DOI: 10.1111/j.1541-0420.2010.01520.x

Control of Population Stratification by Correlation-Selected Principal Components

Seunggeun Lee,* Fred A. Wright, and Fei Zou*****

Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
*email: slee@bios.unc.edu
**email: fwright@bios.unc.edu
***email: fzou@bios.unc.edu

Summary. In genome-wide association studies, population stratification is recognized as producing inflated type I error due to the inflation of test statistics. Principal component-based methods applied to genotypes provide information about population structure, and have been widely used to control for stratification. Here we explore the precise relationship between genotype principal components and inflation of association test statistics, thereby drawing a connection between principal component-based stratification control and the alternative approach of genomic control. Our results provide an inherent justification for the use of principal components, but call into question the popular practice of selecting principal components based on significance of eigenvalues alone. We propose a new approach, called EigenCorr, which selects principal components based on both their eigenvalues and their correlation with the (disease) phenotype. Our approach tends to select fewer principal components for stratification control than does testing of eigenvalues alone, providing substantial computational savings and improvements in power. Analyses of simulated and real data demonstrate the usefulness of the proposed approach.

Key words: Genomic control; GWAS; PCA; Population stratification.

1. Introduction

In tests of genetic association among unrelated individuals, it is recognized that population stratification can result in test statistics with inflated apparent significance, resulting in overall type I error that is far above the nominal level. The method of *genomic control* (Devlin and Roeder, 1999; Devlin, Roeder, and Wasserman, 2001) was among the first attempts to address this problem, and is straightforward. For (chi-square) statistics at numerous markers measuring association with phenotype, an estimate is obtained for the inflation of test statistics beyond that expected under the null hypothesis and assuming no stratification. Then the test statistics are all adjusted by the inflation factor. However, a typical genome-wide association scan (GWAS) tests a very large number of single nucleotide polymorphism (SNP) markers, requiring a stringent significance threshold. In this setting, genomic control can fail to properly control the type I error (Devlin, Bacanu, and Roeder, 2004; Marchini et al., 2004), in part because variance inflation is not constant across the SNPs.

Alternatively, the principal component (PC) approach (Price et al., 2006) uses PCs computed from all genotypes as covariates in phenotype–genotype regression or in stratified analyses. Results from numerous studies indicate that the PC values often reflect known substructure and ancestry (Price et al., 2008). A great advantage of the PC approach is its sensitivity, effectively adjusting test statistics for only those markers contributing to the stratification. However, several challenges remain for effective PC-based stratification control. The primary challenge lies in choosing which PCs to include as covariates. Clearly not all PCs can be included, as there are as many PCs as there are individuals under study. Price et al. (2006) originally suggested using the 10 PCs with the highest eigenvalues. These investigators later proposed using the Tracy–Widom (TW) statistic (Patterson, Price, and Reich, 2006) to assess significance of eigenvalues to select PCs. However, this approach may detect a very large number of PCs as significant, with uncertain impact on the analysis. Moreover, the precise contribution of each PC to the overall type I error has not been established. As we shall see below, it is entirely possible for a relatively low-ranked PC to have a greater impact on type I error than a higher-ranked PC.

The article is arranged as follows. In Section 2, we establish a relationship between PCs and the average of the test statistics. Based on this relationship, we propose a new method, EigenCorr, to select PCs based on their eigenvalues and their correlations with the phenotype. The explicit use of phenotypes in stratification control has been anticipated in previous work, e.g., Epstein, Allen, and Satten (2007) and Kimmel et al. (2007). However, EigenCorr provides a direct connection to the type I error inflation introduced by PCs. A straightforward generalization applies when only a subset of markers are used for stratification control. In Section 3, we demonstrate the usefulness of EigenCorr via simulation and real GWAS analysis. In Section 4, we conclude with a discussion of implications and future directions.

2. Materials and Methods

Let g_{ij} be the genotype of SNP i and individual j, where $i = 1, \ldots, M$ and $j = 1, \ldots, N$. We define a normalized genotype x_{ij} as $x_{ij} = (g_{ij} - \bar{g}_{i.})/\sqrt{\sum_{j=1}^{N}(g_{ij} - \bar{g}_{i.})^2/N}$, where $\bar{g}_{i.} = \sum_{j=1}^{N} g_{ij}/N$. Let $\mathbf{X}$ be the resulting $M \times N$ normalized genotype matrix, and $\mathbf{x}_{i.}$ the ith row of $\mathbf{X}$. We have $\sum_{j=1}^{N} x_{ij} = 0$ and $\sum_{j=1}^{N} x_{ij}^2 = 1$. For mathematical precision in later development, the normalization used here is slightly different from that used in Price et al. (2006), but produces nearly identical PCs. From the singular value decomposition we obtain $\mathbf{X} = \mathbf{UDP}^T$, where $\mathbf{D}$ is an $N \times N$ diagonal matrix of ordered singular values with jth diagonal element d_j, $\mathbf{U}$ is an $M \times N$ loading matrix, and $\mathbf{P}$ is the $N \times N$ normalized PC matrix. Let $\boldsymbol{p}_{.j}$ be the jth column of $\mathbf{P}$ (*i.e.*, the jth PC), for $j \in \{1, \ldots, N\}$. Note that $\boldsymbol{p}_{.j}^T\boldsymbol{p}_{.j} = 1$, $\boldsymbol{p}_{.j}^T\boldsymbol{p}_{.k} = 0$ for $k \neq j$, and $\boldsymbol{p}_{.j}^T\mathbf{1} = 0$ for $j \in \{1, \ldots, N-1\}$ where $\mathbf{1} = \{1, \ldots, 1\}$. Finally, we use $\boldsymbol{y}$ to denote the vector of N phenotypes.

