An analytical comparison of the principal component method and the mixed effects model for genetic association studies

Kai Wang

Department of Biostatistics, College of Public Health,

University of Iowa, Iowa City, IA 52242

Yingwei Peng

Department of Community Health and Epidemiology

Department of Mathematics and Statistics

Queens University, Kingston, ON K7L 3N6, Canada

Received _____; accepted _____

Corresponding author: Kai Wang, PhD, Department of Biostatistics, N322 CPHB, College of Public Health, University of Iowa, Iowa City, IA 52242. E-mail: kai-wang@uiowa.edu, phone: (319) 384-1594, fax: (319) 384-1591.

ABSTRACT

The principal component method and the mixed effects model represent two popular approaches for controlling for population structure and cryptic relatedness in genetic association studies. There are quite few studies comparing their performance. However, these comparisons are typically conducted using simulation studies and their implications are therefore limited. We report an analytical study of these two approaches in the presence of cryptic relatedness and population structure in terms of their validity and efficiency. We show that in the presence of cryptic relatedness, both methods are valid but the mixed effects model is more powerful. In the presence of population structure, both methods can be invalid and be conservative or anti-conservative. Conditions under which they are valid are provided. These conclusions are demonstrated through examples and simulation studies.

Subject headings: principal component, mixed effects model, population stratification, cryptic relatedness, genetic association

Introduction

Population stratification is a well-known confounding factor in genetic association studies. It can lead to spurious association if not dealt with appropriately. It exists even in populations that seem to be homogeneous, for instance, European American[1] and Han Chinese[2,3]. Nowadays more and more large consortia and collaborations are routinely formed in order to identify genetic factors of small effect. For instance, the GENEVA project[4] involves 14 participating studies covering a wide range of primary phenotypes. Each study consists of thousands of study subjects. It is not surprising that population stratification is a pressing issue to be addressed in analyzing data from large-scale multi-center studies. Indeed, the GENEVA coordinating center has developed its own version of software package to handle this issue (https://www.genevastudy.org/Accomplishments/software).

The genomic control method (GC)[5–9] is a popular method for handling population stratification. It modifies the Cochran-Armitage (CA) test for trend by a deflating factor. This factor is estimated using markers that are known to be unassociated with the phenotype (null markers). The structured population method[10–12] tries to infer the subpopulation structure first. Subsequent analyses are then conducted within each subpopulation and the results are summarized. The third method is to create surrogates for population stratification using null markers. For example, the principal component analysis (PCA) method[13] uses the first few principal components of the matrix of relatedness as covariates in a regression analysis. Similarly, one can use the first few components from a partial least-squares regression analysis[14]. Given the huge amount of genome-wide SNP genotype data, it is feasible to estimate the degree of relatedness of study subjects[15–18]. This possibility has resulted in novel approaches to genome-wide association studies (GWAS). The mixed effects model method[19–22] can take into account fixed effects such as age and gender while modeling population structure and cryptic relatedness as random

effects. Preliminary studies have found that these methods perform better than methods that do not model relatedness of study subjects[20,21,23]. Most of these methods have been implemented in various computer programs such as STRUCTURE[24,25], ADMIXTURE[26], EIGENSTRAT[13], ROADTRIPS[27], EMMAX[20], and TASSEL[19,21]. A review of statistical methods is[28].

The PCA method and the mixed effects model are the most popular methods in genetic association studies. There are quite few comparisons regarding the performance of these two methods. However, these comparisons are almost all conducted via simulation studies[23,29–31]. It is not clear whether the conclusions reached in such studies are applicable beyond the simulated situations. To avoid the limitations of simulation studies, we analytically compare the PCA and the mixed effects model in terms of their validity and the efficiency in testing genetic effects.

In what follows, we will first describe the PCA method and the mixed effects model. These two methods are then compared analytically first in the case of relatedness and then in the case of population stratification. The main results are presented in two propositions. Examples are used to illustrate the implication of these propositions. Simulation studies are then conducted to demonstrate the conclusions.

Cryptic relatedness and population stratification

Let y denote the value of a quantitative trait for a study subject (the case of dichotomous y will be discussed later). It is assumed that y is determined by the following additive polygenetic model[32]

$$y = u + \sum_j h_j + \epsilon$$

where u is the mean value of y and h_j is the contribution of the j-th genetic factor to trait y. The variance of y, denoted by σ_y^2, is the sum of the variance of the polygenic effect, denoted by σ_G^2, and the variance of the environmental effect, denoted by σ_e^2. That is,

$$\sigma_y^2 = \sigma_G^2 + \sigma_e^2.$$

In the presence of cryptic relatedness, the polygenic effect of the study subjects are correlated even though they seem to be otherwise. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ be an $n \times 1$ vector of trait values on n study subjects and $\mathbf{S}$ the matrix of pair-wise relatedness such as identical by state (IBS) or Balding-Nichols similarity. Without loss of generality, assume that $\mathbf{S}$ has been standardized such that its diagonal elements are all equal to 1. The variance matrix of $\mathbf{y}$ is

$$\mathbf{\Omega} = \sigma_G^2 \mathbf{S} + \sigma_e^2 \mathbf{I}. \tag{1}$$

The matrix $\mathbf{S}$ is generally unknown unless the genealogical information is given. However, given the large amount of SNPs in a genome-wide association studies, it can be reliably estimated[33–38]. For this reason, it is assumed that the $\mathbf{S}$ is known hereafter. Furthermore, σ_G^2 and σ_e^2 are segregation parameters that measure the contribution of the genetic factors and the environment factor. They can be estimated beforehand and will be treated as known hereafter. This approach has been used elsewhere (e.g., the program EMMAX[20]).

If there exists population stratification, the value of u is no longer the same for all individuals. Instead, it is population-dependent. Let u_k be its value for the k-th population. These u_ks are not observable otherwise the problem would be trivial: one can use indicator variables for subpopulations. The variation in u_ks contributes to phenotypic variation and inter-individual correlation since subjects from the same population share a common u_k. That is, population stratification induces correlation among study subjects. Let $\mathbf{1}_k$ denote the vector of 1s whose length is the same as the number subjects from the k-th population. Let K be the total number of populations in the data. Suppose that the subjects are

organized in such a way that subjects from the same subpopulation are indexed next to each other, the variance matrix of $\mathbf{y}$ becomes

$$\mathbf{\Omega}^* = \sigma_u^2 \tilde{\mathbf{S}} + \sigma_G^2 \mathbf{S} + \sigma_e^2 \mathbf{I}$$

where $\tilde{\mathbf{S}} = \text{diag}(\mathbf{1}_1\mathbf{1}_1', \mathbf{1}_2\mathbf{1}_2', \ldots, \mathbf{1}_K\mathbf{1}_K')$ is the induced relatedness matrix by population stratification and σ_u^2 is the variance of elements in vector $(u_1\mathbf{1}_1', \ldots, u_K\mathbf{1}_K')'$. The matrix $\mathbf{\Omega}^*$ can be written $\sigma_{G'}^2\mathbf{S}^* + \sigma_e^2\mathbf{I}$ where

$$\sigma_{G'}^2 = \sigma_u^2 + \sigma_G^2$$

and

$$\mathbf{S}^* = \frac{\sigma_u^2}{\sigma_{G'}^2}\tilde{\mathbf{S}} + \frac{\sigma_G^2}{\sigma_{G'}^2}\mathbf{S}.$$

One can estimate $\mathbf{S}^*$ and $\sigma_{G'}^2$ from genomic data but not σ_u^2, σ_G^2, and $\mathbf{S}$ individually. The effect of population stratification confounds with the polygenic effect.

In summary, the phenotype data $\mathbf{y}$ can be modeled in the following way. Let $\mathbf{g}$ denote the $n \times 1$ vector of the genotype scores at the single-nucleotide polymorphism (SNP) being tested for association. Each component of $\mathbf{g}$ is a genotype score. It is the number of copies of a reference allele an individual has and assumes value 0, 1, or 2. Let β be the effect size of one unit change in genotype score. The phenotype vector $\mathbf{y}$ follows the following model:

$$\mathbf{y} = \mathbf{u} + \beta\mathbf{g} + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{\Omega})$. If there is no population stratification, $\mathbf{u}$ is a vector of constants. That is, $\mathbf{u} = u\mathbf{1}$ for some u. If there is population stratification, $\mathbf{u}$ is segment-wise constant: $\mathbf{u} = (u_1\mathbf{1}_1', \ldots, u_K\mathbf{1}_K)'$. In either case, the distribution of $\mathbf{y}$ is multivariate normal with mean vector $\mathbf{u} + \beta\mathbf{g}$ and variance matrix $\mathbf{\Omega} = \sigma_G^2\mathbf{S} + \sigma_e^2\mathbf{I}$.

The PCA method and the mixed effects model

The analysis model is typically different from the true model. Given the phenotype data $\mathbf{y}$, genotype data $\mathbf{g}$, relatedness matrix $\mathbf{S}$ (or matrix $\mathbf{S}^*$ if there is population stratification), and variance-covariance matrix $\mathbf{\Omega}$, we describe the PCA method and the mixed effects model below.

