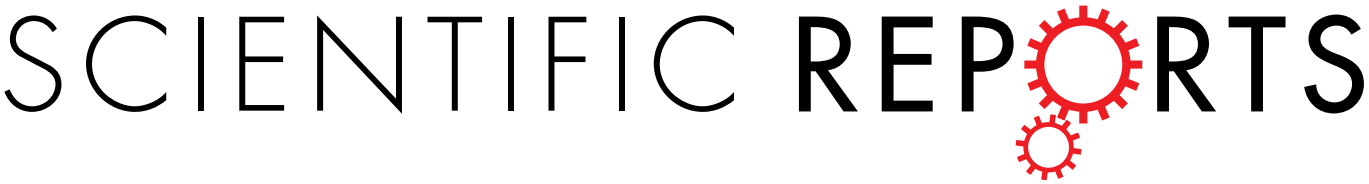

OPEN

Power Calculation of Multi-step Combined Principal Components with Applications to Genetic Association Studies

Received: 12 January 2016
Accepted: 28 April 2016
Published: 18 May 2016

Zhengbang Li[1,*], Wei Zhang[2,*], Dongdong Pan[3] & Qizhai Li[2]

Principal component analysis (PCA) is a useful tool to identify important linear combination of correlated variables in multivariate analysis and has been applied to detect association between genetic variants and human complex diseases of interest. How to choose adequate number of principal components (PCs) to represent the original system in an optimal way is a key issue for PCA. Note that the traditional PCA, only using a few top PCs while discarding the other PCs, might significantly lose power in genetic association studies if all the PCs contain non-ignorable signals. In order to make full use of information from all PCs, Aschard and his colleagues have proposed a multi-step combined PCs method (named mCPC) recently, which performs well especially when several traits are highly correlated. However, the power superiority of mCPC has just been illustrated by simulation, while the theoretical power performance of mCPC has not been studied yet. In this work, we attempt to investigate theoretical properties of mCPC and further propose a novel and efficient strategy to combine PCs. Extensive simulation results confirm that the proposed method is more robust than existing procedures. A real data application to detect the association between gene TRAF1-C5 and rheumatoid arthritis further shows good performance of the proposed procedure.

Identification of genetic variants associated with human complex diseases can help investigators further understand genetic structure of diseases of interest. Compared with single-marker analysis, which tests every marker individually and is commonly employed in genome-wide association study, multiple-marker test has been well appreciated because of its potentially improved statistical power. Statistical methods for multiple-marker analysis can be summarized as synthesizing single-marker test statistics such as Hotelling's T^2 test[1–3] and summation of squared univariate test[4,5], weighted Fourier transformation[6], variance-components score test[7], principal components regression method[8–10], and Kernel-machine-based test[11]. Performances of these methods have been explored by intensive computer simulations[1,12,13]. Their results showed that when the number of SNPs is relatively large, variance-component-based methods and principal components regression methods were found to have competitive power.

As it is well known that principal component analysis (PCA) is a useful tool to search for important characteristics among correlated variables. A key issue in developing an effective PCA model is choosing an adequate number of principal components (PCs) to represent the system in an optimal way. Taking advantage of the size of variances, Hocking[14] provided a firm rule for retaining PCs in the framework of regression models. Usually, in PCA, investigators only used a few top principal components and discarded the other PCs. Recently, some investigators have illustrated that commonly used method for choosing PCs is not always reasonable. In fact, as early as in 1982, Jollife[15] showed an interesting counter-intuitive phenomenon that principal components explaining a small amount of variances can be as important as those explaining a large amount of variances when analyzing non-genetic data. Aschard *et al.*[16] confirmed this phenomenon when analyzing genetic data and proposed a

[1]School of Mathematics and Statistics & Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, 430079, China. [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China. [3]Department of Statistics, Yunnan University, Kunming, 650091, China. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Z.L. (email: lizhengbang@mail.ccnu.edu.cn)

called multi-step combined principal component (mCPC) strategy. However, the performance of mCPC strongly depends on how to partition all PCs.

Without loss of generality, suppose a random vector T follows a multivariate normal distribution with a $m \times 1$ mean vector μ and known $m \times m$ covariance matrix V. We want to test the null hypothesis H_0: $\mu = 0$. Therefore, a Chi-squared statistic can be used for testing H_0. However, when m is large, which is fairly common in genome-wide association studies, Chi-squared test might substantially lose power due to its large degrees of freedom. To reduce degrees of freedom, PCA is recommended. Based on orthogonal decomposition, we have $V = Q\Lambda Q^\tau$, where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0, Q = (q_1 \colon q_2 \colon \cdots \colon q_m)$, and q_i is called as the eigenvector corresponding to the eigenvalue $\lambda_i, i = 1, 2, \cdots, m$. Define $Z_i = (q_i^\tau T)^2/\lambda_i$ for $i = 1, 2, \cdots, m$. We note that Z_i is related to the ith PC for $i = 1, 2, \cdots, m$. Under $H_0, Z_i \sim \chi_1^2$, a central Chi-squared distribution with 1 degree of freedom for $i = 1, 2, \cdots, m$. Under the alternative hypothesis, $Z_i \sim \chi_1^2(\Omega_i)$, a noncentral Chi-squared distribution with 1 degree of freedom and non-centrality parameter $\Omega_i = (q_i^\tau \mu)^2/\lambda_i$, for $i = 1, 2, \cdots, m$.

For simplicity with $m = 2$, we consider a linear model with a normally distributed phenotype Y, which depends on two scaled genotypes G_1 and G_2 that are also normally distributed with mean 0 and variance 1. So the phenotype can be expressed as: $Y = \beta_0 + G_1\beta_1 + G_2\beta_2 + \varepsilon$, where ε is the random error term which is distributed from the standard normal distribution. For this general model, the principal components of these two genotypes are $PC_1 = \frac{\sqrt{2}}{2}G_1 + \frac{\sqrt{2}}{2}G_2$, and $PC_2 = \frac{\sqrt{2}}{2}G_1 - \frac{\sqrt{2}}{2}G_2$. By some algebras, we can get $G_1 = \frac{\sqrt{2}}{2}PC_1 + \frac{\sqrt{2}}{2}PC_2$, and $G_2 = \frac{\sqrt{2}}{2}PC_1 - \frac{\sqrt{2}}{2}PC_2$. Phenotype Y can be reexpressed as: $Y = \beta_0 + PC_1\left(\frac{\sqrt{2}}{2}\beta_1 + \frac{\sqrt{2}}{2}\beta_2\right) + G_2\left(\frac{\sqrt{2}}{2}\beta_1 - \frac{\sqrt{2}}{2}\beta_2\right) + \varepsilon$, which indicates that PC_2 may be very important as $\frac{\sqrt{2}}{\sqrt{2}}\beta_1 - \frac{\sqrt{2}}{\sqrt{2}}\beta_2$ is large, although the variance of PC_2 is less than that of PC_1. So we can not discard any PCs arbitrarily. In order to test H_0: $\mu = 0$, Aschard *et al.*[17] proposed a multi-step combined principal component (mCPC) as following $\mathrm{mCPC}(k) = -2[\ln(1 - F_k(Z_1 + Z_2 + \cdots + Z_k)) + \ln(1 - F_{m-k}(Z_{k+1} + Z_{k+2} + \cdots + Z_m))]$, where $F_k(\cdot)$ is cumulative distribution function of a central Chi-squared random variable with k degrees of freedom. Moreover, they used simulation to compare the power of various PCA-based strategies when analyzing up to 100 correlated traits, and showed that their method with combining the signals across all PCs could have greater power. However, there has not been an in-depth study of the theoretical properties of mCPC in Aschard *et al.*'s paper[16]. Obviously, Aschard *et al.*[16] find an unusual way to fully utilize all PCs. Another key issue is to decide the value of k. A commonly used method for selecting k is based on cumulative contribution rates, which are equal to $\sum_{i=1}^k \lambda_i / \sum_{i=1}^m \lambda_i \times 100\%$, and denoted by c_k for $k = 1, \cdots, m$, respectively. Let $k_c = \min\{k \sum_{i=1}^k \lambda_i / \sum_{i=1}^m \lambda_i \geq c;\ k = 1, \cdots, m.\}$, for any $c \in [0, 1]$. Aschard *et al.*[16] followed the traditional way to use mCPC (k) with k being determined by cumulative contribution rate of 80%.