THEOREM 1. *Let* $\gamma_j = \boldsymbol{p}_{.j}^T\boldsymbol{y}$. *We have* $\sum_{i=1}^{M}(\boldsymbol{x}_{i.}^T\boldsymbol{y})^2 = \sum_{j=1}^{N}\gamma_j^2\lambda_j$,*where* $\lambda_j = d_j^2$ *is the* j*th eigenvalue of* $\mathbf{X}^T\mathbf{X}$.

The proof is given in Web Appendix A. As γ_j is an inner product between the (normalized) $\boldsymbol{p}_{.j}$ and $\boldsymbol{y}$, it is easy to show that

$$\gamma_j = \sqrt{\sum_{k=1}^{N}(y_k - \bar{y})^2} \times \mathrm{corr}(\boldsymbol{p}_{.j}, \boldsymbol{y}), \qquad (1)$$

where "corr" is the Pearson correlation coefficient. Thus γ_j is proportional to the correlation between the phenotype $\boldsymbol{y}$ and the jth PC. We emphasize that the correlation is a *sample* quantity, observable from the data. Similarly, each term $\boldsymbol{x}_{i.}^T\boldsymbol{y}$ in the equality is proportional to the correlation between the genotype at SNP i and the phenotype. The importance of the result lies in the explicit connection between these M genotype–phenotype correlations to the N PC–phenotype correlations.

2.1 *Relationship between Genomic Control and Principal Components*

Here we obtain explicit results for the relevant test statistics applied in association mapping. Theorem 1 technically applies to score test statistics. However, we later demonstrate that the results apply to other common choices of test statistic.

Quantitative traits. For continuous quantitative phenotype Y, we assume a simple linear regression model at each SNP i:$y_j = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_j$, $\epsilon_j \sim N(0, \sigma^2)$. To test an association of the SNP i with the phenotype, we can use the following score test statistic:

$$S_i = \frac{\left(\boldsymbol{x}_{i.}^T\boldsymbol{y}\right)^2}{\sum_{j=1}^{N}(y_j - \bar{y})^2/N}. \qquad (2)$$

By Theorem 1,

$$\frac{1}{M}\sum_{i=1}^{M} S_i = \left\{M\sum_{j=1}^{N}(y_j - \bar{y})^2/N\right\}^{-1} \times \sum_{j=1}^{N}\gamma_j^2\lambda_j = \frac{N}{M}\sum_{j=1}^{N}\mathrm{corr}^2(\boldsymbol{p}_{.j}, \boldsymbol{y})\lambda_j. \qquad (3)$$

The observed mean of all score test statistics across the M SNPs is thus proportional to the sum of the squared PC–phenotype correlations multiplied by their respective eigenvalues.

In a justification of genomic control, it has been argued that under certain models of population stratification, the inflation of test statistics should be similar across all "null" SNPs (Devlin et al., 2001). This and related work (Devlin and Roeder, 1999) compared the sample median of test statistics to the chi-square median value, for an estimated inflation factor $\hat{\tau} = \mathrm{median}(S)/0.456$, to be robust to outlying test statistics which presumably correspond to "alternative" SNPs. Other work (Reich and Goldstein, 2001; Devlin et al., 2004) has used the sample mean $\hat{\tau} = \bar{S} = (1/M)\sum_{i=1}^{M} S_i$ directly. In practice, the median and mean typically give similar results, as the proportion of alternative SNPs is typically small. Using either approach, for each i a new statistic $S_i' = S_i/\hat{\tau}$ is then compared to χ_1^2.