Let the eigenvalues of matrix $\mathbf{S}$ be denoted by (in descending order of their values) $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and corresponding eigenvectors denoted by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The number of non-zero eigenvalues is equal to the rank of $\mathbf{S}$. By definition, these eigenvalues and eigenvectors satisfy

$$\mathbf{S}\mathbf{x}_i = \lambda_i \mathbf{x}_i$$

and $\mathbf{x}_i'\mathbf{x}_i = 1$, $\mathbf{x}_i'\mathbf{x}_j = 0$ for $i \neq j$. The matrix $\mathbf{S}$ is substituted by $\mathbf{S}^*$ if there is population stratification.

The PCA method uses the eigenvectors corresponding to the m largest eigenvalues (a.k.a. axes of variation[13]) as covariates and fits the following model:

$$\begin{aligned} \mathbf{y} &= \alpha_0 \mathbf{1} + \sum_{i=1}^{m} \alpha_i \mathbf{x}_i + \beta \mathbf{g} + \boldsymbol{\delta} \\ &= \mathbf{Z}\boldsymbol{\alpha} + \beta \mathbf{g} + \boldsymbol{\delta} \end{aligned}$$

where $\mathbf{Z} = (\mathbf{1}, \mathbf{X}_m)$ with $\mathbf{X}_m = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ and m is a predetermined number. Note that if $\mathbf{1}$ happens to be a eigenvector, then this eigenvector needs to be removed in order to avoid collinearity with the intercept term (see Example 1 to be introduced later for an example). The estimate of β is

$$\tilde{\beta} = \frac{\mathbf{g}'\mathbf{P}_1\mathbf{y}}{\mathbf{g}'\mathbf{P}_1\mathbf{g}}$$

where

$$\mathbf{P}_1 = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$$

is the projection matrix that projects a $n \times 1$ vector into a space orthogonal to the space spanned by the columns of matrix $\mathbf{Z}$. The variance of $\tilde{\beta}$ is

$$Var(\tilde{\beta}) = \frac{\mathbf{g}'\mathbf{P}_1\mathbf{\Omega}\mathbf{P}_1\mathbf{g}}{(\mathbf{g}'\mathbf{P}_1\mathbf{g})^2}.$$

For the mixed effects model method, the fitted model is

$$\mathbf{y} = \beta_0\mathbf{1} + \beta\mathbf{g} + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{\Omega}^*)$ with $\mathbf{\Omega}^* = \sigma_u^2\tilde{\mathbf{S}} + \mathbf{\Omega}$. Let $\mathbf{A}$ be the matrix generated from the Cholesky decomposition of matrix $\mathbf{\Omega}^*$ that satisfies $\mathbf{\Omega}^* = \mathbf{A}\mathbf{A}'$. The inverse of $\mathbf{\Omega}^*$ is $(\mathbf{\Omega}^*)^{-1} = (\mathbf{A}')^{-1}\mathbf{A}^{-1}$. Since $\mathbf{\Omega}^*$ is taken to be known, the mixed effects model estimate of β is the same as its generalized least square estimate. To find out the latter, rewrite (3) as

$$\mathbf{A}^{-1}\mathbf{y} = \beta_0\mathbf{A}^{-1}\mathbf{1} + \beta\mathbf{A}^{-1}\mathbf{g} + \mathbf{A}^{-1}\boldsymbol{\epsilon}.$$

The variance matrix of $\mathbf{A}^{-1}\boldsymbol{\epsilon}$ is $\mathbf{A}^{-1}(\mathbf{A}\mathbf{A}')(\mathbf{A}')^{-1} = \mathbf{I}$. So the mixed effects model estimate of β is

$$\hat{\beta} = \frac{\mathbf{g}'\mathbf{P}_2\mathbf{y}}{\mathbf{g}'\mathbf{P}_2\mathbf{g}}$$

where

$$\begin{aligned}\mathbf{P}_2 &= (\mathbf{A}')^{-1}[\mathbf{I} - \mathbf{A}^{-1}\mathbf{1}[(\mathbf{A}^{-1}\mathbf{1})'\mathbf{A}^{-1}\mathbf{1}]^{-1}(\mathbf{A}^{-1}\mathbf{1})']\mathbf{A}^{-1} \\ &= (\mathbf{\Omega}^*)^{-1} - (\mathbf{\Omega}^*)^{-1}\mathbf{1}(\mathbf{1}'(\mathbf{\Omega}^*)^{-1}\mathbf{1})^{-1}\mathbf{1}'(\mathbf{\Omega}^*)^{-1}.\end{aligned}$$

It is easy to see that $\mathbf{P}_2$ satisfies $\mathbf{P}_2\mathbf{1} = \mathbf{0}$. Since the variance of $\mathbf{y}$ in the generating model is $\mathbf{\Omega}$, the variance of $\hat{\beta}$ is

$$Var(\hat{\beta}) = \frac{\mathbf{g}'\mathbf{P}_2\mathbf{\Omega}\mathbf{P}_2\mathbf{g}}{(\mathbf{g}'\mathbf{P}_2\mathbf{g})^2}.$$

If there is no population stratification, then $\mathbf{\Omega}^* = \mathbf{\Omega}$. It is straightforward to verify that $\mathbf{P}_2\mathbf{\Omega}\mathbf{P}_2 = \mathbf{P}_2$. In this case, the variance of $\hat{\beta}$ becomes

$$Var(\hat{\beta}) = \frac{1}{\mathbf{g}'\mathbf{P}_2\mathbf{g}}.$$

Main Results

We are interested in which method, the PCA method or the mixed effects model method, is valid in the presence of cryptic relatedness or population stratification. And if so, which method is more efficient. The main results of this report is summarized in the following two propositions.

Proposition 1 *If there exists cryptic relatedness but not population stratification (that is,* $\mathbf{u} = u\mathbf{1}$ *in (2) for some value* u*) and the PCA method contains an intercept term, both* $\tilde{\beta}$ *and* $\hat{\beta}$ *are unbiased estimates of* β*. However, the former has a larger variance than the latter does. That is,* $Var(\tilde{\beta}) \geq Var(\hat{\beta})$*. Consequently, the mixed effects model is more efficient.*

The proof of unbiasedness is straightforward. $\tilde{\beta}$ is unbiased because $\mathbf{P}_1\mathbf{1} = \mathbf{0}$ and

$$\begin{aligned} E(\tilde{\beta}) &= (\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}[\mathbf{g}'\mathbf{P}_1 E(\mathbf{y})] \\ &= (\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}[\mathbf{g}'\mathbf{P}_1(\mathbf{u} + \beta\mathbf{g})] \\ &= u(\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}\mathbf{g}'\mathbf{P}_1\mathbf{1} + \beta \\ &= \beta. \end{aligned}$$

Since $\mathbf{P}_2\mathbf{1} = \mathbf{0}$, the unbiasedness of $\hat{\beta}$ can be shown in a similar way:

$$\begin{aligned} E(\hat{\beta}) &= (\mathbf{g}'\mathbf{P}_2\mathbf{g})^{-1}[\mathbf{g}'\mathbf{P}_2 E(\mathbf{y})] \\ &= (\mathbf{g}'\mathbf{P}_2\mathbf{g})^{-1}[\mathbf{g}'\mathbf{P}_2(u\mathbf{1} + \beta\mathbf{g})] \\ &= \beta. \end{aligned}$$

The second part can be shown by using the Schwarz inequality as follows. Since $\mathbf{\Omega}$ is positive definite, define inner product $\langle \cdot, \cdot \rangle$ such that $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{\Omega}\mathbf{y}$. Notice that

$\mathbf{P}_2\mathbf{\Omega}\mathbf{P}_2 = \mathbf{P}_2$ and

$$\begin{aligned}\mathbf{P}_2\mathbf{\Omega}\mathbf{P}_1 &= (\mathbf{I} - \mathbf{\Omega}^{-1}\mathbf{1}(\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1})^{-1}\mathbf{1}')\mathbf{P}_1 \\ &= \mathbf{P}_1,\end{aligned}$$

we have

$$\begin{aligned}(\mathbf{g}'\mathbf{P}_2\mathbf{g}) \cdot (\mathbf{g}'\mathbf{P}_1\mathbf{\Omega}\mathbf{P}_1\mathbf{g}) &= \langle \mathbf{P}_2\mathbf{g}, \mathbf{P}_2\mathbf{g}\rangle \cdot \langle \mathbf{P}_1\mathbf{g}, \mathbf{P}_1\mathbf{g}\rangle \\ &\geq (\langle \mathbf{P}_2\mathbf{g}, \mathbf{P}_1\mathbf{g}\rangle)^2 \\ &= (\mathbf{g}'\mathbf{P}_2\mathbf{\Omega}\mathbf{P}_1\mathbf{g})^2 \\ &= (\mathbf{g}'\mathbf{P}_1\mathbf{g})^2.\end{aligned}$$

Using the definition of $Var(\tilde{\beta})$ and $Var(\hat{\beta})$, it immediately follows that $Var(\tilde{\beta}) \geq Var(\hat{\beta})$.