In this work, we focus on the theoretical power of mCPC and find that the maximum power of mCPC is related to the maximum noncentral parameters under alternative hypothesis. We also find that the noncentral parameter corresponding to the top PC (the first PC which corresponds to the largest eigenvalue) is greater than 0 under most scenarios and those of other PCs do not possess this property when only a few means of all PCs are non-zero under alternative and the correlation coefficients among original variables are relatively large. Herein we propose a method tCPC. Based on numerical results, the tCPC is more powerful than the existing procedures under most of the considered scenarios.

Results

Theoretical Properties of mCPC (k). For the multiple genetic variants association studies, the above random vector can be written as $T = (T_1, T_2, \cdots, T_m)$, where T_i is the statistic that is used to test for the association between the phenotype of interested and the ith genetic variants, $i = 1, 2, \cdots, m$. V is the covariance matrix of the random vector T. Through the eigen-decomposition of the covariance matrix, we have $V = Q\Lambda Q^\tau$, where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0, Q = (q_1 \colon q_2 \colon \cdots \colon q_m)$, and q_i is the eigenvector corresponding to the eigenvalue $\lambda_i, i = 1, 2, \cdots, m$. Then we can obtain transformed statistics as $Z_i = (q_i^\tau T)^2/\lambda_i, i = 1, 2, \cdots, m$. Furthermore, under H_0, Z_i follows a central Chi-squared distribution with 1 degree of freedom for $i = 1, 2, \cdots, m$. Under the alternative hypothesis, $Z_i \sim \chi_1^2(\Omega_i)$, a noncentral Chi-squared distribution with 1 degree of freedom and non-centrality parameter $\Omega_i, i = 1, 2, \cdots, m$.

For $i = 1, 2, \cdots$, let $F_i^{-1}(\cdot)$ be the inverse function of $F_i(\cdot)$. Note that for any given $x \in [0, 1]$, $\mathrm{mCPC}(k) = F_2^{-1}[F_k(\sum_{i=1}^k Z_i)] + F_2^{-1}[F_{m-k}(\sum_{i=k+1}^m Z_i)]$. Under H_0, both $F_k(\sum_{i=1}^k Z_i)$, and $F_{m-k}(\sum_{i=k+1}^m Z_i)$ follow uniform distribution on [0, 1] and they are independent to each other. So both $F_2^{-1}[F_k(\sum_{i=1}^k Z_i)]$, and $F_2^{-1}[F_{m-k}(\sum_{i=k+1}^m Z_i)]$ follow Chi-square distributions with 2 degrees of freedom, then mCPC (k) follows a central Chi-squared distribution with 4 degrees of freedom.

According to Sankaran[17], probability density function of a noncentral Chi-squared distribution with d degrees of freedom and non-centrality parameter ξ is $f_{d,\xi}(x) = \frac{e^{-\frac{1}{2}(x+\xi)}}{2^{\frac{1}{2}d}} \sum_{i=0}^\infty \frac{x^{\frac{1}{2}d+i}\xi^i}{2^{2i} i! \Gamma\left(\frac{1}{2}d + i\right)}$. Denote $\eta_1 = \sum_{i=1}^k \Omega_i$, and $\eta_2 = \sum_{i=k+1}^m \Omega_i$. For any $x > 0$, and $k \in \{1, 2, \cdots, m\}$, the probability density function of $F_2^{-1}[F_k(\sum_{i=1}^k Z_i)]$ is

$$h_1(x, k) = \frac{\frac{1}{2}\exp\left(-\frac{1}{2}x\right) f_{k,\eta_1}\{F_k^{-1}[F_2(x)]\}}{g_{1,k}\{F_k^{-1}[F_2(x)]\}} I_{\{x>0\}}, \tag{1}$$

where $g_{1,k}(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} I_{\{x>0\}}$, and the probability density function of $F_2^{-1}[F_{m-k}(\sum_{i=k+1}^m Z_i)]$ is

$$h_2(x,\,k) = \frac{\frac{1}{2}e^{-\frac{1}{2}x} f_{m-k,\eta_2}\{F^{-1}_{m-k}[F_2(x)]\}}{g_{2,m-k}\{F^{-1}_{m-k}[F_2(x)]\}} I_{\{x>0\}}, \tag{2}$$

with $g_{2,m-k}(x) = \frac{1}{2^{\frac{m-k}{2}}\Gamma\left(\frac{m-k}{2}\right)} x^{\frac{m-k}{2}-1} e^{-\frac{x}{2}} I_{\{x>0\}}$.

Let $C_{1-\alpha}$ be $1-\alpha$ quantile of a central Chi-squared distribution with 4 degrees of freedom. The power of mCPC (k) under the significance level α is

$$\begin{aligned}
\beta(k,\,\eta_1,\,\eta_2) &= \int_0^\infty dx_1 \int_{C_{1-\alpha}-x_1}^\infty h_1(x_1)h_2(x_2)dx_2 \\
&= \int_0^\infty \left[\int_{C_{1-\alpha}}^\infty h_1(x_1)h_2(x_2-x_1)dx_1\right]dx_2 \\
&\quad \int_0^\infty \left[\int_{C_{1-\alpha}}^\infty \frac{\frac{1}{2}e^{-\frac{1}{2}x_1} f_{k,\eta_1}\{F_k^{-1}[F_2(x_1)]\}}{g_{1,k}\{F_k^{-1}[F_2(x_1)]\}} \right. \\
&\quad \left. \times \frac{\frac{1}{2}e^{-\frac{1}{2}(x_2-x_1)} f_{m-k,\eta_2}\{F^{-1}_{m-k}[F_2(x_2-x_1)]\}}{g_{2,m-k}\{F^{-1}_{m-k}[F_2(x_2-x_1)]\}} dx_1\right] dx_2 \\
&\quad \int_0^\infty \left[\int_{C_{1-\alpha}}^\infty \frac{\frac{1}{4}e^{-\frac{1}{2}x_2} f_{k,\eta_1}\{F_k^{-1}[F_2(x_1)]\}}{g_{1,k}\{F_k^{-1}[F_2(x_1)]\}} \right. \\
&\quad \left. \times \frac{f_{m-k,\eta_2}\{F^{-1}_{m-k}[F_2(x_2-x_1)]\}}{g_{2,m-k}\{F^{-1}_{m-k}[F_2(x_2-x_1)]\}} dx_1\right] dx_2, \\
&\quad k \in \{1,\,2,\,\cdots,\,m\}.
\end{aligned} \tag{3}$$

