{r} + \frac{(\widetilde{q}_2 + \theta^2\widetilde{q}_1) - (\widetilde{q}_2 + \theta\widetilde{q}_1)^2}{s},$$

and $\widetilde{p}_i = \widetilde{q}_i = n_i/n$.

For any given $\theta \in [0, 1]$, Z_θ has the asymptotical standard normal distribution under the null hypothesis. The optimal θ for the recessive, the additive, and the dominant models are 0, 0.5, and 1, respectively. So, we can use Z_0, $Z_{1/2}$, and Z_1 to detect association when the genetic models are the recessive, the additive, and the dominant, respectively. However, the true genetic model is usually uncertain in practice. The trend test under a given genetic model may lose power when the model is misspecified. So a robust test is preferable.

The deviation from HWE in cases can be used to detect the genetic model[5]. Denote

$\varDelta_C = \widehat{p}_2 - (\widehat{p}_2 + \widehat{p}_1/2)^2$, $T_C = \varDelta_C / \{\widehat{\mathrm{var}}_0(\varDelta_C)\}^{1/2}$, where

$$\begin{aligned}
&\widehat{\mathrm{var}}_0(\varDelta_C)\\
=&\frac{1}{r}(\widetilde{p}_2 - 5\widetilde{p_2^2} + 8\widetilde{p_2^3} - 4\widetilde{p_2^4} + \frac{1}{4}\widetilde{p_1^3} - \frac{1}{4}\widetilde{p_1^4} - 2\widetilde{p}_1\widetilde{p}_2 + 3\widetilde{p_1^2}\widetilde{p}_2 + 9\widetilde{p}_1\widetilde{p_2^2} - 6\widetilde{p_1^2}\widetilde{p_2^2} - 8\widetilde{p}_1\widetilde{p_2^3} - 2\widetilde{p_1^3}\widetilde{p}_2)\\
&-\frac{1}{r^2}(2\widetilde{p}_2 - 12\widetilde{p_2^2} + 20\widetilde{p_2^3} - 10\widetilde{p_2^4} - \frac{3}{8}\widetilde{p_1^2} + \widetilde{p_1^3} - \frac{5}{8}\widetilde{p_1^4} - 6\widetilde{p}_1\widetilde{p}_2 + 9\widetilde{p_1^2}\widetilde{p}_2 + 24\widetilde{p}_1\widetilde{p_2^2} - 15\widetilde{p_1^2}\widetilde{p_2^2}\\
&-20\widetilde{p}_1\widetilde{p_2^3} - 5\widetilde{p_1^3}\widetilde{p}_2) + \frac{1}{r^3}(\widetilde{p}_2 - 7\widetilde{p_2^2} + 12\widetilde{p_2^3} - 6\widetilde{p_2^4} + \frac{1}{16}\widetilde{p}_1 - \frac{7}{16}\widetilde{p_1^2} + \frac{3}{4}\widetilde{p_1^3} - \frac{3}{8}\widetilde{p_1^4} - 4\widetilde{p}_1\widetilde{p}_2 +\\
&6\widetilde{p_1^2}\widetilde{p}_2 + 15\widetilde{p}_1\widetilde{p_2^2} - 9\widetilde{p_1^2}\widetilde{p_2^2} - 12\widetilde{p}_1\widetilde{p_2^3} - 3\widetilde{p_1^3}\widetilde{p}_2)\\
=&\frac{1}{r}(\widetilde{p}_2 - 5\widetilde{p_2^2} + 8\widetilde{p_2^3} - 4\widetilde{p_2^4} + \frac{1}{4}\widetilde{p_1^3} - \frac{1}{4}\widetilde{p_1^4} - 2\widetilde{p}_1\widetilde{p}_2 + 3\widetilde{p_1^2}\widetilde{p}_2 + 9\widetilde{p}_1\widetilde{p_2^2} - 6\widetilde{p_1^2}\widetilde{p_2^2} - 8\widetilde{p}_1\widetilde{p_2^3} - 2\widetilde{p_1^3}\widetilde{p}_2)\\
&+O(r^{-2}).
\end{aligned}$$