It is a well known textbook result that $\hat{\beta}$ is the best linear unbiased estimator of β among generalized linear models and thus has the smallest variance. However, this proposition is not trivially true because in this situation there are more covariates available to the PCA model than to the mixed effects model. These variables are the principal components from the relatedness matrix $\mathbf{S}$. Actually, these covariates do not need to be principal components of $\mathbf{S}$. From the proof procedure one can see that no specific properties of $\{\mathbf{x}_k\}$ being principal components are used. This proposition remains true even when $\{\mathbf{x}_k\}$ are not eigenvectors as long as the intercept term is included.

Proposition 2 *In the presence of population stratification (i.e.,* $\mathbf{u} = (u_1\mathbf{1}_1', u_2\mathbf{1}_2', \ldots, u_K\mathbf{1}_K')'$ *in (2)), We have*

1. $E(\tilde{\beta}) = (\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}(\mathbf{g}'\mathbf{P}_1\mathbf{u}) + \beta$. *That is,* $\tilde{\beta}$ *is a biased estimator of* β *unless* $\mathbf{u}$ *satisfies* $\mathbf{g}'\mathbf{P}_1\mathbf{u} = 0$. *The bias is* $E(\tilde{\beta}) - \beta = (\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}(\mathbf{g}'\mathbf{P}_1\mathbf{u})$. *Specifically, this bias is 0 if* $\mathbf{u}$ *is a linear combination of the column vectors of matrix* $\mathbf{Z}$.

2. $E(\hat{\beta}) = (\mathbf{g}'\mathbf{P}_2\mathbf{g})^{-1}(\mathbf{g}'\mathbf{P}_2\mathbf{u}) + \beta$. *That is,* $\hat{\beta}$ *is a biased estimator of* β *unless* $\mathbf{u}$ *satisfies* $\mathbf{g}'\mathbf{P}_2\mathbf{u} = 0$. *The bias is* $E(\hat{\beta}) - \beta = (\mathbf{g}'\mathbf{P}_2\mathbf{g})^{-1}(\mathbf{g}'\mathbf{P}_2\mathbf{u})$.

The expected value of $\tilde{\beta}$ can be computed directly as follows:

$$\begin{aligned} E(\tilde{\beta}) &= (\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}[\mathbf{g}'\mathbf{P}_1 E(\mathbf{y})] \\ &= (\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}[\mathbf{g}'\mathbf{P}_1(\mathbf{u} + \beta\mathbf{g})] \\ &= (\mathbf{g}'\mathbf{P}_1\mathbf{g})^{-1}(\mathbf{g}'\mathbf{P}_1\mathbf{u}) + \beta. \end{aligned}$$

Similarly, the expected value of $\hat{\beta}$ is $(\mathbf{g}'\mathbf{P}_2\mathbf{g})^{-1}[\mathbf{g}'\mathbf{P}_2\mathbf{u}] + \beta$. It is not clear whether $Var(\tilde{\beta})$ or $Var(\hat{\beta})$ is larger because matrix $\mathbf{P}_2$ involves the additional relatedness matrix $\tilde{\mathbf{S}}$ induced by population stratification.

This proposition indicates that both methods can be biased. The direction of the bias is determined by $\mathbf{g}'\mathbf{P}_1\mathbf{u}$ and $\mathbf{g}'\mathbf{P}_2\mathbf{u}$, respectively. This observation has two implications. First, if there is genetic association (i.e., $\beta = 0$), both methods are invalid. Second, if there is association, the both methods can be conservative or anti-conservative. The extent of which is affected by not only the direction of the bias in the estimate of β and the true value of β, but also the variance of the estimate of β. It is unclear which method has smaller bias in β estimate or smaller variance in β estimate. We will investigate this issue more in the examples introduced next and in the simulation studies later.

So far our focus is on continuous traits. When the trait is dichotomous, a common approach is to code the trait values as 0s and 1s and treat them as if they were continuous. For instance,[27],[39], and[20]. This approach is asymptotically equivalent to the test based on a logistic regression. The conclusion of Proposition 1 and Proposition 2 applies to such recoded binary trait values.

Examples

To better understand the main results presented in the previous section, we now consider some examples. In each example, we develop more explicit expressions for the bias and the variance of the estimate of β for both the PCA method and the mixed effects model.

Example 1

We first assume that the relatedness between individuals are the same for all pairs. That is, the matrix $\mathbf{S}$ is

$$\begin{aligned}\mathbf{S} &= \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} \\ &= (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}'\end{aligned}$$

where $\rho < 1$ measure the common relatedness between any two subjects. In addition, we assume that there is no population stratification. Hence $\mathbf{u} = u\mathbf{1}$ in (2) for a value u.

Since there is no population stratification, the variance matrix used for the mixed effects model is $\mathbf{\Omega}$ instead of $\mathbf{\Omega}^*$. Because $\mathbf{\Omega} = \sigma_G^2\mathbf{S} + \sigma_e^2\mathbf{I} = \theta^2\mathbf{I} + \rho\sigma_G^2\mathbf{1}\mathbf{1}'$, where $\theta = \sqrt{\sigma_G^2(1-\rho) + \sigma_e^2}$, there is

$$\mathbf{\Omega}^{-1} = \frac{1}{\theta^2}\left(\mathbf{I} - \frac{\rho\sigma_G^2}{\theta^2 + n\rho\sigma_G^2}\mathbf{1}\mathbf{1}'\right).$$

Since

$$\begin{aligned}
\mathbf{g}'\mathbf{\Omega}^{-1}\mathbf{g} &= \frac{1}{\theta^2}\left(\mathbf{g}'\mathbf{g} - \frac{\rho\sigma_G^2(\mathbf{1}'\mathbf{g})^2}{\theta^2 + n\rho\sigma_G^2}\right), \\
\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{g} &= \frac{1}{\theta^2}\left(1 - \frac{n\rho\sigma_G^2}{\theta^2 + n\rho\sigma_G^2}\right)(\mathbf{1}'\mathbf{g}) \\
&= \frac{\mathbf{1}'\mathbf{g}}{\theta^2 + n\rho\sigma_G^2},
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1} &= \frac{n}{\theta^2}\left(1 - \frac{n\rho\sigma_G^2}{\theta^2 + n\rho\sigma_G^2}\right) \\
&= \frac{n}{\theta^2 + n\rho\sigma_G^2},
\end{aligned}$$

we have

$$\begin{aligned}
[Var(\hat{\beta})]^{-1} &= \mathbf{g}\mathbf{P}_2\mathbf{g} \\
&= \mathbf{g}[\mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}\mathbf{1}(\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{\Omega}^{-1}]\mathbf{g} \\
&= \mathbf{g}'\mathbf{\Omega}^{-1}\mathbf{g} - \frac{(\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{g})^2}{\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1}} \\
&= \frac{\mathbf{g}'\mathbf{g}}{\theta^2} - \frac{\rho\sigma_G^2(\mathbf{1}'\mathbf{g})^2}{\theta^2(\theta^2 + n\rho\sigma_G^2)} - \frac{(\mathbf{1}'\mathbf{g})^2}{n(\theta^2 + n\rho\sigma_G^2)} \\
&= \frac{\mathbf{g}'\mathbf{g}}{\theta^2} - \frac{(\mathbf{1}'\mathbf{g})^2}{n\theta^2} \\
&= \frac{n\sigma_g^2}{\theta^2}
\end{aligned}$$

with

$$\sigma_g^2 = n^{-1}\mathbf{g}'\mathbf{g} - (n^{-1}\mathbf{1}'\mathbf{g})^2.$$

That is,

$$Var(\hat{\beta}) = \frac{\theta^2}{n\sigma_g^2}.$$

The variance of $\tilde{\beta}$ depends on the principal components included in the regression. The largest characteristic root of matrix $\mathbf{S}$ is $\lambda_1 = 1 + (n-1)\rho$ with associated characteristic

vector $\mathbf{x}_1 = (n^{-1/2}, n^{-1/2}, \ldots, n^{-1/2})'$. This eigenvector confounds with intercept term and needs to be excluded. All other eigenvalues are equal to $1 - \rho$. The corresponding eigenvectors are not unique. One set of eigenvectors is

$$\begin{aligned}
\mathbf{x}_2 &= \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, \ldots, 0\right)', \\
&\vdots \\
\mathbf{x}_i &= \left(\frac{1}{\sqrt{(i-1)i}}, \ldots, \frac{1}{\sqrt{(i-1)i}}, -\frac{i-1}{\sqrt{(i-1)i}}, 0, \ldots, 0\right)', \\
&\vdots \\
\mathbf{x}_n &= \left(\frac{1}{\sqrt{(n-1)n}}, \ldots, \frac{1}{\sqrt{(n-1)n}}, -\frac{n-1}{\sqrt{(n-1)n}}\right)'.
\end{aligned}$$

Switching any two elements in these vectors at the same time yields another set of eigenvectors.