Based on the above notations, the non-centrality parameter of the distribution of Z_1 which corresponds to the first PC is $\Omega_1 = (q_1^\tau \mu)^2/\lambda_1$, where μ is the mean vector of T under the alternative hypothesis. $\Omega_1 = 0$ if and only if the mean vector μ belongs to the space that expanded by the other $m-1$ eigenvectors q_2, q_3, $\cdots$, q_m, that is $\mu \in \{a_2q_3 + a_3q_3 + \cdots + a_mq_m\}$, a_2, a_3, $\cdots$, a_m are $m-1$ real numbers. However, for a m-dimensional space, $\Pr(\mu \in \{a_2q_3 + a_3q_3 + \cdots + a_mq_m\}) = 0$. Hence, the non-centrality parameter of the Chi-squared distribution of the statistic Z_1 is not equal to 0 almost everywhere. Besides this, and since the top PC possesses the largest variation among all PCs and $k = 1$ is a boundary point of the set consisting of 1, 2, $\cdots$, m, herein we propose to use the following strategy (named tCPC) to combine all PCs

$$\text{tCPC} = \text{mCPC}(1) = -2\,[\,\ln(1 - F_1(Z_1)) + \ln(1 - F_{p-1}(Z_2 + \cdots + Z_m))\,]. \tag{4}$$

Under null hypothesis of no association at any locus, tCPC follows a central Chi-squared distribution with 4 degrees of freedom.

Simulation Settings and Numerical Results. In this subsection, we conduct simulation studies to compare powers between tCPC to some exiting approaches such as Hotelling's T^2 test (HT)[1–3], ordinary PCA ($\text{oPC}(k_{0.8}) = Z_1 + \cdots + Z_{k_{0.8}}$), summation of squared univariate test statistic (SSU)[4], sequence kernel association test (SKAT)[11] and multi-step combine principal component test mCPC ($k_{0.8}$)[17].

Consider testing association between m genetic variants (or SNPs) and a complex human disease. Let $T = (T_1,\,T_2,\,\cdots,\,T_m)^\tau$ be a test statistic, where τ means the transpose of a vector or matrix. For example, we can construct T using the method in Chatterjee *et al.*[18] to detect genetic association between m SNPs and a binary trait as

$$T = \frac{n_1 n_2}{n_1 + n_2}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} g_{i,1} - \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} g_{i,1},\,\cdots,\,\frac{1}{n_1}\sum_{i=1}^{n_1} g_{i,m} - \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} g_{i,m}\right)^\tau, \tag{5}$$

where $g_{i,j}$ denotes the genotype of the jth SNP for the ith individual and n_1 and n_2 are the sample size of the case group and control group, respectively. Under the null hypothesis that these m SNPs are not associated with the disease of interest, T follows a multivariate normal distribution $N(\mu_{m\times 1},\,V_{m\times m})$ asymptotically with the mean vector $\mu_{m\times 1}$ and covariance matrix $V_{m\times m} = \frac{n_1 n_2}{n_1 + n_2} V_G$, in which V_G is the pooled-sample covariance matrix of all SNPs.

In order to obtain T and $V_{m\times m}$, we first generate a latent vector with length of 20 from a multivariate normal distribution with covariance structures of a compound symmetry with equal pairwise correlation ρ. Then, this latent vector is dichotomized to yield a haplotype with predesignated minor allele frequency (MAF). We repeat the above process 100,000 times to form a large population. Without loss of generality, we designate the first SNP as disease-causal SNP with MAF being p, and other SNPs as noncausal SNPs with MAFs all being q. Both sizes

40.5 cmType	ρ	β_1	p	q	HT	oPC (0.8)	SSU	SKAT	mCPC ($k_{0.8}$)	tCPC
40.1 cmI 40.5 cm error	0.20	0	0.20	0.20	0.047	0.047	0.044	0.048	0.044	0.048
	0.50	0	0.20	0.20	0.051	0.052	0.043	0.049	0.051	0.045
	0.80	0	0.20	0.20	0.053	0.049	0.051	0.052	0.050	0.051
	0.95	0	0.20	0.20	0.046	0.053	0.049	0.052	0.047	0.047
	0.20	ln 1.6	0.05	0.30	0.448	0.054	0.140	0.446	0.610	0.404
	0.20	ln 1.4	0.20	0.20	0.754	0.736	0.772	0.762	0.700	0.766
	0.20	ln 1.4	0.30	0.05	0.884	0.922	0.996	0.886	0.856	0.996
	0.50	ln 1.6	0.05	0.30	0.450	0.106	0.288	0.430	0.524	0.496
40.8 cm Power	0.50	ln 1.4	0.20	0.20	0.724	0.756	0.810	0.816	0.702	0.856
	0.50	ln 1.4	0.30	0.05	0.886	0.928	0.996	0.894	0.874	0.988
	0.80	ln 1.6	0.05	0.30	0.450	0.144	0.352	0.440	0.456	0.528
	0.80	ln 1.4	0.20	0.20	0.740	0.778	0.918	0.918	0.754	0.924
	0.80	ln 1.4	0.30	0.05	0.872	0.942	0.984	0.818	0.910	0.964
	0.95	ln 1.6	0.05	0.30	0.484	0.246	0.300	0.368	0.500	0.520
	0.95	ln 1.4	0.20	0.20	0.736	0.940	0.972	0.972	0.924	0.948
	0.95	ln 1.4	0.30	0.05	0.880	0.996	0.962	0.684	0.988	0.920

Table 1. Empirical type I error rates and powers of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$) and tCPC for Constant correlations.

of case samples and control samples are set to be 1,000. Case or control status of one subject is generated from a logistic regression model

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = 0)} = \beta_0 + \beta_1 g_{i,1} + \cdots + \beta_{20} g_{i,20}, \tag{6}$$

with $\beta_0 = -1.5$, $\beta_1 \in \{\ln(1.4), \ln(1.6)\}$ denoting the log odds ratio for the disease-causal SNP, and $\beta_2 = \cdots = \beta_{20} = 0$ denoting log odds ratios for the non-causal SNPs, where $Y_i = 1$ or 0 represents disease or healthy status of the ith individual, $i = 1, \cdots, 2000$. The nominal significance level is 0.05 throughout the whole simulation, and the number of replicates is 1,000. All parameter settings and their relevant results are displayed in Table 1. Table 1 shows that all these tests can control type I error rate correctly. For example, when the correlation coefficient of these 20 SNPs are equal to $\rho = 0.50$, $\beta_1 = 0$ and $p = q = 0.20$, the empirical type I error rates of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$), and tCPC are 0.051, 0.052, 0.043, 0.049, 0.051, and 0.045, respectively. The results of power comparison shows that tCPC performs more robustly than the other methods. For example, when the correlation coefficients of these 20 SNPs are uniformly equal to 0.20 and $\beta_1 = \ln 1.4$, $p = q = 0.20$, the empirical powers of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$), and tCPC are 0.754, 0.736, 0.772, 0.766, 0.700, and 0.766, respectively. The empirical power of tCPC is a little lower than that of SSU in this scenario. However, when $\rho = 0.50$, $\beta_1 = \ln 1.4$, and $p = q = 0.20$, the empirical powers of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$), and tCPC are 0.724, 0.756, 0.810, 0.816, 0.702, and 0.856, respectively. It is obvious that tCPC performs the best among all the considered procedures in this setting.