To summarize, the results above provide a direct relationship between the mean version of the genomic control quantity $\hat{\tau}$ (left-hand side of (3)) and the PC–phenotype correlations and eigenvalues. This relationship is more than a simple curiosity. Although it is known that distinct subpopulations can be represented using PCs (Price et al., 2008), we are not aware that a natural relationship has been previously established between the PCs and the testing procedures. Moreover, (3) is exact (not based on expectations), holding for any $\mathbf{X}$ and $\boldsymbol{y}$, regardless of the underlying population substructure and the proportion of alternative SNPs. Thus the right-hand side of (3) is subject to the same sampling variation as $\bar{S}$. We also note that, to the extent that increases in $\bar{S}$ above 1 determine type I error inflation, the equation specifically highlights the terms $\mathrm{corr}^2(\boldsymbol{p}_{.j}, \boldsymbol{y})\lambda_j$ as contributors to this inflation. PCs contributing meaningfully to this inflation must have appreciable values for both λ_j *and* $\mathrm{corr}^2(\boldsymbol{p}_{.j}, \boldsymbol{y})$. Due to sampling variation, PCs will exhibit sample $\mathrm{corr}^2(\boldsymbol{p}_{.j}, \boldsymbol{y}) > 0$ for each j, even if the PCs are truly uncorrelated with the population from which $\boldsymbol{y}$ is drawn. The eigenvalues $\lambda_1, \ldots, \lambda_{N-1}$ are also nonzero. Thus we must consider the magnitude of the terms, as well as sampling variation. Our general approach in the later sections will be to (i) re-rank the PCs by the terms $\mathrm{corr}^2(\boldsymbol{p}_{.j}, \boldsymbol{y})\ \lambda_j$, (ii) test for the statistical significance of each of the terms, and (iii) control for stratification using only those PCs with significant terms.

Case–control traits. We now establish that the relationships described above apply to case–control studies, with $Y = 0$ and $Y = 1$ corresponding to controls and cases, respectively. We use the logistic regression model for SNP i: $\log[P(Y = 1)/\{1 - P(Y = 1)\}] = \beta_{0i} + \beta_{1i}x_{ij}$, which is conditional on the sampling scheme. With N_1 cases and N_0 controls, the score test statistic is $S_i = (\boldsymbol{x}_{i.}^T\boldsymbol{y})^2/\{(N_0N_1)/N^2\}$. It is simple to show that $(N_0N_1)/N^2 = \sum_{j=1}^{N}(y_j - \bar{y})^2/N$, and so by comparison with (2), we see that (3) directly applies.

We have now demonstrated a direct connection between the genomic control inflation factor and the PCs, but the two correction approaches are fundamentally different. In genomic control, the inflation of test statistics is effectively assumed to be constant across all null SNPs. However, it can be shown that regression analyses with the top PCs used as covariates is equivalent to adjusting the inflation effect of each SNP separately (see Web Appendix B for details).

2.2 *EigenCorr: An Eigenvalue and Correlation-based PC Selection Procedure*

Using the result that the effect of $\boldsymbol{p}_{.j}$ on $\bar{S}$ is proportional to $\gamma_j^2\lambda_j$, we propose to select PCs based on the $\gamma_j^2\lambda_j$, which we call the EigenCorr scores. To determine the impact of a given PC, we describe two procedures, which differ in their underlying assumptions.

(1) *EigenCorr1:* We adopt the null hypothesis that the population correlation of the PCs and phenotypes is zero and that there is no population substructure. We are able to directly estimate the null distribution of EigenCorr scores using the TW approximation (Johnstone, 2001; Patterson et al., 2006) and the Fisher z-transformation applied to sample correlations (Fisher, 1921). That is, $\text{corr}(\boldsymbol{p}_{.j}, \boldsymbol{y})$ approximately follows the distribution of $(e^{2Z} - 1)/(e^{2Z} + 1)$, where $Z \sim N(0, 1/(N - 3))$. Using these approximations to the distributions of (independent) γ_j^2 and λ_j, we obtain null distributions for $\gamma_j^2\lambda_j$ by simulation, resulting in p-values for each EigenCorr score. The process proceeds sequentially. We simulate a random λ_1^* from the distribution of $(\xi T + \mu)\sum_{k=1}^{N-1}\lambda_k/(N-1)$, where T is a TW random variable, and ξ and μ are as described in Web Appendix C. We then simulate γ_1^* using the Fisher z-transformation to obtain p-values for $\gamma_1^2\lambda_1$. After excluding the first eigenvalue, we set $N = N - 1$, recompute μ and ξ, and follow the same procedure sequentially to obtain p-values for the remaining EigenCorr scores. Significant PCs are selected by the p-values, acknowledging multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995).

(2) *EigenCorr2:* In EigenCorr1, we assumed no population substructure. Although this assumption underlies TW testing (Patterson et al., 2006), we can relax the assumption, recognizing that significant eigenvalues alone are not sufficient to produce inflation of type I error. For EigenCorr2, we assume only that the PCs are uncorrelated with the population phenotype distribution. Here we treat the λ_j values as fixed, and use the Fisher z approximation to compute p-values for high values of γ_j^2. Although EigenCorr2 is simpler than EigenCorr1, and is shown to perform well in later simulations, both approaches may have value in different situations. In particular, EigenCorr1 may have an advantage in situations where few eigenvalues are truly significant. For either approach, our experience indicates that a relatively small number of PCs will be chosen for stratification control, which is desirable for both computational and statistical simplicity.

In the current practice of PC-based stratification control, investigators are often concerned that the inclusion of SNPs in high-linkage disequilibrium can produce misleading results. Thus many investigators "thin" out SNPs so that only a subset with lower correlations is used to generate PCs (e.g., Fellay et al., 2007), which is a special case of the *shrinkage PC* approach (Zou et al., 2010). We have shown (Web Appendix F) that EigenCorr procedures can be applied, but to the EigenCorr scores based on the weighted PCs. Further, simulations show that even when genomic control is performed using all SNPs, but PCs are calculated using a thinned set of SNPs, the approximate relationship still holds (Web Appendix G).