Excluding $\mathbf{x}_1$, redefine $\mathbf{X}_m$ by $\mathbf{X}_m = (\mathbf{x}_2, \ldots, x_m)$, the matrix formed by the first $m-1$ eigenvectors. To compute $\mathbf{P}_1$, we note that

$$\begin{aligned}
\mathbf{Z}'\mathbf{Z} &= \begin{pmatrix} n & \mathbf{1}'\mathbf{X}_k \\ \mathbf{X}_k'\mathbf{1} & \mathbf{X}_k'\mathbf{X}_k \end{pmatrix} \\
&= \begin{pmatrix} n & 0 \\ 0 & \mathbf{I}_k \end{pmatrix}.
\end{aligned}$$

The inverse of $\mathbf{Z}'\mathbf{Z}$ is

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{pmatrix} 1/n & 0 \\ 0 & \mathbf{I}_m \end{pmatrix}.$$

Therefore, $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = n^{-1}\mathbf{1}\mathbf{1}' + \mathbf{X}_m\mathbf{X}_m'$ and

$$\begin{aligned}
\mathbf{P}_1 &= \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \\
&= (\mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}') - \mathbf{X}_m\mathbf{X}_m'.
\end{aligned}$$

Since $\mathbf{P}_1$ is idempotent and $\mathbf{1}'\mathbf{P}_1 = \mathbf{0}$, we have $\mathbf{P}_1\mathbf{\Omega}\mathbf{P}_1 = \theta^2\mathbf{P}_1$ and

$$\begin{aligned} Var(\tilde{\beta}) &= \frac{\mathbf{g}'\mathbf{P}_1\mathbf{\Omega}\mathbf{P}_1\mathbf{g}}{(\mathbf{g}'\mathbf{P}_1\mathbf{g})^2} \\ &= \frac{\theta^2}{\mathbf{g}'\mathbf{P}_1\mathbf{g}} \\ &= \frac{\theta^2}{\mathbf{g}'\mathbf{g} - n^{-1}(\mathbf{1}'\mathbf{g})^2 - \mathbf{g}'\mathbf{X}_m\mathbf{X}_m'\mathbf{g}} \\ &= \frac{\theta^2}{n\sigma_g^2 - \mathbf{g}'\mathbf{X}_m\mathbf{X}_m'\mathbf{g}} \\ &= \frac{\theta^2}{n\sigma_g^2 - [(\mathbf{g}'\mathbf{x}_2)^2 + \ldots + (\mathbf{g}'\mathbf{x}_m)^2]}, \end{aligned}$$

where $\sigma_g^2 = n^{-1}\mathbf{g}'\mathbf{g} - (n^{-1}\mathbf{1}'\mathbf{g})^2$ is the variance of the components of vector $\mathbf{g}$. One can see from this relationship that adding more PCs increases the variance of $\tilde{\beta}$ unless the added PCs are orthogonal to $\mathbf{g}$ (i.e., each of the added ones satisfies $\mathbf{g}'\mathbf{x}_m = 0$). In addition, for the same number of PCs used (i.e., the value of m is fixed), different choices of PCs could result in a difference in efficiency. The ratio $Var(\tilde{\beta})/Var(\hat{\beta})$ is

$$\frac{n\sigma_g^2}{n\sigma_g^2 - \mathbf{g}'\mathbf{X}_m\mathbf{X}_m'\mathbf{g}}$$

which is 1 if and only if $\mathbf{X}_m'\mathbf{g} = \mathbf{0}$. This happens if no principal component is used. We note that this ratio does not depend on ρ.

Example 2

Next we assume that there are K populations. There are n_k subjects in population k. The vector $\mathbf{u}$ in (2) is $\mathbf{u} = (u_1\mathbf{1}_1', \ldots, u_K\mathbf{1}_K')'$. The variance of the elements in vector $\mathbf{u}$ is denoted by σ_u^2. That is, $\sigma_u^2 = n^{-1}\mathbf{u}'\mathbf{u} - (n^{-1}\mathbf{1}'\mathbf{u})^2$. Assuming that there is no cryptic relatedness, we have $\mathbf{\Omega} = \sigma_e^2\mathbf{I}$ and $\mathbf{\Omega}^* = \sigma_u^2\tilde{\mathbf{S}} + \mathbf{\Omega}$ with $\tilde{\mathbf{S}} = \text{diag}(\mathbf{1}_1\mathbf{1}_1', \ldots, \mathbf{1}_K\mathbf{1}_K')$, the relatedness matrix induced by population stratification. The matrix $\mathbf{\Omega}^*$ is block diagonal: $\mathbf{\Omega}^* = \text{diag}(\mathbf{\Omega}_1^*, \ldots, \mathbf{\Omega}_k^*, \ldots, \mathbf{\Omega}_K^*)$ where $\mathbf{\Omega}_k^* = \sigma_e^2\mathbf{I}_k + \sigma_u^2\mathbf{1}_k\mathbf{1}_k'$. The largest eigenvalue for $\mathbf{1}_k\mathbf{1}_k'$ is n_k with corresponding eigenvector $\mathbf{1}_k$, $k = 1, \ldots, K$. All other eigenvalues are equal to 0.

Each eigenvalue λ for matrix $\tilde{\mathbf{S}}$ is a root to the equation $|\tilde{\mathbf{S}} - \lambda\mathbf{I}| = 0$. This implies

$$\begin{aligned} |\mathrm{diag}(\mathbf{1}_1\mathbf{1}_1', \ldots, \mathbf{1}_k\mathbf{1}_k', \ldots, \mathbf{1}_K\mathbf{1}_K') - \lambda\mathbf{I}| &= \prod_{k=1}^{K} |\mathbf{1}_k\mathbf{1}_k' - \lambda\mathbf{I}_k| \\ &= 0. \end{aligned}$$

That is, the eigenvalues of $\tilde{\mathbf{S}}$ consist of those of $\mathbf{1}_k\mathbf{1}_k', k = 1, \ldots, K$. So the only non-zero eigenvalues are $n_k, k = 1, \ldots, K$. It is straightforward to verify that the eigenvector corresponding to n_1 is $(\mathbf{1}_1', \mathbf{0}_{(-1)}')'$, where the subscript (-1) implies the part of the vector excluding the first n_1 elements. Eigenvectors corresponding to eigenvalues $n_k, k = 2, \ldots, K$ can be similarly constructed. If $n_1 = n_2 = \cdots = n_K$, the vector $\mathbf{1} = (\mathbf{1}_1', \mathbf{1}_2', \ldots, \mathbf{1}_K')'$ is also associated with the largest eigenvalue of $\tilde{\mathbf{S}}$ which is the common value n_1. However, it confounds with the intercept term and can not be used in the PCA method.

Assume that $n_1 > n_k, k = 2, \ldots, K$. The eigenvector associated with the largest eigenvalue of $\tilde{\mathbf{S}}$ is $\mathbf{x}_1 = (\mathbf{1}_1', \mathbf{0}_{(-1)}')'$. Let's use only this eigenvector in the principal components method (i.e. $m = 1$). The matrix $\mathbf{P}_1$ turns out to be

$$\mathbf{P}_1 = \mathbf{I} - \begin{pmatrix} n_1^{-1}\mathbf{1}_1\mathbf{1}_1' & \mathbf{0}_1\mathbf{0}_{(-1)}' \\ \mathbf{0}_{(-1)}'\mathbf{0}_1 & (n - n_1)^{-1}\mathbf{1}_{(-1)}\mathbf{1}_{(-1)}' \end{pmatrix}.$$

Furthermore, it can be shown that

$$\begin{aligned} \mathbf{g}'\mathbf{P}_1\mathbf{u} &= (n - n_1)Cov(\mathbf{u}_{(-1)}, \mathbf{g}_{(-1)}) \\ \mathbf{g}'\mathbf{P}_1\mathbf{g} &= n_1\sigma_{g_1}^2 + (n - n_1)\sigma_{g_{(-1)}}^2 \end{aligned}$$

where

$$Cov(\mathbf{u}_{(-1)}, \mathbf{g}_{(-1)}) = (n - n_1)^{-1} \sum_{i \neq 1} n_i u_i \bar{g}_i - (n - n_1)^{-2} \left(\sum_{i \neq 1} n_i u_i \right) \cdot \left(\sum_{i \neq 1} n_i \bar{g}_i \right)$$

is the covariance between $\mathbf{u}_{(-1)}$ and $\mathbf{g}_{(-1)}$ and

$$\sigma_{g_{(-1)}}^2 = (n - n_1)^{-1}\mathbf{g}_{(-1)}'\mathbf{g}_{(-1)} + (n - n_1)^{-2}(\mathbf{g}_{(-1)}'\mathbf{1}_{(-1)})^2.$$

is the variance of $\mathbf{g}_{(-1)}$. So the bias of $\tilde{\beta}$ is

$$\frac{\mathbf{g}'\mathbf{P}_1\mathbf{u}}{\mathbf{g}'\mathbf{P}_1\mathbf{g}} = \frac{(n-n_1)Cov(\mathbf{u}_{(-1)}, \mathbf{g}_{(-1)})}{n_1\sigma^2_{g_1} + (n-n_1)\sigma^2_{g_{(-1)}}}.$$

We note that this bias does not depend on σ^2_e. If there are only two populations (i.e., $K = 2$), $\mathbf{u}_{(-1)} = u_2\mathbf{1}_2$ is a vector of constants and $Cov(\mathbf{u}_{(-1)}, \mathbf{g}_{(-1)}) = 0$. That is, $\tilde{\beta}$ is unbiased. Another explanation of this fact is that $\mathbf{u}$ is a linear combination of $\mathbf{1}$ and $\mathbf{x}_1$, two vectors contained in $\mathbf{Z}$. If there are more than 2 populations, this conclusion is no longer true.