Next we consider a decreasing correlation structure. As a preliminary step, a latent vector with length of 20 is generated from a multivariate normal distribution with covariance matrix being $(\rho^{|i-j|})_{20\times 20}$. Other simulation settings are similar as above and are shown in Table 2. As presented in Table 2, the empirical powers of all the tests are close to the nominal significance level which indicates that they can control type I error rate correctly. For instance, when $\rho = 0.50$, $\beta_1 = 0$ and $p = q = 0.20$, the empirical type I error rates of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$), and tCPC are 0.036, 0.032, 0.04, 0.044, 0.038, and 0.042, respectively. For power comparison, tCPC still performs more robustly than the other methods. For example, when correlations of these 20 SNPs are decreasing with distance as $\rho = 0.80$, $\beta_1 = \ln 1.4$, $p = 0.20$, and $q = 0.20$, the empirical powers of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$), and tCPC are 0.77, 0.846, 0.772, 0.762, 0.788, and 0.836, respectively. The empirical power of tCPC is a little lower than that of oPC ($k_{0.8}$) in this case. However, when $\rho = 0.95$, $\beta_1 = \ln 1.4$, $p = 0.20$, and $q = 0.20$, the powers of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$) and tCPC are 0.746, 0.898, 0.858, 0.858, 0.86 and 0.864, respectively. It indicates that tCPC gives the maximum power among the six methods in this scenario. Compared Tables 1 and 2 comprehensively, we can see that, when linkage disequilibrium extents among all SNPs are relatively strong, tCPC performs more robustly than existing statistical methods.

Applications to gene TRAF1-C5 associated with Rheumatoid Arthritis. We apply tCPC and the other five existing tests to detect the association between gene TRAF1-C5 and rheumatoid arthritis using the data from the Genetic Analysis Workshop 16[19]. Our goal is to detect whether there is an association between gene TRAF1-C5 and rheumatoid arthritis. This gene has been reported to be deleterious previously[20]. There are 2,062 subjects including 868 cases and 1,194 controls in this study. The gene TRAF1-C5 consists of 38 SNPs. The p-values of HT, oPC (0.8), SSU, SKAT, mCPC ($k_{0.8}$) and tCPC of detecting associations between gene TRAF1-C5 and rheumatoid arthritis are 5.21×10^{-5}, 7.58×10^{-3}, 5.95×10^{-4}, 6.50×10^{-5}, 7.56×10^{-5} and 3.75×10^{-5}, respectively. If we use the p-value threshold of 5×10^{-5} as the moderate association at the genome-wide level

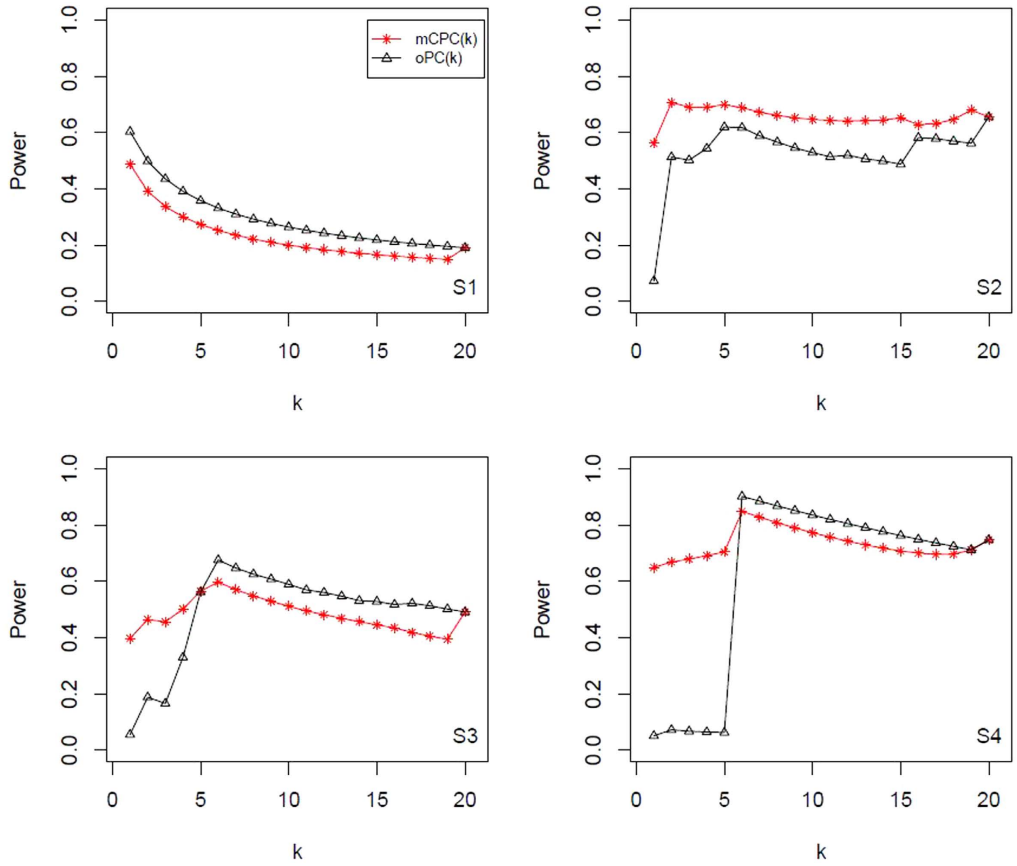

Figure 1. Powers of mCPC (k) and oPC (k) under significant level $\alpha = 0.05$ for Scenarios (S1) to (S4).

40.5 cmType	ρ	β_1	p	q	HT	oPC (0.8)	SSU	SKAT	mCPC ($k_{0.8}$)	tCPC
40.1 cm I 40.5 cm error	0.20	0	0.20	0.20	0.062	0.048	0.054	0.052	0.052	0.040
	0.50	0	0.20	0.20	0.036	0.032	0.040	0.044	0.038	0.042
	0.80	0	0.20	0.20	0.052	0.038	0.042	0.044	0.056	0.046
	0.95	0	0.20	0.20	0.036	0.046	0.052	0.054	0.042	0.048
	0.20	ln 1.6	0.05	0.30	0.470	0.060	0.110	0.470	0.650	0.386
	0.20	ln 1.4	0.20	0.20	0.746	0.756	0.766	0.738	0.700	0.688
	0.20	ln 1.4	0.30	0.05	0.886	0.926	0.998	0.890	0.864	0.998
	0.50	ln 1.6	0.05	0.30	0.490	0.074	0.126	0.466	0.568	0.390
40.8 cm Power	0.50	ln 1.4	0.20	0.20	0.722	0.784	0.742	0.706	0.686	0.662
	0.50	ln 1.4	0.30	0.05	0.882	0.914	0.996	0.862	0.856	0.996
	0.80	ln 1.6	0.05	0.30	0.482	0.120	0.148	0.350	0.478	0.400
	0.80	ln 1.4	0.20	0.20	0.770	0.846	0.772	0.762	0.788	0.836
	0.80	ln 1.4	0.30	0.05	0.888	0.946	0.994	0.852	0.912	0.988
	0.95	ln 1.6	0.05	0.30	0.480	0.192	0.266	0.358	0.464	0.510
	0.95	ln 1.4	0.20	0.20	0.746	0.898	0.858	0.858	0.860	0.864
	0.95	ln 1.4	0.30	0.05	0.866	0.972	0.982	0.732	0.962	0.936

Table 2. Empirical type-1 error rates and powers of HT, oPC ($k_{0.8}$), SSU, SKAT, mCPC ($k_{0.8}$) and tCPC for decreasing correlations.

as Burton *et al.*[21], only the proposed tCPC can detect the moderate-strong association signal between the gene TRAF1-C5 and rheumatoid arthritis.