3. Simulations and Real-Data Analysis

We investigated the performance of the proposed EigenCorr approach in applications to simulated data and two real GWAS datasets.

3.1 *Simulation Studies*

Simulation 1: We simulated 1000 samples from five subpopulations with 20,000 uncorrelated SNPs, with 210 samples from each of the first four subpopulations, and the remaining 160 samples from subpopulation 5. For each SNP, the overall minor allele frequency was uniform from 0.05 to 0.5, and F_{st} uniform from 0.01 to 0.04. From these values, the minor allele frequency for each subpopulation was generated according to the Balding–Nichols model (Balding and Nichols, 1995). PC analysis showed that the top four PCs were significant according to the TW test, with $\boldsymbol{p}_{.4}$ specifically distinguishing subpopulation 5 from the others. To simulate population stratification, we generated a disease phenotype from a logistic model with log(odds ratio) = 1.6 between subpopulation 5 and the remaining samples. Therefore, increases in type I error resulting from population stratification arise entirely from the differing disease prevalence in the subpopulations. Figure 1 shows eigenvalues and EigenCorr scores of the first 10 PCs. The TW test selected the top four PCs as significant at $p < 0.01$, because its selection is entirely eigenvalue based, while only $\boldsymbol{p}_{.4}$ was identified, correctly, by EigenCorr. This simulation is illustrative of the intended advantage of EigenCorr scores.

Simulation 2: Simulations based on a real dataset. To investigate type I error and power for PC-based methods, we simulated phenotypes based on a real schizophrenia GWAS from the Genetic Association Information Network (GAIN) consortium [Version 2, Accession number: phs000021.v2.p1] (Sanders et al., 2008). In this manner we intended to reflect the genetic complexity encountered in real studies. The General Research Use African American data consist of 1904 samples with 845,814 SNPs, and was downloaded from dbGap at National Center for Biotechnology Information (NCBI; `ncbi.nlm.nih.gov`). After several data-filtering steps (See Web Appendix D), we obtained a final dataset with 1835 samples and 96,346 SNPs. TW testing identified 91 significant PCs with $p < 0.01$, which also corresponded to FDR control at 0.1. In contrast, the informal (but widely used) "scree plot" method of simply inspecting eigenvalues would

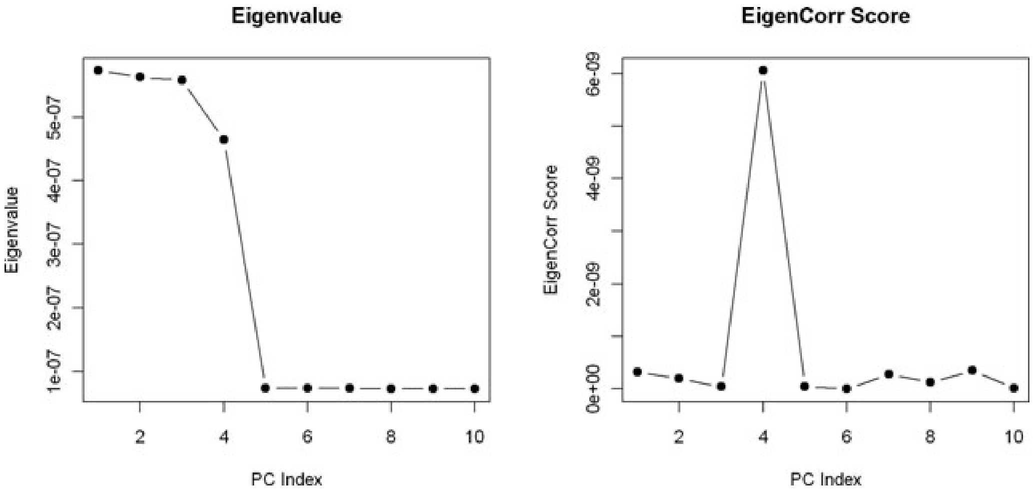

Figure 1. Illustration of the EigenCorr scores for simulation1. The left panel presents the first 10 eigenvalues and the right panel presents the first 10 EigenCorr scores. The importance of PC4 is clearly indicated by the EigenCorr score.

result in choosing two PCs (See Web Figure 1). A different version of the TW test is implemented in the GEM software EigenSoft, TW–GEM estimates the effective number of markers only once and uses it for all PCs when computing TW statistics.