Since $\mathbf{\Omega} = \sigma^2_e\mathbf{I}$, there is $\mathbf{P}_1\mathbf{\Omega}\mathbf{P}_1 = \sigma^2_e\mathbf{P}_1$. So the variance of $\tilde{\beta}$ is

$$\begin{aligned}\frac{\mathbf{g}'\mathbf{P}_1\mathbf{\Omega}\mathbf{P}_1\mathbf{g}}{(\mathbf{g}'\mathbf{P}_1\mathbf{g})^2} &= \frac{\sigma^2_e}{\mathbf{g}'\mathbf{P}_1\mathbf{g}} \\ &= \frac{\sigma^2_e}{n_1\sigma^2_{g_1} + (n-n_1)\sigma^2_{g_{(-1)}}}.\end{aligned}$$

Note that this variance does not depend on σ^2_u.

We now investigate the bias of $\hat{\beta}$. To this end, we note first that the inverse $\mathbf{\Omega}^{-1}$ is also block-diagonal with the kth block equal to

$$\mathbf{\Omega}_k^{-1} = \frac{1}{\sigma^2_e}\left[\mathbf{I}_k - \frac{\sigma^2_u\alpha_k}{n_k}\mathbf{1}_k\mathbf{1}'_k\right], k = 1, \ldots, K,$$

where $\alpha_k = 1/(\sigma^2_e/n_k + \sigma^2_u)$. To compute the bias of $\hat{\beta}$, which is $(\mathbf{g}'\mathbf{P}_2\mathbf{g})^{-1}(\mathbf{g}'\mathbf{P}_2\mathbf{u})$, we note

that $\mathbf{\Omega}_k^{-1}\mathbf{1}_k = n_k^{-1}\alpha_k\mathbf{1}_k$, $\mathbf{1}_k'\mathbf{\Omega}_k^{-1}\mathbf{1}_k = \alpha_k$. For simplicity, assume that $K = 2$. We have

$$\begin{aligned}
\mathbf{P}_2\mathbf{u} &= \left(\mathbf{\Omega}^{-1} - \frac{\mathbf{\Omega}^{-1}\mathbf{1}\mathbf{1}'\mathbf{\Omega}^{-1}}{\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1}}\right)\cdot\begin{pmatrix} u_1\mathbf{1}_1 \\ u_2\mathbf{1}_2 \end{pmatrix} \\
&= \begin{pmatrix} u_1\mathbf{\Omega}_1^{-1}\mathbf{1}_1 \\ u_2\mathbf{\Omega}_2^{-1}\mathbf{1}_2 \end{pmatrix} - \frac{u_1\mathbf{1}_1'\mathbf{\Omega}_1^{-1}\mathbf{1}_1 + u_2\mathbf{1}_2'\mathbf{\Omega}_2^{-1}\mathbf{1}_2}{\mathbf{1}_1'\mathbf{\Omega}_1^{-1}\mathbf{1}_1 + \mathbf{1}_2'\mathbf{\Omega}_2^{-1}\mathbf{1}_2}\begin{pmatrix} \mathbf{\Omega}_1^{-1}\mathbf{1}_1 \\ \mathbf{\Omega}_2^{-1}\mathbf{1}_2 \end{pmatrix} \\
&= \frac{u_1 - u_2}{\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1}}\begin{pmatrix} (\mathbf{1}_2'\mathbf{\Omega}_2^{-1}\mathbf{1}_2)\mathbf{\Omega}_1^{-1}\mathbf{1}_1 \\ -(\mathbf{1}_1'\mathbf{\Omega}_1^{-1}\mathbf{1}_1)\mathbf{\Omega}_2^{-1}\mathbf{1}_2 \end{pmatrix} \\
&= \frac{\alpha_1\alpha_2(u_1 - u_2)}{\alpha_1 + \alpha_2}\begin{pmatrix} n_1^{-1}\mathbf{1}_1 \\ -n_2^{-1}\mathbf{1}_2 \end{pmatrix} \\
\mathbf{g}'\mathbf{P}_2\mathbf{u} &= \frac{\alpha_1\alpha_2(u_1 - u_2)(\bar{g}_1 - \bar{g}_2)}{\alpha_1 + \alpha_2}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbf{g}_k'\mathbf{g}_k &= n_k(\sigma_{g_k}^2 + \bar{g}_k^2) \\
\mathbf{g}_k'\mathbf{\Omega}_k^{-1}\mathbf{g}_k &= \frac{\mathbf{g}_k'\mathbf{g}_k}{\sigma_e^2} - \frac{\sigma_u^2 n_k\alpha_k\bar{g}_k^2}{\sigma_e^2} \\
&= \frac{n_k\sigma_{g_k}^2}{\sigma_e^2} + \alpha_k\bar{g}_k^2 \\
\mathbf{g}_k'\mathbf{\Omega}_k^{-1}\mathbf{1}_k &= \alpha_k\bar{g}_k \\
\mathbf{g}'\mathbf{P}_2\mathbf{g} &= \mathbf{g}'\mathbf{\Omega}^{-1}\mathbf{g} - \frac{(\mathbf{g}'\mathbf{\Omega}^{-1}\mathbf{1})^2}{\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1}} \\
&= \frac{n_1\sigma_{g_1}^2 + n_2\sigma_{g_2}^2}{\sigma_e^2} + \alpha_1\bar{g}_1^2 + \alpha_2\bar{g}_2^2 - \frac{(\mathbf{g}'\mathbf{\Omega}^{-1}\mathbf{1})^2}{\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1}} \\
&= \frac{n_1\sigma_{g_1}^2 + n_2\sigma_{g_2}^2}{\sigma_e^2} + \frac{\alpha_1\alpha_2(\bar{g}_1 - \bar{g}_2)^2}{\mathbf{1}'\mathbf{\Omega}^{-1}\mathbf{1}} \\
&= \frac{n_1\sigma_{g_1}^2 + n_2\sigma_{g_2}^2}{\sigma_e^2} + \frac{\alpha_1\alpha_2(\bar{g}_1 - \bar{g}_2)^2}{\alpha_1 + \alpha_2}.
\end{aligned}$$

So the bias is

$$\frac{\mathbf{g}'\mathbf{P}_2\mathbf{u}}{\mathbf{g}'\mathbf{P}_2\mathbf{g}} = \frac{(u_1 - u_2)(\bar{g}_1 - \bar{g}_2)}{\frac{n_1\sigma_{g_1}^2 + n_2\sigma_{g_2}^2}{\sigma_e^2}(1/\alpha_1 + 1/\alpha_2) + (\bar{g}_1 - \bar{g}_2)^2}.$$

Generally, if there are more that two populations (i.e., $K > 2$), it can be shown that

$$\begin{aligned} \mathbf{g}'\mathbf{P}_2\mathbf{u} &= \frac{\sum_k n_k \bar{g}_k \sum_{l \neq k} n_l (u_k - u_l)}{(\sum_k \alpha_k)(\prod_i n_k/\alpha_k)} \\ &= \frac{\sum_{k<l} n_k n_l (u_k - u_l)(\bar{g}_k - \bar{g}_l)}{(\sum_k \alpha_k)(\prod_k n_k/\alpha_k)} \end{aligned}$$

and

$$\begin{aligned} \mathbf{g}'\mathbf{P}_2\mathbf{g} &= \sum_k [n_k \sigma^2_{g_k}/\sigma^2_e + \alpha_k \bar{g}^2_k] - \frac{(\sum_k \alpha_k \bar{g}_k)^2}{\sum_k \alpha_k} \\ &= \frac{1}{\sigma^2_e} \sum_k n_k \sigma^2_{g_k} + \frac{\sum_{k<l} \alpha_k \alpha_l (\bar{g}_k - \bar{g}_l)^2}{\sum_k \alpha_k} \end{aligned}$$

from which one would be able to compute the bias $(\mathbf{g}'\mathbf{P}_2\mathbf{u})/(\mathbf{g}'\mathbf{P}_2\mathbf{g})$. Both the bias and the variance of $\hat{\beta}$ depend on σ^2_e.