We can use T_C to classify the underlying genetic models: the recessive model if $T_C > \varphi$; the additive model if $|T_C| < \varphi$; and the dominant model if $T_C < -\varphi$, where φ is a pre-specified threshold, here we set $\varphi = \varPhi^{-1}(0.95)$, the 95% quantile of a standard normal distribution. Based on the chosen model, we can use the optimal θ to construct the trend test. So the final test is

$$T = Z_0 \times I_{\{T_C > \varphi\}} + Z_{1/2} \times I_{\{|T_C| < \varphi\}} + Z_1 \times I_{\{T_C < -\varphi\}}.$$

2.2 Theoretical Properties

Since T is constructed by choosing the genetic model, we call T to be model-embedded trend test (MET). To obtain the p-value of the MET, we need to derive the asymptotic distribution of the MET, which is given in Theorem 2.1.

Theorem 2.1 *Assume that* $\frac{r}{n} \to \eta \in (0,1)$, *as* $n \to \infty$. $\varLambda_\theta^{-\frac{1}{2}}[(Z_\theta, T_C)^\tau - (\mu_{1\theta}, \mu_2)^\tau]$ *asymptotically follows a bivariate normal distribution* $N(\mathbf{0}, \boldsymbol{I}_2)$, *with*

$$\mu_{1\theta} = \frac{(p_2 + \theta p_1) - (q_2 + \theta q_1)}{\{\widehat{\mathrm{var}}_0(U_\theta)\}^{1/2}}, \quad \mu_2 = \frac{p_2 - (p_2 + p_1/2)^2}{\{\widehat{\mathrm{var}}_0(\varDelta_C)\}^{1/2}}, \quad \varLambda_\theta = \begin{pmatrix} \delta_{1\theta}^2 & \zeta_\theta \\ \zeta_\theta & \delta_2^2 \end{pmatrix},$$

and $\delta_{1\theta}^2 = \frac{\widehat{\mathrm{var}}_1(U_\theta)}{\widehat{\mathrm{var}}_0(U_\theta)}, \zeta_\theta = \widehat{\mathrm{cov}}\,(Z_\theta, T_C) = \frac{\widehat{\mathrm{cov}}(U_\theta, \varDelta_C)}{\{\widehat{\mathrm{var}}_0(U_\theta)\widehat{\mathrm{var}}_0(\varDelta_C)\}^{1/2}}, \delta_2^2 = \frac{\widehat{\mathrm{var}}_1(\varDelta_C)}{\widehat{\mathrm{var}}_0(\varDelta_C)}$, *where*

$$\begin{aligned}
\widehat{\mathrm{var}}_1\,(U_\theta) &= \frac{(\widehat{p}_2 + \theta^2\widehat{p}_1) - (\widehat{p}_2 + \theta\widehat{p}_1)^2}{r} + \frac{(\widehat{q}_2 + \theta^2\widehat{q}_1) - (\widehat{q}_2 + \theta\widehat{q}_1)^2}{s},\\
\widehat{\mathrm{var}}_1(\varDelta_C) &= \frac{1}{r}\Bigg(\widehat{p}_2 - 5\widehat{p_2^2} + 8\widehat{p_2^3} - 4\widehat{p_2^4} + \frac{1}{4}\widehat{p_1^3} - \frac{1}{4}\widehat{p_1^4} - 2\widehat{p}_1\widehat{p}_2 + 3\widehat{p_1^2}\widehat{p}_2\\
&\qquad +9\widehat{p}_1\widehat{p_2^2} - 6\widehat{p_1^2}\widehat{p_2^2} - 8\widehat{p}_1\widehat{p_2^3} - 2\widehat{p_1^3}\widehat{p}_2\Bigg),\\
\widehat{\mathrm{cov}}\,(U_\theta, \varDelta_C) &= \frac{1}{r}\Bigg(\widehat{p}_2 - 3\widehat{p_2^2} + 2\widehat{p_2^3} - \frac{1}{2}\theta\widehat{p_1^2} + \frac{1}{2}\theta\widehat{p_1^3} - (2\theta + 1)\widehat{p}_1\widehat{p}_2\\
&\qquad + \left(2\theta + \frac{1}{2}\right)\widehat{p_1^2}\widehat{p}_2 + (2\theta + 2)\widehat{p}_1\widehat{p_2^2}\Bigg).
\end{aligned}$$

Proof The proof is divided into three parts as follows.