We used $\boldsymbol{p}_{.1}, \boldsymbol{p}_{.2}, \boldsymbol{p}_{.5}$, and $\boldsymbol{p}_{.10}$ to generate association of strata with a simulated phenotype. For quantitative traits, under the null hypothesis for SNP association, we simulated phenotypes according to $y_j = \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10} + \epsilon_j$. Here the ϵ_j were generated from $N(0, 1)$, and η_1, η_2, η_5, η_{10} generated as independent and identically normally distributed so as to contribute the half of the variability of $\boldsymbol{y}$. Note that for all simulations, the genotypes were fixed, but the phenotypes were simulated prospectively, producing variation in the PCs selected by EigenCorr. On average, EigenCorr1 and EigenCorr2 selected 3.52 and 3.51 PCs, respectively, out of the first 200 PCs at FDR level 0.1. These PCs mostly overlapped and reflected the true PC stratification. Similarly, for a dichotomous trait, we simulated case–control status from the model

$$\log[P(Y_j = 1)/\{1 - P(Y_j = 1)\}]$$
$$= \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10},$$

where the coefficients η_1, η_2, η_5, η_{10} were randomly generated from a normal distribution such that the variance of the log odds ratio equaled 4.0.

For the alternative hypothesis, given the genotype x at a causal SNP, quantitative traits were generated from $y_j = \beta x + \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10} + \epsilon_j$, where $\beta = 0.15$, and dichotomous traits were generated from

$$\log[P(Y_j = 1)/\{1 - P(Y_j = 1)\}]$$
$$= \beta x + \eta_1 p_{j1} + \eta_2 p_{j2} + \eta_5 p_{j5} + \eta_{10} p_{j10},$$

with $\beta = \log(1.6) = 0.47$. The same η distributions used in the null models were applied.

To investigate the type I error and power, we ran 10,000 simulations for each model. In each null simulation a single SNP was chosen at random for investigation and p-value computation. Table 1 provides the empirical type I errors. We compared seven different approaches for PC selection: (1) no PCs for adjustment; (2) the four PCs that reflected the true population stratification; (3) the first two PCs, selected by scree plot examination; (4) the 91 PCs selected by the TW test; (5) the 12 PCs selected by TW–GEM; (6) the PCs selected by EigenCorr1; and (7) the PCs selected by EigenCorr2. The "true PC" approach can be viewed as a gold standard. The p-values were computed from the score statistics for the additive genotype effect. The results from likelihood ratio testing were similar (See Web Table 1).

From the table we can see that the "no-adjustment" and scree plot methods result in improper control of type I error. As expected, use of the four known confounding PCs properly controlled type I error. The 91 top PCs from the TW test controlled the type I error well for the quantitative trait, although the model was highly overparameterized. However, these 91 PCs result in somewhat inflated type I error for the dichotomous trait. In practice, investigators might be reluctant to fit so many covariates, but the absence of a principled alternate procedure based solely on eigenvalues makes it difficult to prescribe an alternative, when so many eigenvalues are clearly significant. The fewer PCs selected by the TW–GEM test controlled type I error in these data, as did the PCs from both EigenCorr methods. The statistical power of the models chosen by TW–GEM and both EigenCorr procedures were comparable to the gold standard of known PCs, while the TW procedure resulted in somewhat reduced power.

The average estimates of the genetic effect β are shown in Web Table 2. For the quantitative traits, all seven methods gave essentially unbiased estimates. The estimates from the logistic regression, however, were biased downward with adjustment by zero PCs or by the two scree-based PCs. Estimates were upwardly biased when using the 91 TW PCs. The presence of bias and poorly controlled type I error are well-known features of logistic regression if a large number of unnecessary (null) covariates are included in the analysis (Lubin, 1981).

Table 1

Performance of the methods for 10,000 *GWAS simulations, evaluated at a null or alternative SNP. Values in the table represent type I error (for the null simulations) or power (for the alternative simulations) from the score test. The simulation setups are described in "Simulations and Real-Data Analysis."*

Nominal significance α	No adjustment	Known counfounding PCs	Scree method	TW	TW–GEM	EigenCorr1	EigenCorr2
			Quantitative trait				
Null							
0.05	0.1661	0.0501	0.0789	0.0499	0.0497	0.0508	0.0508
10^{-2}	0.0782	0.0104	0.0229	0.0106	0.0099	0.0104	0.0105
Alternative							
10^{-2}	0.6871	0.7221	0.7110	0.6983	0.7205	0.7201	0.7191
10^{-4}	0.3901	0.3743	0.3777	0.3388	0.3713	0.3708	0.3703
10^{-6}	0.1877	0.1369	0.1487	0.1155	0.1345	0.1346	0.1334
			Case–control trait				
NULL							
0.05	0.1642	0.0490	0.0704	0.0561	0.0483	0.0492	0.0500
10^{-2}	0.0770	0.0095	0.0188	0.0118	0.0099	0.0096	0.0097
Alternative							
10^{-2}	0.7260	0.8532	0.7835	0.8464	0.8528	0.8524	0.8519
10^{-4}	0.4593	0.6315	0.5152	0.6193	0.6269	0.6294	0.6293
10^{-6}	0.2611	0.3943	0.2840	0.3832	0.3909	0.3917	0.3915

Table 2

Family-wise error rates (FWER) for the minimum score statistic p-values in 1000 *simulated whole genome scans, with population structure as described in the text. Values represent FWER when applying significance level* α *to each of the* 810,264 *SNPs. A significance level* $\alpha = 6.17 \times 10^{-8}$ *corresponds to Bonferroni control of FWER* $\leqslant 0.05$.