Discussion

Principal component analysis is a common tool to grasp important features of correlated variables and has been applied in genetic association studies. In principal component analysis, cumulative contribution rate of 80% or

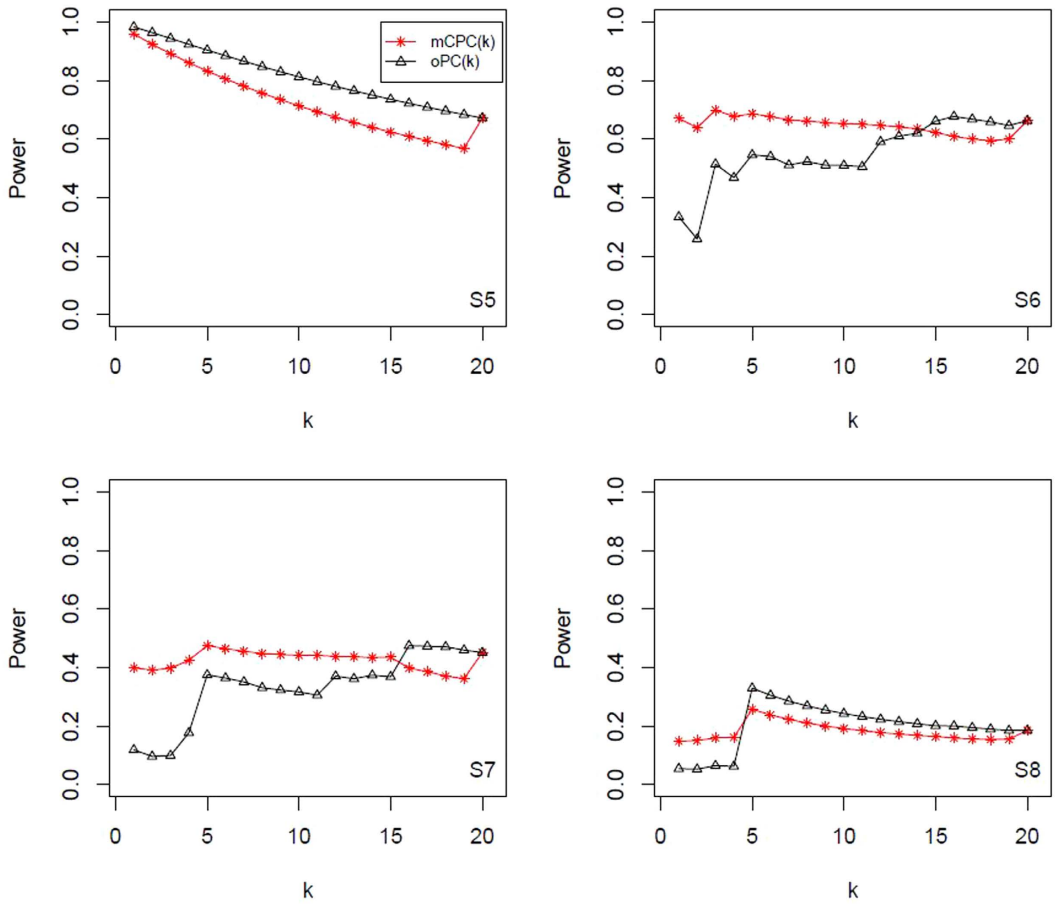

Figure 2. Powers of mCPC (k) and oPC (k) under significant level $\alpha = 0.05$ for Scenarios (S5) to (S8).

Covariance matrix	Scenarios	Mean vectors
Uniform correlation with $\rho = 0.8$	(S1)	$\mu_1 = \cdots = \mu_{20} = 2$
	(S2)	$\mu_1 = \cdots = \mu_{10} = 0.8, \mu_{11} = \cdots = \mu_{20} = 0$
	(S3)	$\mu_1 = \cdots = \mu_5 = 0.8, \mu_6 = \cdots = \mu_{20} = 0$
	(S4)	$\mu_1 = 2, \mu_2 = \cdots = \mu_{20} = 0$
Uniform correlation with $\rho = 0.2$	(S5)	$\mu_1 = \cdots = \mu_{20} = 2$
	(S6)	$\mu_1 = \cdots = \mu_{10} = 1.5, \mu_{11} = \cdots = \mu_{20} = 0$
	(S7)	$\mu_1 = \cdots = \mu_5 = 1.5, \mu_6 = \cdots = \mu_{20} = 0$
	(S8)	$\mu_1 = 2, \mu_2 = \cdots = \mu_{20} = 0$
Decreasing correlation with $\rho = 0.8$	(S9)	$\mu_1 = \cdots = \mu_{20} = 1.5$
	(S10)	$\mu_1 = \cdots = \mu_{10} = 1.5, \mu_{11} = \cdots = \mu_{20} = 0$
	(S11)	$\mu_1 = \cdots = \mu_5 = 2, \mu_6 = \cdots = \mu_{20} = 0$
	(S12)	$\mu_1 = 2.5, \mu_2 = \cdots = \mu_{20} = 0$
Decreasing correlation with $\rho = 0.2$	(S13)	$\mu_1 = \cdots = \mu_{20} = 1$
	(S14)	$\mu_1 = \cdots = \mu_{10} = 1.5, \mu_{11} = \cdots = \mu_{20} = 0$
	(S15)	$\mu_1 = \cdots = \mu_5 = 2, \mu_6 = \cdots = \mu_{20} = 0$
	(S16)	$\mu_1 = 2.5, \mu_2 = \cdots = \mu_{20} = 0$

Table 3. Parameter settings about means and covariance matrices.