(i) Firstly, since (r_0, r_1, r_2) and (s_0, s_1, s_2) independently follow the trinomial distributions, based on the conditions of theorem and the central limit theorem, we have

$$\begin{aligned}&\left(\sqrt{r}\right)^{-1}\left[(r_0, r_1, r_2)^\tau - (rp_0, rp_1, rp_2)^\tau\right] \xrightarrow{D} N(\mathbf{0}, \Lambda_r),\\ &\left(\sqrt{s}\right)^{-1}\left[(s_0, s_1, s_2)^\tau - (sq_0, sq_1, sq_2)^\tau\right] \xrightarrow{D} N(\mathbf{0}, \Lambda_s),\end{aligned}$$

where

$$\Lambda_r = \begin{pmatrix} p_0(1-p_0) & -p_0p_1 & -p_0p_2 \\ -p_0p_1 & p_1(1-p_1) & -p_1p_2 \\ -p_0p_2 & -p_1p_2 & p_2(1-p_2) \end{pmatrix}, \quad \Lambda_s = \begin{pmatrix} q_0(1-q_0) & -q_0q_1 & -q_0q_2 \\ -q_0q_1 & q_1(1-q_1) & -q_1q_2 \\ -q_0q_2 & -q_1q_2 & q_2(1-q_2) \end{pmatrix}.$$

(ii) Next, we aim to derive the asymptotic distribution of $(U_\theta, \Delta_C)^\tau$. For any t_1, t_2 satisfying $t_1^2 + t_2^2 = 1$,

$$(t_1, t_2)\begin{pmatrix} U_\theta \\ \Delta_C \end{pmatrix} = (t_1, t_2)\begin{pmatrix} (0 \ \ \theta \ \ 1)\begin{pmatrix} \widehat{p}_0 \\ \widehat{p}_1 \\ \widehat{p}_2 \end{pmatrix} - (0 \ \ \theta \ \ 1)\begin{pmatrix} \widehat{q}_0 \\ \widehat{q}_1 \\ \widehat{q}_2 \end{pmatrix} \\ \widehat{p}_2 - (\widehat{p}_2 + \widehat{p}_1/2)^2 \end{pmatrix}.$$

Since $\frac{r}{n} \to \eta$ as $n \to \infty$, it can be obtained that $\frac{r}{s} \to \frac{\eta}{1-\eta}$. Using the Delta method, as $\sqrt{r}\left[(\widehat{p}_0, \widehat{p}_1, \widehat{p}_2)^\tau - (p_0, p_1, p_2)^\tau\right] \xrightarrow{D} N(\mathbf{0}, \Lambda_r)$, $\sqrt{s}\left[(\widehat{q}_0, \widehat{q}_1, \widehat{q}_2)^\tau - (q_0, q_1, q_2)^\tau\right] \xrightarrow{D} N(\mathbf{0}, \Lambda_s)$, we have

$$\sqrt{r}(t_1, t_2)\begin{pmatrix} U_\theta - [(p_2 + \theta p_1) - (q_2 + \theta q_1)] \\ \Delta_C - \left[p_2 - (p_2 + p_1/2)^2\right] \end{pmatrix} \xrightarrow{D} N(0, \xi),$$

where

$$\xi = (t_1, t_2)\ \Omega\ (t_1, t_2)^\tau = (t_1, t_2)\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}(t_1, t_2)^\tau,$$

and

$$\begin{aligned}\Omega_{11} &= (p_2 + \theta^2 p_1) - (p_2 + \theta p_1)^2 + \frac{\eta}{1-\eta}\left((q_2 + \theta^2 q_1) - (q_2 + \theta q_1)^2\right),\\ \Omega_{12} &= \xi_{21} = p_2 - 3p_2^2 + 2p_2^3 - \frac{1}{2}\theta p_1^2 + \frac{1}{2}\theta p_1^3 - (2\theta + 1)p_1p_2 + \left(2\theta + \frac{1}{2}\right)p_1^2p_2 + (2\theta + 2)p_1p_2^2,\\ \Omega_{22} &= p_2 - 5p_2^2 + 8p_2^3 - 4p_2^4 + \frac{1}{4}p_1^3 - \frac{1}{4}p_1^4 - 2p_1p_2 + 3p_1^2p_2 + 9p_1p_2^2 - 6p_1^2p_2^2 - 8p_1p_2^3 - 2p_1^3p_2.\end{aligned}$$

Therefore, $\sqrt{r}\left[(U_\theta, \Delta_C)^\tau - \left((p_2 + \theta p_1) - (q_2 + \theta q_1), p_2 - (p_2 + p_1/2)^2\right)^\tau\right] \xrightarrow{D} N(\mathbf{0}, \Omega)$.