Nominal significance α	No adjustment	Known confounding PCs	Scree method	TW	TW–GEM	EigenCorr1	EigenCorr2
			Quantitative trait				
10^{-6}	0.9870	0.4460	0.8600	0.4460	0.4570	0.4450	0.4450
10^{-7}	0.9470	0.0620	0.5600	0.0680	0.0620	0.0660	0.0660
6.17×10^{-8}	0.9230	0.0380	0.5150	0.0400	0.0420	0.0400	0.0400
			Case–control trait				
10^{-6}	0.9970	0.4850	0.9120	0.7140	0.4840	0.4910	0.4910
10^{-7}	0.9690	0.0640	0.6800	0.1350	0.0620	0.0630	0.0630
6.17×10^{-8}	0.9620	0.0290	0.6410	0.0850	0.0390	0.0330	0.0330

To investigate the FWER under more stringent testing thresholds, we used the same setup to run 1000 null whole-genome scans using all 810,264 SNPs. Here the minimum GWAS p-value in each simulation was compared to prespecified thresholds (Table 2). The threshold of 6.15×10^{-8} corresponds to Bonferroni adjustment for FWER $= 0.05$. The precise FWER-controlling threshold is difficult to predetermine, due to the effects of SNP correlation and the use of asymptotic p-values at extreme thresholds. Nonetheless, for each threshold, the "known PC" situation can serve as a gold standard for comparison. Table 2 shows that the large number of PCs from the TW test inflates the FWER for dichotomous traits, doubling or tripling the error compared to using known PCs. The no-adjustment and scree inspection methods failed to control FWER for both continuous and dichotomous traits, but because too *few* PCs are used. For the simulation setup, we find that EigenCorr1, EigenCorr2, and TW–GEM offer reasonable FWER control. However, the real-data example below shows that for some datasets TW–GEM can fail to detect PCs which have a substantial impact on type I error.

3.2 *Real-Data Analysis*

We next applied the EigenCorr methods to two schizophrenia GWAS studies: (i) the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) dataset of Sullivan et al. (2008); (ii) the GAIN consortium dataset described earlier, using the actual study phenotypes. The CATIE analysis is described in detail here, with the GAIN analysis described in Web Appendix E.

3.3 *The CATIE Schizophrenia Data*

CATIE Schizophrenia GWAS data were obtained from the National Institute of Mental Health (NIMH) Genetic Repository (`nimhgenetics.org`), with 1492 samples (741 cases and