90% is commonly adopted to choose PCs. However, this adoption is not always suitable since PCs with low contribution rate might be much more strongly correlated with the outcome than those with large contribution rate. To overcome this drawback, a mCPC method was developed recently[16]. In this study, we explored theoretical powers of mCPC deeply and find out that the maximum power of mCPC depends on the maximum noncentral

	S1–S4		S1	S2	S3	S4	S5–S8		S5	S6	S7	S8
i	λ_i	c_i (%)	Ω_i	Ω_i	Ω_i	Ω_i	λ_i	c (%)	Ω_i	Ω_i	Ω_i	Ω_i
1	16.2	81%	4.94	0.20	0.05	0.01	4.8	24%	1.67	2.34	0.59	0.04
2	0.2	82%	0.00	4.92	1.56	0.28	0.2	28%	0.00	0.00	0.00	0.00
3	0.2	83%	0.00	0.68	0.07	0.00	0.2	32%	0.00	3.61	0.19	0.21
4	0.2	84%	0.00	1.26	2.46	0.00	0.2	36%	0.00	0.02	1.31	0.00
5	0.2	85%	0.00	1.88	3.83	0.00	0.2	40%	0.00	1.74	3.09	4.29
6	0.2	86%	0.00	0.58	2.68	17.22	0.2	44%	0.00	0.45	0.25	0.00
7	0.2	87%	0.00	0.01	0.01	0.00	0.2	48%	0.00	0.00	0.16	0.00
8	0.2	88%	0.00	0.08	0.09	0.00	0.2	52%	0.00	0.65	0.00	0.00
9	0.2	89%	0.00	0.06	0.13	0.00	0.2	56%	0.00	0.2	0.16	0.00
10	0.2	90%	0.00	0.09	0.06	0.00	0.2	60%	0.00	0.38	0.16	0.00
11	0.2	91%	0.00	0.08	0.01	0.00	0.2	64%	0.00	0.28	0.05	0.00
12	0.2	92%	0.00	0.47	0.19	0.00	0.2	68%	0.00	2.11	1.47	0.00
13	0.2	93%	0.00	0.07	0.08	0.00	0.2	72%	0.00	0.80	0.09	0.00
14	0.2	94%	0.00	0.16	0.01	0.00	0.2	76%	0.00	0.62	0.49	0.00
15	0.2	95%	0.00	0.07	0.25	0.00	0.2	80%	0.00	1.39	0.14	0.00
16	0.2	96%	0.00	2.41	0.07	0.00	0.2	84%	0.00	0.80	2.48	0.12
17	0.2	97%	0.00	0.23	0.39	0.00	0.2	88%	0.00	0.12	0.23	0.00
18	0.2	98%	0.00	0.10	0.09	0.00	0.2	92%	0.00	0.10	0.23	0.00
19	0.2	99%	0.00	0.12	0.01	0.00	0.2	96%	0.00	0.00	0.00	0.00
20	0.2	100%	0.00	2.73	0.01	1.50	0.2	100%	0.00	0.79	0.06	0.13

Table 4. Eigenvalues, cumulative contribution rates and non-centrality parameters for Scenarios (S1) to (S8).

	S9–S12		S9	S10	S11	S12	S13–S16		S13	S14	S15	S16
i	λ_i	c_i (%)	Ω_i	Ω_i	Ω_i	Ω_i	λ_i	c (%)	Ω_i	Ω_i	Ω_i	Ω_i
1	7.23	36.1%	5.97	1.49	0.45	0.01	1.49	7.50%	11.60	6.51	1.25	0.01
2	4.32	57.8%	0.00	2.26	1.65	0.08	1.46	14.8%	0.00	6.60	3.54	0.05
3	2.45	70.0%	0.64	0.16	2.37	0.20	1.42	21.8%	1.25	0.70	4.31	0.11
4	1.47	77.4%	0.00	0.21	1.49	0.39	1.36	28.6%	0.00	0.00	2.95	0.19
5	0.96	82.1%	0.20	0.05	0.21	0.62	1.29	35.1%	0.42	0.24	1.06	0.27
6	0.67	85.5%	0.00	0.82	0.12	0.88	1.22	14.2%	0.00	0.79	0.11	0.37
7	0.50	88.0%	0.09	0.02	0.57	1.14	1.15	47.0%	0.20	0.11	0.00	0.45
8	0.38	89.9%	0.00	0.35	0.32	1.38	1.08	52.4%	0.00	0.01	0.01	0.53
9	0.31	91.4%	0.05	0.01	0.00	1.56	1.02	57.5%	0.10	0.06	0.02	0.58
10	0.26	92.7%	0.00	0.65	0.56	1.68	0.96	62.2%	0.00	0.31	0.24	0.61
11	0.22	93.8%	0.02	0.01	1.11	1.72	0.90	66.8%	0.06	0.03	0.45	0.62
12	0.19	94.7%	0.00	0.42	0.62	1.67	0.86	71.0%	0.00	0.03	0.33	0.60
13	0.17	95.6%	0.01	0.00	0.01	1.55	0.81	75.1%	0.03	0.02	0.07	0.55
14	0.15	96.3%	0.00	0.56	0.34	1.36	0.78	79.0%	0.00	0.17	0.01	0.48
15	0.14	97.0%	0.00	0.00	0.80	1.12	0.75	82.7%	0.01	0.01	0.07	0.39
16	0.13	97.7%	0.00	0.46	0.41	0.86	0.72	86.3%	0.00	0.05	0.44	0.30
17	0.12	98.3%	0.00	0.00	0.00	0.59	0.70	89.8%	0.00	0.00	0.00	0.21
18	0.12	98.9%	0.00	0.53	0.47	0.35	0.69	93.3%	0.00	0.11	0.10	0.12
19	0.11	99.4%	0.00	0.00	0.93	0.16	0.68	96.7%	0.00	0.00	0.18	0.06
20	0.11	100%	0.00	0.49	0.46	0.04	0.67	100%	0.00	0.07	0.09	0.01

Table 5. Eigenvalues, cumulative contribution rates and non-centrality parameters for Scenarios (S9) to (S16).

parameters of Chi-squared distributions for all PCs under the alternative hypothesis. However, it is difficult to obtain this information beforehand in practice. In view of this, we propose a novel and robust strategy to combine PCs. We also propose a test for genome-wide association studies and compare powers of this test to mCPC ($k_{0.8}$) and some other existing procedures such as Hotelling's T^2 test (HT), oPC ($k_{0.8}$) SSU and SKAT by extensive simulations. All simulation results show that our proposed procedure is more robust than mCPC, HT, oPC ($k_{0.8}$), SSU and SKAT. Results of real data analysis further demonstrates good performances of our proposed test. We suggest researchers to employ our robust strategy when they consider using principal component analysis method in the future.