(iii) Finally, we derive the asymptotic distribution of $(Z_\theta, T_C)^\tau$. From the definitions of Z_θ and T_C, we have

$$\begin{aligned}
&[(Z_\theta, T_C)^\tau - (\mu_{1\theta}, \mu_2)^\tau] \\
&= \begin{pmatrix} \{\widehat{\text{var}}_0(U_\theta)\}^{-1/2} & 0 \\ 0 & \{\widehat{\text{var}}_0(\varDelta_C)\}^{-1/2} \end{pmatrix} \begin{pmatrix} U_\theta - [(p_2 + \theta p_1) - (q_2 + \theta q_1)] \\ \varDelta_C - \left[p_2 - (p_2 + p_1/2)^2\right] \end{pmatrix} \\
&= \begin{pmatrix} \{r\widehat{\text{var}}_0(U_\theta)\}^{-1/2} & 0 \\ 0 & \{r\widehat{\text{var}}_0(\varDelta_C)\}^{-1/2} \end{pmatrix} \sqrt{r} \begin{pmatrix} U_\theta - [(p_2 + \theta p_1) - (q_2 + \theta q_1)] \\ \varDelta_C - \left[p_2 - (p_2 + p_1/2)^2\right] \end{pmatrix}.
\end{aligned}$$

So, $\varLambda^{-\frac{1}{2}}\left[(Z_\theta, T_C)^\tau - (\mu_{1\theta}, \mu_2)^\tau\right]$ asymptotically follows a bivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_2)$, with

$$\varLambda = \begin{pmatrix} \{\xi_{11}\}^{-1/2} & 0 \\ 0 & \{\xi_{22}\}^{-1/2} \end{pmatrix} \begin{pmatrix} \varOmega_{11} & \varOmega_{12} \\ \varOmega_{21} & \varOmega_{22} \end{pmatrix} \begin{pmatrix} \{\xi_{11}\}^{-1/2} & 0 \\ 0 & \{\xi_{22}\}^{-1/2} \end{pmatrix},$$

where

$$\begin{aligned}
\xi_{11} &= \left[\frac{\eta}{1-\eta}p_2 + q_2 + \theta^2\left(\frac{\eta}{1-\eta}p_1 + q_1\right)\right] - (1-\eta)\left[\frac{\eta}{1-\eta}p_2 + q_2 + \theta\left(\frac{\eta}{1-\eta}p_1 + q_1\right)\right]^2, \\
\xi_{22} &= (\eta p_2 + (1-\eta)q_2) - 5\left(\eta p_2 + (1-\eta)q_2\right)^2 + 8\left(\eta p_2 + (1-\eta)q_2\right)^3 - 4\left(\eta p_2 + (1-\eta)q_2\right)^4 \\
&\quad + \frac{1}{4}\left(\eta p_1 + (1-\eta)q_1\right)^3 - \frac{1}{4}(\eta p_1 + (1-\eta)q_1)^4 \\
&\quad -2\left(\eta p_1 + (1-\eta)q_1\right)\left(\eta p_2 + (1-\eta)q_2\right) + 3\left(\eta p_1 + (1-\eta)q_1\right)^2\left(\eta p_2 + (1-\eta)q_2\right) \\
&\quad +9\left(\eta p_1 + (1-\eta)q_1\right)\left(\eta p_2 + (1-\eta)q_2\right)^2 - 6\left(\eta p_1 + (1-\eta)q_1\right)^2\left(\eta p_2 + (1-\eta)q_2\right)^2 \\
&\quad -8\left(\eta p_1 + (1-\eta)q_1\right)\left(\eta p_2 + (1-\eta)q_2\right)^3 - 2\left(\eta p_1 + (1-\eta)q_1\right)^3\left(\eta p_2 + (1-\eta)q_2\right).
\end{aligned}$$