Example 3

The magnitude of pair-wise relatedness in a population does not need to be constant. For instance, if there are two nuclear families from a single population, the relatedness matrix may be modeled as a block diagonal matrix consisting of two blocks, one block for each nuclear family. On the other hand, such a block diagonal matrix can also be explained as a relatedness matrix for individuals from different populations (Example 2). Different interpretation implies different estimates of the genetic effect size and the variance of these estimates. The technical details are omitted since the computation is similar as those in Example 1 and Example 2.

Simulation studies

The data-generating model used in the simulation studies is (2), which is

$$\mathbf{y} = \mathbf{u} + \beta\mathbf{g} + \boldsymbol{\epsilon}.$$

The value of σ_e^2 is fixed at 1. Each component of $\boldsymbol{\epsilon}$ are independently and identically normally distributed with mean 0 and variance σ_e^2. The number of simulation replications in an experiment is 1000. The matrix $\boldsymbol{\Omega}$, vector $\mathbf{u}$, and the genotype score vector $\mathbf{g}$ are fixed. In each replication, $\boldsymbol{\epsilon}$ is randomly generated.

The first simulation study corresponds to Example 1 with $\mathbf{u} = \mathbf{1}$. There are 500 subjects in a sample. The genotype score is 0 for the first 50, 1 the next 150, and 2 the last 300. The relatedness between any pair of subjects is $\rho = 0.5$. The polygenic variance is $\sigma_G^2 = 1$. The first two eigenvectors from matrix $\mathbf{S}$ are used. The first one is $\mathbf{x}_1 = \mathbf{1}$. The choice of the second eigenvector has great impact on the variance of $\tilde{\beta}$. To illustrate this point, we choose $\mathbf{x}_2$ to be proportional to a vector whose first 200 elements are equal to 1, the next 200 are equal to -1, and the last 100 are equal to 0. The ratio $Var(\tilde{\beta})/Var(\hat{\beta})$ is equal to $n\sigma_g^2/(n\sigma_g^2 - (\mathbf{g}'\mathbf{x}_2)^2) = n\sigma_g^2/(n\sigma_g^2 - (-12.5)^2) = 3.252$. In comparison, if $\mathbf{x}_2$ is chosen to be a vector whose first element is $1/\sqrt{2}$, last element $-1/\sqrt{2}$, and all other elements 0, this ratio would be merely 1.009. Figure 1 presents the simulation result with $\beta = 0$. It shows that $\tilde{\beta}$ has larger variance and the distributions of the squared t-statistic from both methods conform with the 1-df chi-square distribution. To investigate the power loss due to the use of principal components, we also consider the case where $\beta = 0.1$ (Figure 2). $Var(\tilde{\beta})$ continue to have larger variance. The mixed effects model is more powerful as indicated by its tendency of having larger (squared) t-statistic.

[Figure 1 about here.]

[Figure 2 about here.]

In the second simulation study, it is assumed that data are from 4 populations. Accordingly, the vector $\mathbf{u}$ has four segments. The elements of $\mathbf{u}$ are equal to 1 for segment 1, 2 for segment 2, 3 for segment 3, and 4 for segment 4. Hence the matrix $\mathbf{\Omega}$ is $\sigma_u^2 \cdot \mathbf{S} + \mathbf{I}$ where $\mathbf{S} = \text{diag}(\mathbf{1}_1\mathbf{1}_1', \mathbf{1}_2\mathbf{1}_2', \mathbf{1}_3\mathbf{1}_3', \mathbf{1}_4\mathbf{1}_4')$. Sample size and Genotype counts for each population is shown in Table 1. We first investigate the validity of the PCA method and the mixed effects model method when there is genetic effect does not exist (i.e., $\beta = 0$). We first used the first two principal components of $\mathbf{S}$ in the PCA method (Figure 3) and then the first four principal components (Figure 4). The first two principal components are indicator vectors for population 1 and population 4, the populations have the largest number of subjects. The first four principal components are indicator vectors for each of the four populations. Due to their collinearity with the regression intercept, using four of them is equivalent to using any three of them.

[Table 1 about here.]

The PCA method in Figure 3 is clearly invalid. Since the rank of $\mathbf{S}$ is 4, there are still some population structure unexplained by the first two principal components. The PCA method treats the residuals as iid while they are actually not. Bias in $\tilde{\beta}$ and inflation of the squared t-statistic are expected (Figure 3). However, when 4 principal components are used, $\tilde{\beta}$ is no longer biased (Figure 4). In Figure 3 and Figure 4, there is no change to the mixed effects model method.

[Figure 3 about here.]

[Figure 4 about here.]

Discussion

Large scale genetic association studies is a popular means for identifying genetic factors underlying complex human traits. Cryptic relatedness and population stratification are deemed to be issues unavoidable in such studies. In this work, we have conducted an analytical comparison of two popular methods, i.e., the PCA method and the mixed effects model, that are supposed to address these issues. We focused not only on the efficiency of each method, but also on their validity through investigation of the bias and the variance of genetic effect estimates. The findings are enlightening.

If there exists cryptic relatedness but there does not population stratification, the mixed effects model is preferred as it has overall better performance. When there is no association, both methods are valid. However, if there is association, the mixed effects model is more powerful (Proposition 1). If there exists population stratification, both methods can lead to biased estimates of the genetic effect (Proposition 2). In other words, none of them can eliminate the confounding effect of population stratification. However, simulation studies suggest that the mixed effects model is much less affected by population stratification (Figure 3) and is almost as good as the PCA method in which the population stratification is completely eliminated (Figure 4). These findings are consistent to simulation studies reported previously[23,29].

The eigenvectors are useful in graphically representing population stratification. But the PCA method is generally not better than the mixed effects model in controlling for the confounding effect of population stratification for reasons discussed previously. In addition, our research has also revealed some issues not discussed before in using the PCA method. Previous research has focused on the number of principal components to use. The issue of what specific components to use appears to be ignored. As indicated by our examples, it is not uncommon in genetic studies that eigenvalues are repeated. Our research indicates

that different choices of the eigenvalues have an impact on the bias of the genetic effect estimate and efficiency of association testing. This impact can be larger than the impact of the choice of the number of eigenvectors (Example 1).

REFERENCES

1.Catarina D Campbell, Elizabeth L Ogburn, Kathryn L Lunetta, Helen N Lyon, Matthew L Freedman, Leif C Groop, David Altshuler, Kristin G Ardlie, and Joel N Hirschhorn. Demonstrating stratification in a European American population. *Nature Genetics*, 37:868–872, 2005.

2.Shuhua Xu, Xianyong Yin, Shilin Li, Wenfei Jin, Haiyi Lou, Ling Yang, Xiaohong Gong, Hongyan Wang, Yiping Shen, Xuedong Pan, Yungang He, Yajun Yang, Yi Wang, Wenqing Fu, Yu An, Jiucun Wang, Jingze Tan, Ji Qian, Xiaoli Chen, Xin Zhang, Yangfei Sun, Xuejun Zhang, Bailin Wu, and Li Jin. Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet*, 85(6):762–774, 2009.

3.Jieming Chen, Houfeng Zheng, Jin-Xin Bei, Liangdan Sun, Wei-Hua Jia, Tao Li, Furen Zhang, Mark Seielstad, Yi-Xin Zeng, Xuejun Zhang, and Jianjun Liu. Genetic structure of the Han Chinese population revealed by genome-wide snp variation. *Am J Hum Genet*, 85(6):775 – 785, 2009.

4.M. C. Cornelis, A. Agrawal, J. W. Cole, N. H. Hansel, K. C. Barnes, T. H. Beaty, S. N. Bennett, L. J. Bierut, E. Boerwinkle, K. F. Doheny, B. Feenstra, E. Feingold, M. Fornage, C. A. Haiman, E. L. Harris, M. G. Hayes, J. A. Heit, F. B. Hu, J. H. Kang, C. C. Laurie, H. Ling, T. A. Manolio, M. L. Marazita, R. A. Matthias, D. B. Mirel, J Paschall, L. R. Pasquale, E. W. Pugh, J. P. Rice, J Udren, R. M. van Dam, X Wang, J. L. Wiggs, K Williams, K. Yu, and for the GENEVA Consortium. The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet Epidemiol*, 34(4):364–372, 2010.

5.B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

6.S. Bacanu, B. Devlin, and K. Roeder. The power of genomic control. *Am J Hum Genet*, 66:1933–1944, 2000.

7.B. Devlin, K. Roeder, and L. Wasserman. Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, 60:155–166, 2001.

8.S. Bacanu, B. Devlin, and K. Roeder. Association studies for quantitative traits in structured populations. *Genet Epidemiol*, 22:78–93, 2002.

9.K. Wang. Testing for genetic association in the presence of population stratification in genome-wide association studies. *Genet Epidemiol*, 33:637–645, PMID: 19235185 2009.