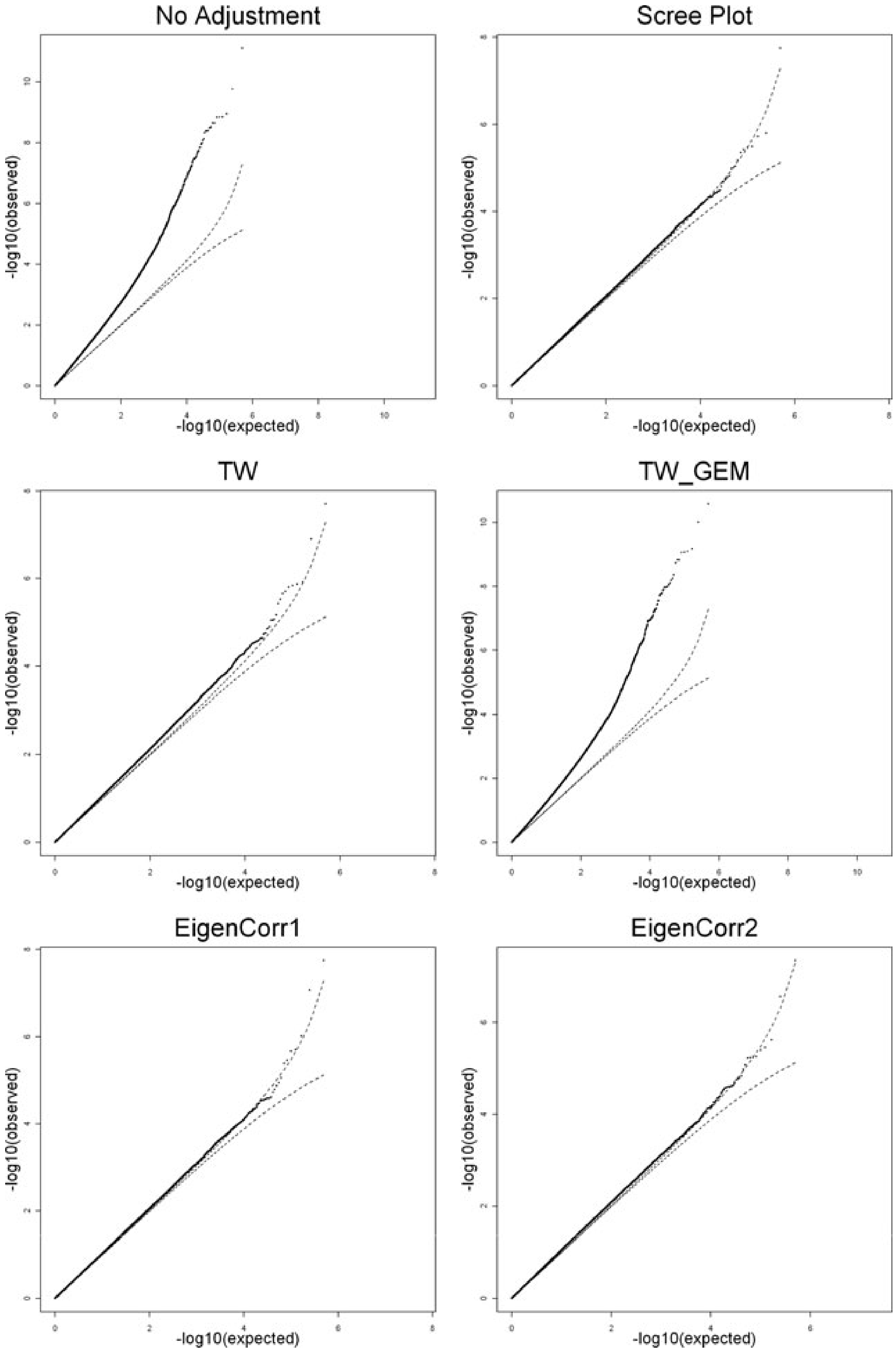

Figure 2. Plots of observed versus expected p-values ($-\log_{10}$ scale) for the CATIE data. The dashed lines indicate 95% prediction bands.

751 controls) and 495,172 SNPs. After applying the same data-filtering steps described in Web Appendix D, 1439 samples remained, with 71,985 thinned SNPs for PC analysis. Web Figure 5 presents sample eigenvalues, their correlations with disease status, and the corresponding EigenCorr scores. Inspection of the scree plot suggested use of the top two PCs. At FDR $= 0.1$, the TW test selected 45 PCs. In contrast, TW–GEM selected only the first PC. Of the top 200 PCs, both EigenCorr1 and EigenCorr2 selected three PCs, with two of them overlapping.

We applied all the PC-selection approaches in Simulation 2, except the true PC approach, as the true confounding PCs are of course unknown. Figure 2 shows the $-\log_{10}$ (p-values) (observed versus expected) from the various approaches. The TW–GEM shows a large deviation from the diagonal, indicating that adjustment by only the first PC was not sufficient to control for stratification. In contrast, the plots for EigenCorr1 and EigenCorr2 suggest proper type I error control. PC 2 has the largest EigenCorr score, with a large eigenvalue and high correlation with the case–control phenotype. In addition, the EigenCorr methods successfully find a short list of SNPs that meet genome-wide significance.

4. Discussion

We have shown that the average inflation of test statistics caused by population stratification can be expressed in terms of genotype PCs, according to their eigenvalues and their correlations with the phenotype. PCs that are uncorrelated with the phenotype have a negligible effect on the inflation of test statistics. The explicit connection of the PCs to genomic control provides insight into the advantages of PC-based methods. Importantly, this natural motivation for PC-based control does not require any assumptions about the nature of population structure. In addition, the results show that PCs computed from genotypes are "natural" variables to consider, even though discrete genotypes are far removed from the normality assumptions that underlie classical multivariate analysis. In addition to the statistical advantages of EigenCorr PC selection, the small number of PCs typically selected has a distinct computational advantage when applied at large scales, for example with expression quantitative trait locus (eQTL) analysis.

For simulations and real data, we have shown that the TW test selects too many PCs. In practice, investigators typically choose few PCs, but informal methods such as scree plot inspection can select too few. Similarly, the TW–GEM test clearly selected too few PCs for CATIE, because it does not approach PC testing sequentially, and can be less powerful for detecting important but lower-ranked PCs. We believe the EigenCorr approaches offer a principled alternative, and tend to correctly select the few most important PCs.

In our analyses, EigenCorr1 and EigenCorr2 produced similar results. EigenCorr1 offers a potential conceptual advantage, in that the EigenCorr score is used directly as a statistic, and lower-ranked PCs are given less emphasis. However, when it is clear that many eigenvalues are highly significant, the null PC assumption may not be realistic, and EigenCorr2 may be preferred.

Other investigators have used the phenotype to identify or construct genotype-based covariates. Epstein et al. (2007) described a general stratification score approach, using partial least squares (PLS) of phenotype on a number of markers. Lee et al. (2008) pointed out that PLS can result in overfitting to the phenotype and reduce power. However, Epstein et al. (2007) and the rejoinder (Epstein, Allen, and Satten, 2008) provide important clarification that desirable stratification control should explain some phenotype variability. This fact was also recognized by Kimmel et al. (2007), who described initial correction of phenotype by a small number of PC clusters before association testing. Zhao, Rebbeck, and Mitra (2009) described a propensity score using (genetic) covariates to predict genotype at a test locus before inclusion in a phenotype–genotype model. This approach, although implemented with relatively few genetic covariates (serving a role analogous to our PCs), potentially avoids overfitting in selecting covariates. However, it is potentially susceptible to inclusion of an excessive number of genetic covariates, which can present a problem for logistic regression.