Note that $r\widehat{\text{var}}_0(U_\theta) = \xi_{11} + o_p(1), r\widehat{\text{var}}_0(\varDelta_C) = \xi_{22} + o_p(1)$ as $\widetilde{p}_i, \widetilde{q}_i$ converging to $\eta p_i + (1-\eta)q_i$. Under the null hypothesis, $\varOmega_{11} = \xi_{11}, \varOmega_{22} = \xi_{22}$. So

$$\begin{aligned}
\varLambda_\theta &= \begin{pmatrix} \delta_{1\theta}^2 & \zeta_\theta \\ \zeta_\theta & \delta_2^2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{\widehat{\text{var}}_1(U_\theta)}{\widehat{\text{var}}_0(U_\theta)} & \frac{\widehat{\text{cov}}(U_\theta, \varDelta_C)}{\left\{\widehat{\text{var}}_0(U_\theta)\widehat{\text{var}}_0(\varDelta_C)\right\}^{1/2}} \\ \frac{\widehat{\text{cov}}(U_\theta, \varDelta_C)}{\left\{\widehat{\text{var}}_0(U_\theta)\widehat{\text{var}}_0(\varDelta_C)\right\}^{1/2}} & \frac{\widehat{\text{var}}_1(\varDelta_C)}{\widehat{\text{var}}_0(\varDelta_C)} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\sqrt{r}}\{\widehat{\text{var}}_0(U_\theta)\}^{-1/2} & 0 \\ 0 & \frac{1}{\sqrt{r}}\{\widehat{\text{var}}_0(\varDelta_C)\}^{-1/2} \end{pmatrix} \begin{pmatrix} r\widehat{\text{var}}_1(U_\theta) & r\widehat{\text{cov}}\,(U_\theta, \varDelta_C) \\ r\widehat{\text{cov}}\,(U_\theta, \varDelta_C) & r\widehat{\text{var}}_1(\varDelta_C) \end{pmatrix} \\
&\quad \begin{pmatrix} \frac{1}{\sqrt{r}}\{\widehat{\text{var}}_0(U_\theta)\}^{-1/2} & 0 \\ 0 & \frac{1}{\sqrt{r}}\{\widehat{\text{var}}_0(\varDelta_C)\}^{-1/2} \end{pmatrix}.
\end{aligned}$$

It can be seen that $\varLambda_\theta$ converges to $\varLambda$ as $n \to \infty$. Therefore, $\varLambda_\theta^{-\frac{1}{2}}\left[(Z_\theta, T_C)^\tau - (\mu_{1\theta}, \mu_2)^\tau\right]$ asymptotically follows a bivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_2)$. ▌

Based on Theorem 2.1, for the observation of T, t, the p-value can be calculated as