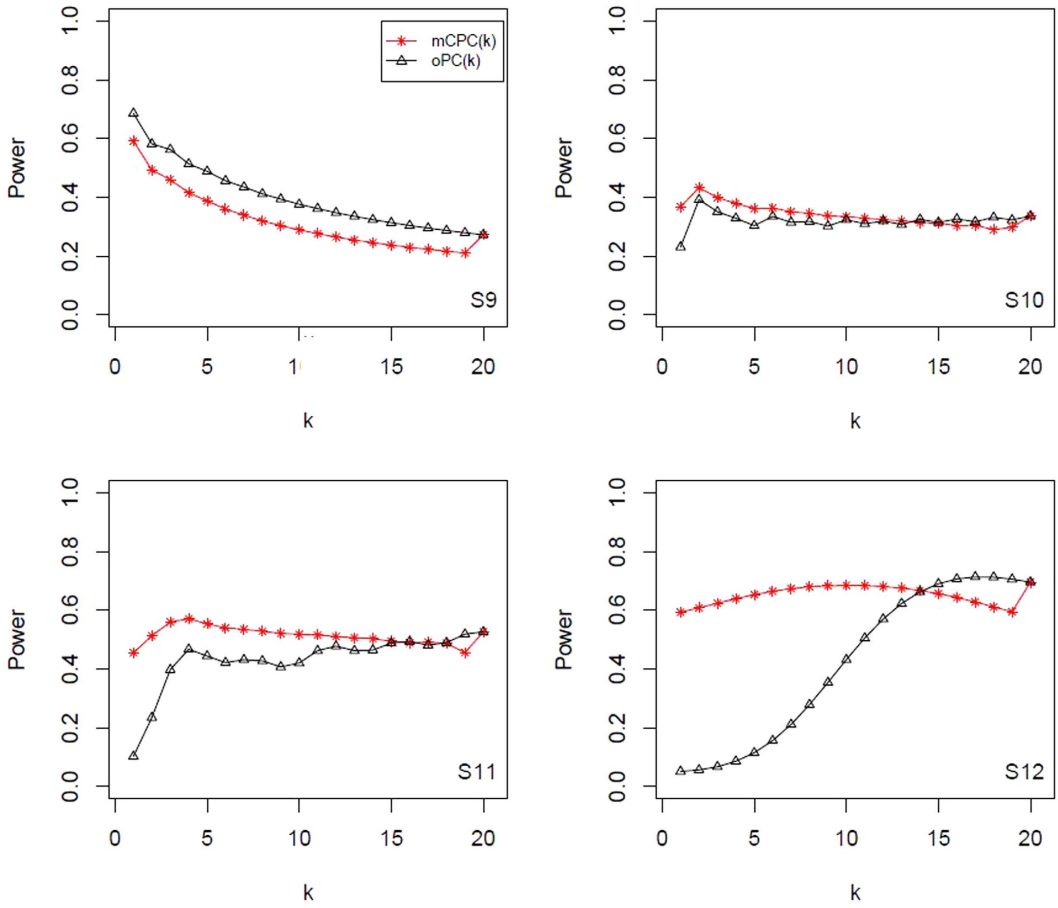

Figure 3. Powers of mCPC (k) and oPC (k) under significant level $\alpha = 0.05$ for Scenarios (S9) to (S12).

It should be noted that our proposed procedure is built upon test by Chatterjee *et al.*[18] which was designed to detect association between a marker and a binary trait. It can be easily extended to other application fields. For instance, it has been used in pleiotropic genetic study to identify deleterious genetic variants associated with multiple traits[16]. In addition, our proposed test can also be used to detect the association between genetic variants and quantitative traits in framework of linear model, and ordinal traits on basis of proportional odds model. If quantitative traits do not follow normal distribution, one can consider constructing a multivariate nonparametric trend test[22] and then employ our proposed strategy to combine them.

Methods

Maximum Powers of mCPC and ordinary PCA over extensive scenarios. For fixed m, the powers of mCPC (k) mainly depend on k, μ and $V_{m\times m}$. We set different mean vectors under the alternative hypothesis among different covariance matrices $V_{m\times m}$. We also consider two types of $V_{m\times m}$: one is that m-dimension variables are uniformly correlated, which means covariance matrix $V_{m\times m}$ is a symmetry positively definite matrix with diagonal elements all being 1 and non-diagonal elements all being ρ; the other is that all correlations among these m variables are decreasing considering the "physical" distance (SNP location), which means $V_{m\times m} = (\rho^{|i-j|})_{m\times m}$. Without loss of generality, ρ is chosen to be 0.8 for strong linkage disequilibrium and 0.2 for weak linkage disequilibrium. Here, we consider $m = 20$. Note that, a test based on ordinary PCA can be gained, which is denoted by oPC (k) with $\mathrm{oPC}(k) = \sum_{i=1}^{k} Z_i = F_k^{-1}[F_k(\sum_{i=1}^{k} Z_i)]$, where $k = 1, 2, \cdots, m$. Obviously, powers of oPC (k) are also affected by k when $\Omega_1, \cdots, \Omega_m$ are given. In order to view powers of mCPC (k) and oPC (k) comprehensively, we set α to be 0.05, and calculate powers of mCPC (k) and oPC (k) by numerical integration in R software under scenarios S1 to S16. All parameter settings about scenario S1 to S16 are displayed in Table 3. We calculate eigenvalues of $V_{m\times m}$, c_i, Ω_i of all scenarios S1 to S16 for $i = 1, \cdots, m$, and display all results in Tables 4 and 5. All power results of mCPC and ordinary PCA are displayed in Figs 1–4. Under the same correlation structure and mean vector, the powers of mCPC (k) and oPC (k) are affected strongly by selection of k. From Tables 3–5 and Figs 1–4, we can find that both the maximum powers of mCPC (k) and oPC (k) are related to the maximum non-centrality parameters of all PCs as k is from 1 to 20. For example, in Scenario S15, the non-centrality parameter of the second PC is the maximum among those of all PCs, and mCPC (2) has the maximum power. In another example, under the scenario S2, the non-centrality parameter of the third PC is the maximum, mCPC (4) has the maximum power, and powers of mCPC (3) and mCPC (4) are close. We can also find out that ordinary way to select k

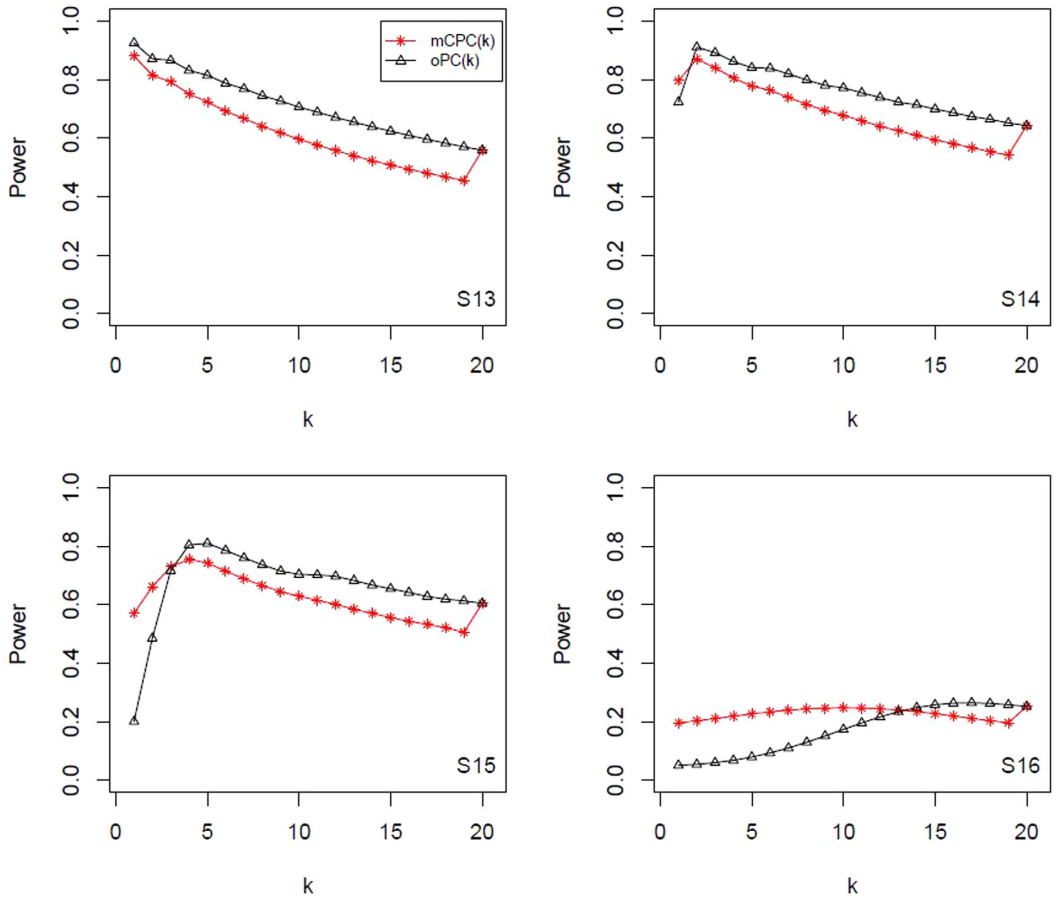

Figure 4. Powers of mCPC (k) and oPC (k) under significant level $\alpha = 0.05$ for Scenarios (S3) to (S16).