10.J. K. Pritchard and N. A. Rosenberg. Use of unlinked genetic markers to detect population stratication in association studies. *Am J Hum Genet*, 65:220–228, 1999.

11.J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *Am J Hum Genet*, 67:170–181, 2000.

12.J. K. Pritchard and P. Donnelly. Case-control studies of association in structured or admixed populatoins. *Theoretical Population Biology*, 60:227–237, 2001.

13.Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratication in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

14.M. P. Epstein, A. S. Allen, and G. A. Satten. A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet*, 80:921–930, 2007.

15.David J. Balding and Richard A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3–12, 1995.

16.Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569, 2010.

17.Joseph E. Powell, Peter M. Visscher, and Michael E. Goddard. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, 11:800–805, 2010.

18.Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet*, 88:76–82, 2011.

19.Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, Stephen Kresovich, and Edward S Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2006.

20.Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-Yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

21.Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas,

and Edward S Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360, 2010.

22.Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–1723, 2008.

23.Alkes L. Price, Noah A. Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11:459–463, 2010.

24.J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

25.Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

26.D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

27.T. Thornton and M. S. McPeek. ROADTRIPS: case–control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet*, 86:172–184, 2010.

28.W Astle and DJ Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, in press, 2010.

29.Chengqing Wu, Andrew DeWan, Josephine Hoh, and Zuoheng Wang. A comparison of association methods correcting for population stratification in case–control studies. *Annals of Human Genetics*, 75(3):418–427, 2011.

30.Matthieu Bouaziz, Christophe Ambroise, and Mickael Guedj. Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide association studies. *PLoS One*, 6(12):e28845, 2011.

31.Nianjun Liu, Hongyu Zhao, Amit Patki, Nita A. Limdi, and David B. Allison. Controlling population structure in human genetic association studies with samples of unrelated individuals. *Stat Interface*, 4(3):317–326, 2011.

32.Kenneth Lange. *Mathematical and statistical methods for genetic analysis*. Springer, second edition, 2002.

33.M. Lynch and K. Ritland. Estimation of pairwise relatedness with molecular markers. *Genetics*, 152:1753–1766, 1999.

34.M. P. Epstein, W. L. Duren, and M. Boehnke. Improved inference of relationship for pairs of individuals. *Am J Hum Genet*, 67:1219–1231, 2000.

35.S. C. Thomas and W. G. Hill. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, 155:1961–1972, 2000.

36.K. Ritland. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res*, 67:175–185, 2009.

37.M. S. McPeek and L. Sun. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet*, 66:1076–1094, 2000.

38.B. G. Milligan. Maximum-likelihood estimation of relatedness. *Genetics*, 163:1153–1167, 2003.

39.Cyril S. Rakovski and Daniel O. Stram. A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors. *PLoS One*, 4(6):e5825, 2009.

40.Jianzhong Ma and Christopher I. Amos. Theoretical formulation of principal components analysis to detect and correct for population stratification. *PLoS One*, 5(9):e12510, 2010.

This manuscript was prepared with the AAS LaTeX macros v5.2.

Fig. 1.— Simulation result for the case of one population. The first two principal components are used for the PCA method. The genetic effect is $\beta = 0$.

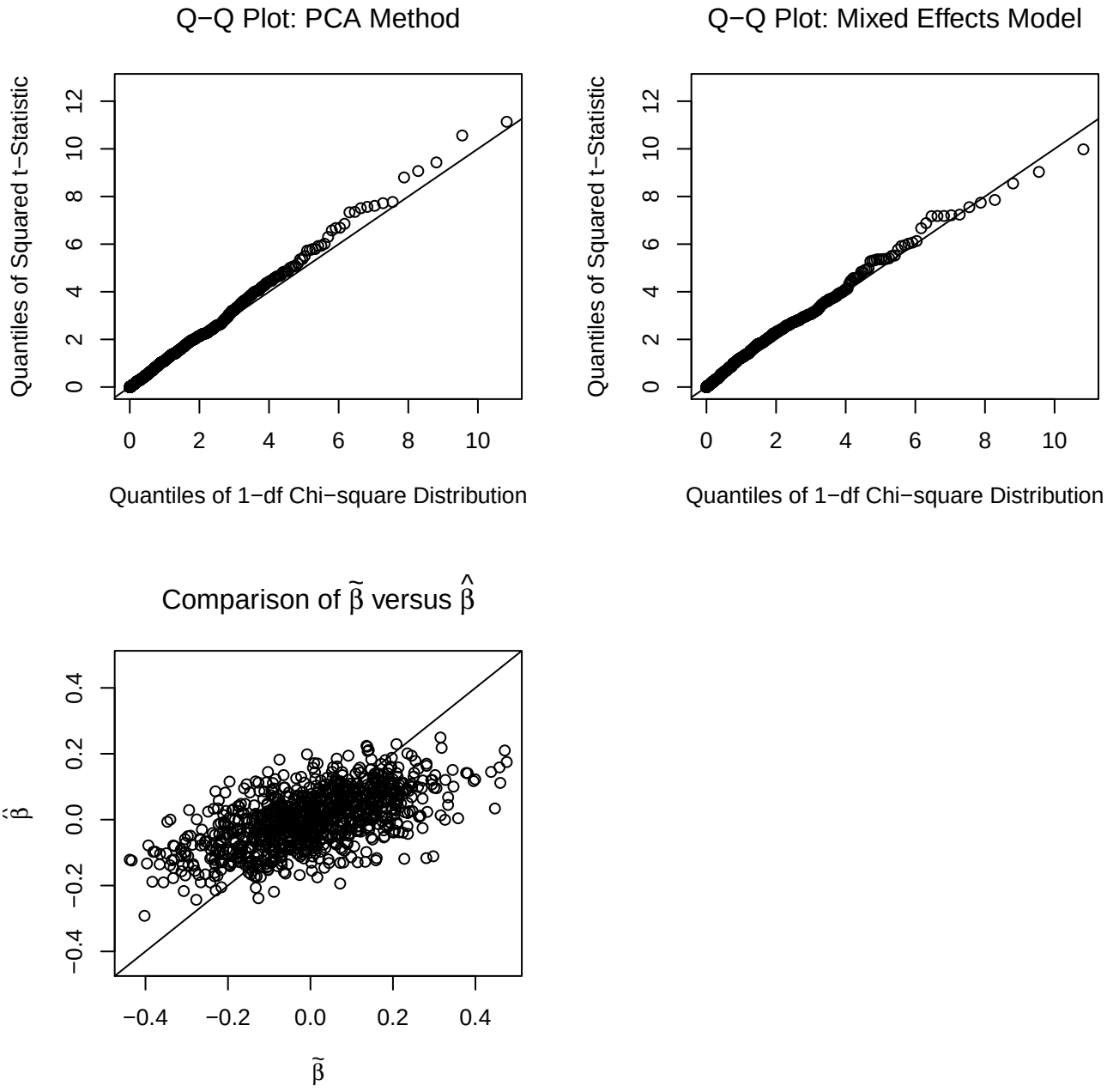

Fig. 2.— Simulation result for the case of one population. The first two principal components are used for the PCA method. The genetic effect is $\beta = 0.1$.

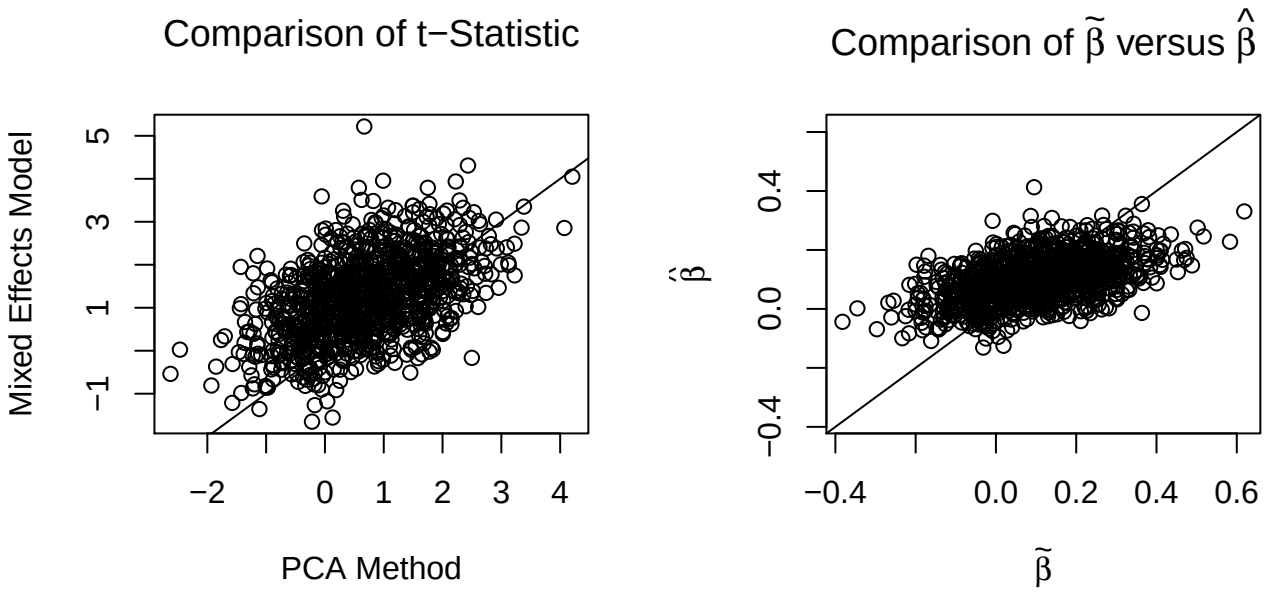

Fig. 3.— Simulation result for the case of four populations. The first two principal components are used in the PCA method. Genetic effect size is $\beta = 0$.