$$\begin{aligned}
&p\text{-value}\\
&= \mathrm{Pr}_{H_0}(T > |t|) + \mathrm{Pr}_{H_0}(T < -|t|)\\
&= \mathrm{Pr}(T|_{H_0} > |t|, T_C > \varphi) + \mathrm{Pr}(T|_{H_0} > |t|, |T_C| < \varphi) + \mathrm{Pr}(T|_{H_0} > |t|, T_C < -\varphi)\\
&\quad + \mathrm{Pr}(T|_{H_0} < -|t|, T_C > \varphi) + \mathrm{Pr}(T|_{H_0} < -|t|, |T_C| < \varphi) + \mathrm{Pr}(T|_{H_0} < -|t|, T_C < -\varphi)\\
&= \left(\int_{|t|}^{+\infty}\int_{\varphi}^{+\infty} f(u,v)dudv + \int_{|t|}^{+\infty}\int_{-\varphi}^{\varphi} f(u,v)dudv + \int_{|t|}^{+\infty}\int_{-\infty}^{-\varphi} f(u,v)dudv\right)\\
&\quad + \left(\int_{-\infty}^{-|t|}\int_{\varphi}^{+\infty} f(u,v)dudv + \int_{-\infty}^{-|t|}\int_{-\varphi}^{\varphi} f(u,v)dudv + \int_{-\infty}^{-|t|}\int_{-\infty}^{-\varphi} f(u,v)dudv\right)\\
&= \int_{\frac{\varphi-\mu_2}{\delta_2}}^{+\infty} \varPhi\left(\frac{-|t| + \zeta_0 w/\delta_2}{\{1-(\zeta_0/\delta_2)^2\}^{1/2}}\right) d\varPhi(w) + \int_{\frac{-\varphi-\mu_2}{\delta_2}}^{\frac{\varphi-\mu_2}{\delta_2}} \varPhi\left(\frac{-|t| + \zeta_{1/2} w/\delta_2}{\{1-(\zeta_{1/2}/\delta_2)^2\}^{1/2}}\right) d\varPhi(w)\\
&\quad + \int_{-\infty}^{\frac{-\varphi-\mu_2}{\delta_2}} \varPhi\left(\frac{-|t| + \zeta_1 w/\delta_2}{\{1-(\zeta_1/\delta_2)^2\}^{1/2}}\right) d\varPhi(w) + \int_{\frac{\varphi-\mu_2}{\delta_2}}^{+\infty} \varPhi\left(\frac{-|t| - \zeta_0 w/\delta_2}{\{1-(\zeta_0/\delta_2)^2\}^{1/2}}\right) d\varPhi(w)\\
&\quad + \int_{\frac{-\varphi-\mu_2}{\delta_2}}^{\frac{\varphi-\mu_2}{\delta_2}} \varPhi\left(\frac{-|t| - \zeta_{1/2} w/\delta_2}{\{1-(\zeta_{1/2}/\delta_2)^2\}^{1/2}}\right) d\varPhi(w) + \int_{-\infty}^{\frac{-\varphi-\mu_2}{\delta_2}} \varPhi\left(\frac{-|t| - \zeta_1 w/\delta_2}{\{1-(\zeta_1/\delta_2)^2\}^{1/2}}\right) d\varPhi(w),
\end{aligned}$$

where $f(u,v)$ is the probability density function of Z_θ and T_C

$$f(u,v) = \left(2\pi|\varLambda_\theta|^{1/2}\right)^{-1} \mathrm{e}^{-\frac{1}{2}(u-\mu_{1\theta}, v-\mu_2)\varLambda_\theta^{-1}(u-\mu_{1\theta}, v-\mu_2)^\tau}.$$

3 Numerical Results

3.1 Simulation Settings

We consider five MAFs as 0.05, 0.15, 0.25, 0.35, 0.45. The disease prevalence k is set to be 0.01. The genotype relative risks are chosen as $\lambda_1 = 1, \lambda_2 = 1.5$ for the recessive model, $\lambda_1 = 1.2, \lambda_2 = 1.4$ for the additive model and $\lambda_1 = 1.3, \lambda_2 = 1.3$ for the dominant model. Assume that the HWE holds in the source population and the genotype frequencies (p_0, p_1, p_2) in cases and (q_0, q_1, q_2) in controls are calculated using the expressions given in (1) and (2). Then 1,000 cases and 1,000 controls are independently generated from the trinomial distributions $\mathrm{Mul}(1000; p_0, p_1, p_2)$ and $\mathrm{Mul}(1000; q_0, q_1, q_2)$, respectively. The nominal significance level $\alpha = 0.05$.

3.2 The Empirical Type I Error Rates

Table 1 shows the empirical type I error rates of the TPP, the TTA, the CHI2, the MAX3, the MIN2, and the MET. Obviously all the values are very close to the nominal significance level of 0.05, which means all the considered methods can control the type I error rates well. For example, when MAF = 0.15 the empirical type I error rates of the TPP, the TTA, the CHI2, the MAX3, the MIN2, and the MET are 0.052, 0.053, 0.051, 0.050, 0.051, and 0.053, respectively.