We view the EigenCorr approach as generally similar to that advocated by Epstein et al. (2007), using PC-based phenotype adjustment in the form of regression covariates. However, the EigenCorr procedure, with a fixed number of PCs to choose from, is intentionally less flexible than procedures such as PLS, and FDR-based testing provides a natural penalty against overfitting. Moreover, in contrast to other procedures, the EigenCorr motivation and approach is explicitly connected to the source of test statistic inflation. The fact that genotype must also be associated with population stratum to create confounding is also implicit in EigenCorr, because each informative marker has an influence on, and will be associated with, at least one eigenvector. We believe that EigenCorr offers an efficient filter to identify the confounding variables of greatest influence.

5. Supplementary Materials

Web Appendices A–G, Figures, and Tables are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

Support was provided in part by NIH R01GM074175, EPA R832720, and a Gillings Innovation Laboratory award. The authors gratefully acknowledge the comments and suggestions of the reviewers and editors, which greatly improved the article. GAIN dataset were obtained from the GAIN database found at `http://view.ncbi.nlm.nih.gov/dbgap`—controlled through dbGaP accession number phs000021.v2.p1. Samples and associated phenotype data for the Linking Genome-Wide Association Study of Schizophrenia were provided by P. Gejman. The principal investigators of the CATIE trial were Jeffrey A. Lieberman, M.D., T. Scott Stroup, M.D., M.P.H., and Joseph P. McEvoy, M.D. The CATIE trial was funded by a grant from the National Institute of Mental Health (N01 MH900001) along with MH074027 (PI PF Sullivan). Genotyping was funded by Eli Lilly and Company.

References

Balding, D. and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96,** 3–12.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* **57,** 289–300.

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

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

Devlin, B., Bacanu, S., and Roeder, K. (2004). Genomic control to the extreme. *Nature Genetics* **36,** 1129–1130.

Epstein, M., Allen, A., and Satten, G. (2007). A simple and improved correction for population stratification in case-control studies. *The American Journal of Human Genetics* **80,** 921–930.

Epstein, M., Allen, A., and Satten, G. (2008). Response to Lee et al. *The American Journal of Human Genetics* **82,** 526–528.

Fellay, J., Shianna, K., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagha, A., Cossarizza, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J. M., Obel, N., Smith, J. P., Wyniger, J., Descombes, P., Antonarakis, S. E., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., and Goldstein, D. B. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* **317,** 944–947.

Fisher, R. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1,** 3–32.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics* **29,** 295–327.

Kimmel, G., Jordan, M., Halperin, E., Shamir, R., and Karp, R. (2007). A randomization test for controlling population stratification in whole-genome association studies. *The American Journal of Human Genetics* **81,** 895–905.

Lee, S., Sullivan, P., Zou, F., and Wright, F. (2008). Comment on a simple and improved correction for population stratification. *The American Journal of Human Genetics* **82,** 524–526.

Lubin, J. (1981). An empirical evaluation of the use of conditional and unconditional likelihoods for case-control data. *Biometrika* **68,** 567–571.

Marchini, J., Cardon, L., Phillips, M., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36,** 512–517.

Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2,** e190.

Price, A., Butler, J., Patterson, N., Capelli, C., Pascali, V., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A., Korkolopoulou, P., Seligsohn, U., Waliszewska, A., Schirmer, C., Ardlie, K., Ramos, A., Nemesh, J., Arbeitman, L., Goldstein, D. B., Reich, D., and Hirschhorn, J. N. (2008). Discerning the ancestry of European Americans in genetic association studies. *PLoS Genetics* **4,** e236.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–909.

Reich, D. and Goldstein, D. (2001). Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology* **20,** 4–16.

Sanders, A., Duan, J., Levinson, D., Shi, J., He, D., Hou, C., Burrell, G., Rice, J., Nertney, D., Olincy, A., Rozic, P., Vinogradov, S., Buccola, N. G., Mowry, B. J., Freedman, R., Amin, F., Black, D. W., Silverman, J. M., Byerley, W. F., Crowe, R. R., Cloninger, C. R., Martinez, M., and Gejman, P. V. (2008). No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: Implications for psychiatric genetics. *American Journal of Psychiatry* **165,** 497–506.

Sullivan, P., Lin, D., Tzeng, J., van den Oord, E., Perkins, D., Stroup, T., Wagner, M., Lee, S., Wright, F., Zou, F., Liu, W., Downing, A. M., Lieberman, J., and Close, S, L. (2008). Genomewide association for schizophrenia in the CATIE study: Results of stage 1. *Molecular Psychiatry* **13,** 570–584.

Zhao, H., Rebbeck, T., and Mitra, N. (2009). A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors. *Genetic Epidemiology* **33,** 679–690.

Zou, F., Lee, S., Knowles, M., and Wright, F. (2010). Quantification of population structure using correlated SNPs by shrinkage principal components. *Human Heredity* **70,** 9–22.

Received June 2009. *Revised July* 2010.
Accepted September 2010.