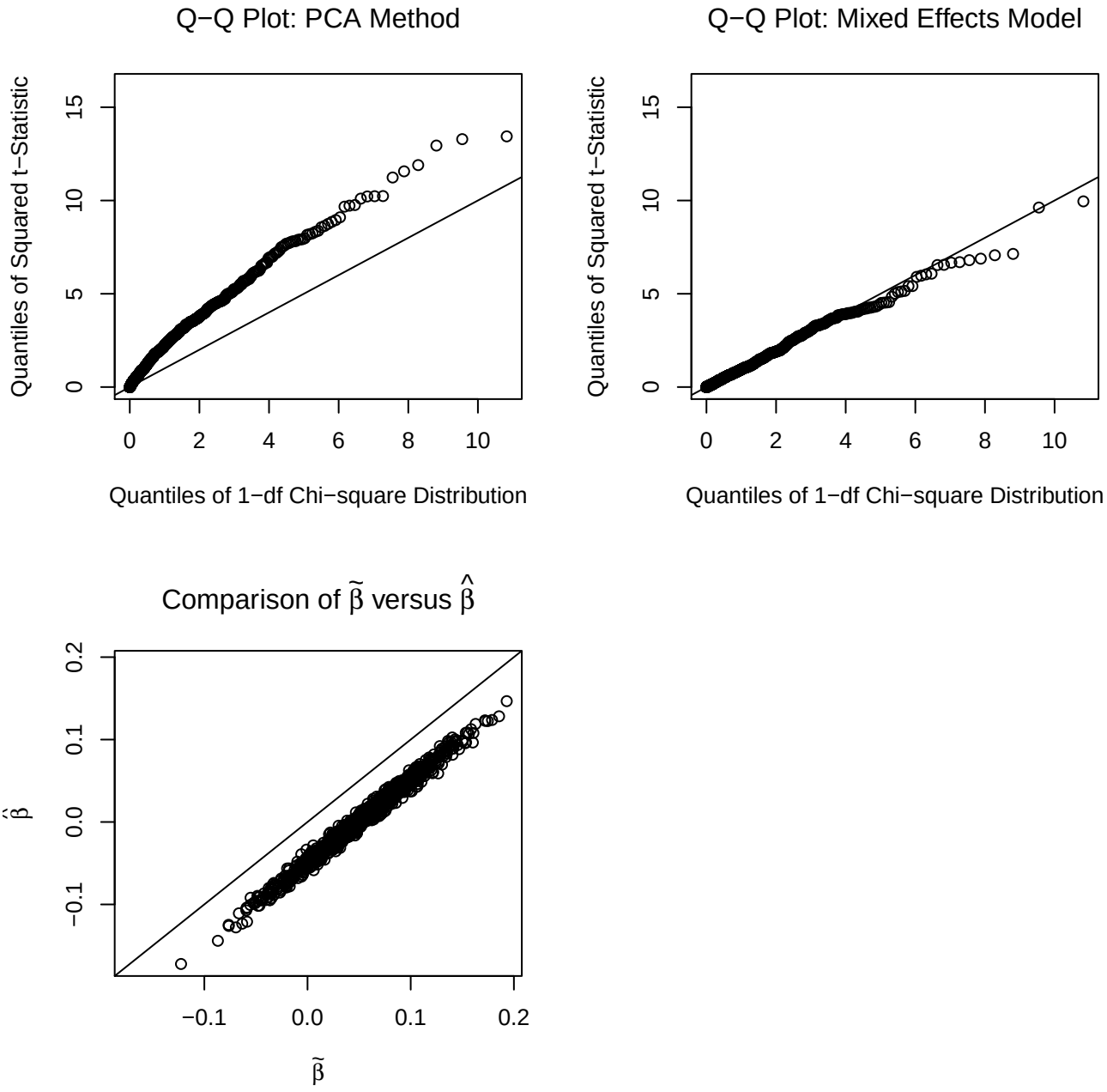

Fig. 4.— Simulation result for the case of four populations. The first four principal components are used in the PCA method. Genetic effect size is $\beta = 0$.

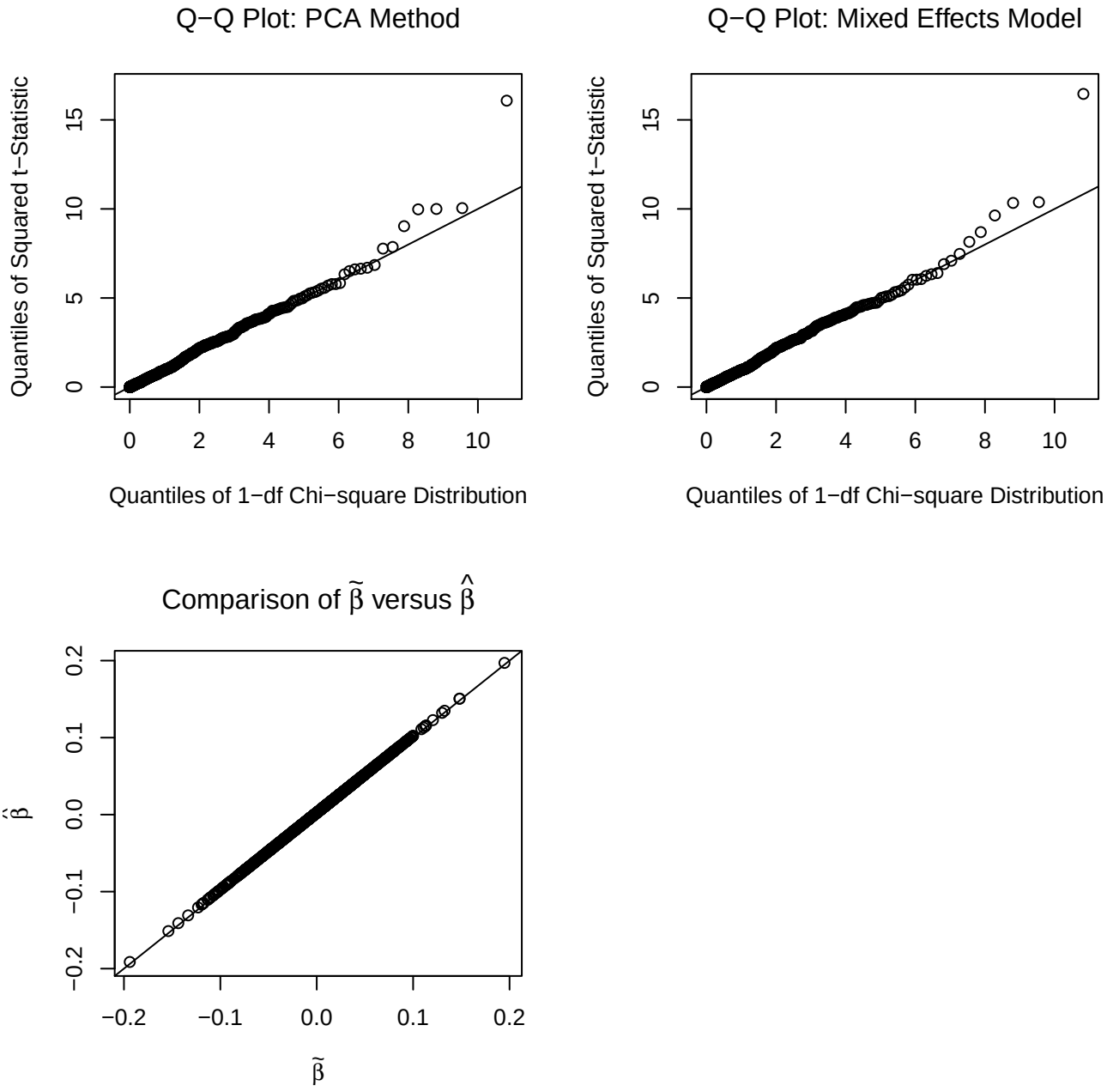

Table 1: Genotype counts in each population for the simulation study involving four populations.

	Genotype Score			
Population	0	1	2	Total
1	240	30	30	$n_1 = 300$
2	20	160	20	$n_2 = 200$
3	20	70	60	$n_3 = 150$
4	70	70	160	$n_4 = 300$