Table 1 The empirical type I error rates of the TPP, the TTA, the CHI2, the MAX3, the MIN2, and the MET (1,000 cases and 1,000 controls, $\varphi = \Phi^{-1}(0.95)$)

MAF	TPP	TTA	CHI2	MAX3	MIN2	MET
0.05	0.045	0.051	0.040	0.035	0.041	0.050
0.15	0.052	0.053	0.051	0.050	0.051	0.053
0.25	0.053	0.052	0.047	0.051	0.053	0.054
0.35	0.050	0.051	0.050	0.049	0.051	0.050
0.45	0.053	0.054	0.052	0.053	0.053	0.054

3.3 The Empirical Power

Figure 1 shows the empirical power of the TPP, the TTA, the CHI2, the MAX3, the MIN2, and the MET under three genetic models. The powers are computed based on 10,000 replicates with $(\lambda_1, \lambda_2) = (1, 1.5)$ for the recessive model, $(\lambda_1, \lambda_2) = (1.2, 1.4)$ for the additive model and $(\lambda_1, \lambda_2) = (1.3, 1.3)$ for the dominant model, respectively. It can be obviously seen that, the MET is more powerful than other procedures for the recessive and dominant models, but slightly less effective than the TTA for the additive model. For example, when MAF $= 0.25$ under the recessive model, the empirical power of the TPP, the TTA, the CHI2, the MAX3, the MIN2, and the MET are 0.630, 0.364, 0.559, 0.578, 0.525, and 0.698, respectively. When MAF $= 0.35$ under the additive model, the corresponding values are 0.738, 0.750, 0.652, 0.700, 0.706, and 0.747, respectively. This is sensible because the TTA is optimal for the additive model. So, the MET has greater efficiency robustness than the TPP, the TTA, the CHI2, the MAX3, and the MIN2.

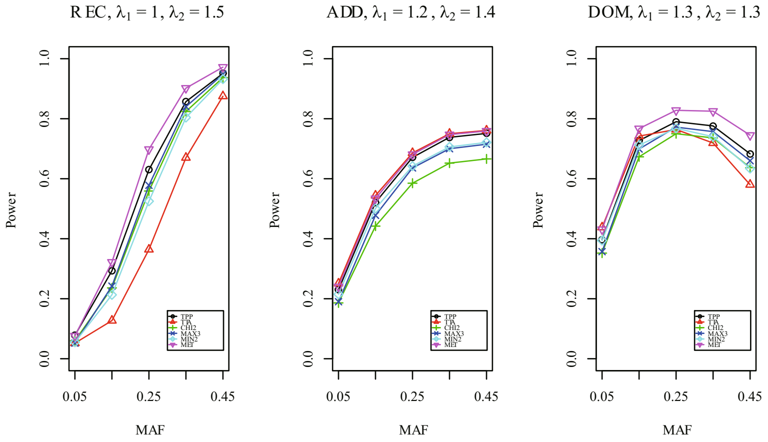

Figure 1 The empirical power of the TPP, the TTA, the CHI2, the MAX3, the MIN2, and the MET based on 10,000 replicates. The number of cases and control are equal and equals to 1,000, the genotype relative risks $(\lambda_1, \lambda_2) = (1, 1.5)$ is for the recessive model (REC), $(\lambda_1, \lambda_2) = (1.2, 1.4)$ is for the additive model (ADD), and $(\lambda_1, \lambda_2) = (1.3, 1.3)$ is for the dominant model (DOM), and $\varphi = \Phi^{-1}(0.95)$

3.4 Application to Detect the Association Between the SNP rs5007171 and the Coronary Artery Disease

To further illustrate the use of the MET, we consider the SNP rs5007171 with the data being from the coronary artery disease genome-wide association study[2]. The SNP has been reported to be moderately associated with the coronary artery disease. For the SNP rs5007171, the genotypes counts are $(r_0, r_1, r_2) = (29, 684, 1173)$ in cases and $(s_0, s_1, s_2) = (126, 966, 1803)$ in controls. The corresponding p-values of the TPP, the TTA, the CHI2, the MAX3, the MIN2, and the MET, are $7.851 \times 10^{-8}, 0.096, 2.166 \times 10^{-7}, 1.570 \times 10^{-7}, 3.741 \times 10^{-7}$, and 7.779×10^{-8}. The p-values of the TTA, the CHI2, the MAX3, and the MIN2 are larger than the genome-wide significance level $1.456 \times 10^{-7} = 0.05/343413$[15]. For the TPP, followed by Zheng and Ng[13], we obtain the adjusted significance level $\alpha^* = 0.037$, and the p-value of the TPP is smaller than $0.037/343413 = 1.077 \times 10^{-7}$. Therefore, both the MET and the TPP can significantly detect the association between the SNP rs5007171 and the coronary artery disease and the MET gives smaller p-value than the TPP.

4 Conclusions

In genetic association analysis, the traditional trend test was usually used for a case-control design under a known genetic model. However, in practice, the underlying genetic model of the disease loci was uncertain. It may lose power when the model was misspecified. Therefore, the robust tests were more suitable under this scenario. In this work, we proposed a model-embedded trend test to handle the genetic model uncertainty. Simulation studies shown the proposed MET was more robust than the existing TPP, TTA, CHI2, MAX3, and MIN2.

Zheng and Ng[13] also considered using model selection to do association analysis. However, to control the type I error rate of their TPP, it needed to calculate the adjusted significance level for each SNP. It was very time-consuming for testing more than 500,000 SNPs in a genome-wide association study. In this work, we proposed to construct an omnibus test statistic, and then derived its asymptotic distribution. We derived an effective formula based on the double-integration to calculate the p-value of the MET. It provided a more convenient and intuitive way to compute the empirical power than the TPP. Meanwhile, it greatly reduced the computational intensity to obtain the p-value compared with the resampling procedure. Sladek, et al.[8] used 10,000 permutations to calculate the p-values of the MAX3. A further topic is to consider the penalized method[16] by putting the genetic model in the logistic regression. The R code to implement the MET is given in the website: http://www.statsci.amss.ac.cn/yjscy/yjy/lqz/201510/t20151026_313273.html.

References

[1] Klein R J, Zeiss C, Chew E Y, et al., Complement factor H polymorphism in age-related macular degeneration, *Science*, 2005, **308**(5720): 385–389.

[2] Craddock N J and Gwilliam R, Wellcome Trust Case Control Consortium (WTCCC) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, 2007, **447**(7145): 661–678.

[3] Zaykin D V, Zhivotovsky L A, Westfall P H, et al., Truncated product method for combining P-values, *Genetic Epidemiology*, 2002, **22**(2): 170–185.

[4] Hu X, Zhang W, Zhang S, et al., Group-combined p-values with applications to genetic association studies, *Bioinformatics*, 2016, **32**(18): btw314.

[5] Zheng G, Zhang W, Xu J, et al., Genetic risks and genetic model specification, *Journal of Theoretical Biology*, 2016, **403**: 68–74.

[6] Skol A D, Scott L J, Abecasis G R, et al., Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies, *Nature Genetics*, 2006, **38**(2): 209–213.

[7] Yeager M, Orr N, and Wang Z, Genome-wide association study of prostate cancer identifies a second risk locus at 8q24, *Nature Genetics*, 2007, **39**(5): 645–649.

[8] Sladek R, Rocheleau G, and Balkau B, A genome-wide association study identifies novel risk loci for type 2 diabetes, *Nature*, 2007, **445**(7130): 881–885.

[9] Wang K and Sheffield V C, A constrained-likelihood approach to marker-trait association studies, *The American Journal of Human Genetics*, 2005, **77**(5): 768–780.

[10] Wigginton J E, Cutler D J, and Abecasis G R, A note on exact tests of Hardy-Weinberg equilibrium, *The American Journal of Human Genetics*, 2005, **76**(5): 887–893.

[11] Schaid D J, Batzler A J, Jenkins G D, et al., Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata, *The American Journal of Human Genetics*, 2006, **79**(6): 1071–1080.

[12] Zhang W and Li Q, Incorporating Hardy-Weinberg equilibrium law to enhance the association strength for ordinal trait genetic study, *Annals of Human Genetics*, 2016, **80**(2): 102–112.

[13] Zheng G and Ng H K T, Genetic model selection in two-phase analysis for case-control association studies, *Biostatistics*, 2008, **9**(3): 391–399.

[14] Sasieni P D, From genotypes to genes: Doubling the sample size, *Biometrics*, 1997, **53**: 1253–1261.

[15] Li Q, Zheng G, Li Z, et al., Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies, *Annals of Human Genetics*, 2008, **72**(3): 397–406.

[16] Lu Y, Zhang R, and Hu B, The adaptive LASSO spline estimation of single-index model, *Journal of Systems Science and Complexity*, 2016, **29**(4): 1100–1111.