to construct mCPC (k) is not always desirable. For example, in Scenario S12, $k_{0.8} = 5$, but oPC (5) has power as low as 0.115. It is verified that mCPC is more desirable than ordinary PCA based on Figs 1–4. It is also verified that the selection of k according to the cumulative contribution rate is not robust. One can just follow the common adoption and choose $c = 80\%$ or 90% for oPC (k_c), but it will result in loss of power substantially under some situations. Furthermore, we draw a conclusion that mCPC (k_c) performs more robust than oPC (k_c) similar, since mCPC (k_c) has reasonable powers over all the considered scenarios. For example, in Scenario S2, these 20 variables are uniformly correlated with $\rho = 0.8$ and $\mu_1 = \cdots = \mu_{10} = 0.8$, $\mu_{11} = \cdots = \mu_{20} = 0$, the power of oPC ($k_{0.8}$) is 0.073, which is far less than the power of mCPC (k_c), which is 0.57.

A novel robust strategy to combine PCs. A further investigation of the maximum powers of mCPC (k) and oPC (k) shows that both of them are related to non-centrality parameters of the Chi-square distributions under the alternative hypothesis. For example, about scenario S16 in Table 5, the non-centrality parameters of all 20 PCs are 0.01, 0.05, 0.11, 0.19, 0.27, 0.37, 0.45, 0.53, 0.58, 0.61, 0.62, 0.60, 0.55, 0.48, 0.38, 0.30, 0.21, 0.12, 0.06 and 0.01 respectively, and non-centrality parameter being 0.62 which belongs to the 11th PC is the largest one among the non-centrality parameters of all 20 PCs. mCPC (10) takes the maximum power with 0.248 and the power of mCPC (11) is 0.247, which is very close to that of mCPC (10). The difference maybe are caused by numerical computing errors. The cumulative contribution rate of the top 10 PCs are 62.25%, which is much less than 80%. It is worth noted that the non-centrality parameters are determined by means and covariance matrix, which are hard to know in practice. Therefore, if we can know some prior information on means and covariance matrix, then the optimal strategy for selection of k become more prone to obtain. Aschard *et al.*[17] proposed to use mCPC (k) with k being determined by cumulative contribution rate of 80%.

As shown above, using 80% cumulative contribution rate might not be a robust strategy, and it will give a very low power in some cases (e.g., Fig. 1). According to numerical results in Scenarios S1 to S16, we propose to use the tCPC method to combine all PCs.

References

1. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* **56,** 18–31 (2003).
2. Xiong, M., Zhao, J. & Boerwinkle, E. Generalized T^2 test for genome association studies. *Am J Hum Genet* **80,** 1257–1268 (2002).
3. Fan, R. & Knapp, M. Genome association studies of complex diseases by case-control designs. *Am J Hum Genet* **72,** 850–868 (2003).

4. Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* **33,** 497–507 (2009).
5. Li, Z. B., Yuan, A., Han, G., Gao, G. M. & Li, Q. Rank-based tests for identifying multiple genetic variants associated with quantitative traits. *Ann Hum Genet* **78,** 306–310 (2014).
6. Wang, T. & Elston, R. C. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet* **80,** 353–360 (2007).
7. Tzeng, J. Y. & Zhang, D. Haplotype-based association analysis via variance-components score test. *Am J Hum Genet* **81,** 927–938 (2007).
8. Gauderman, W. J., Murcray, C., Gilliland, F. & Conti, D. V. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol* **31,** 383–395 (2007).
9. Wang, K. & Abbott, D. A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol* **32,** 108–118 (2008).
10. Zhang, F., Guo, X., Wu, S., Han, J., Liu, Y., Shen, H. & Deng, H. Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses. *PLos One* **7,** e53320 (2012).
11. Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. & Lin, X. H. Powerful SNP-Set analysis for case-control genome-wide association studies. *Am J Hum Genet* **86,** 929–942 (2010).
12. Ballard, D. H., Cho, J. & Zhao, H. Y. Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet epidemiol* **34,** 201–212 (2010).
13. Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* **35,** 606–619 (2011).
14. Hocking, R. R. The analysis and selection of variable in linear regression. *Biometrics* **32,** 1–49 (1976).
15. Jolliffe, I. T. A note on the use of principal components in regression. *J R Stat Soc Ser C* **31,** 300–303 (1982).
16. Aschard, H., Vilhjalmsson, B. J., Greliche, N., Morange, P. E., Tregouet, D. A. & Kraft, P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet* **94,** 662–676 (2014).
17. Sankaran, M. Approximations to the non-central Chi-square distribution. *Biometrika* **50,** 199–204 (1963).
18. Chatterjee, N., Chen, Y. H., Luo, S. & Carroll, R. J. Analysis of Case-Control association studies: SNPs, imputation and haplotypes. *Stat Sci* **24,** 489–502 (2009).
19. Amos, C. I., Chen, W. V., Seldin, M. F., Remmers, E. F., Taylor, K. E., Criswell, L. A., Lee, A. T., Plenge, R. M., Kastner, D. L. & Gregersen, P. K. Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proc* **3,** Suppl 7, S2 (2009).
20. Chen, L., Zhong, M., Chen, W. V., Amos, C. I. & Fan, R. A genome-wide association scan for rheumatoid arthritis data by Hotelling's T^2 tests. *BMC Proc* **3,** Suppl 7, S6 (2009).
21. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
22. Zhang, W. & Li, Q. Nonparametric risk and nonparametric odds in quantitative genetic association studies. *Sci Rep-UK* **5,** 12105 (2015).

Acknowledgements

Z. Li is partially supported by National Nature Science Foundation of China (No. 11401240,11471135), and the self-determined research funds of CCNU from the colleges' basic research of MOE (CCNU15A05038, CCNU15ZD011). D. Pan is partially supported by National Natural Science Foundation of China (No. 11301465), The Youth Program of Applied Basic Research Programs of Yunnan Province (No. 2013FD001) and the Young and Middle-aged Key Teachers Training Program of Yunnan University (No. XT412003). Q. Li is partially supported by National Nature Science Foundation of China, No. 61134013 and 11371353 and the Breakthrough Project of Strategic Priority Program of the Chinese Academy of Sciences, Grant No. XDB13040600.

Author Contributions

Z.L. and Q.L. conceived and designed the method and wrote the main manuscript text, W.Z. and D.P. conducted simulations. W.Z. contributed to the interpretation of all results. All authors reviewed the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Li, Z. *et al.* Power Calculation of Multi-step Combined Principal Components with Applications to Genetic Association Studies. *Sci. Rep.* **6**, 26243; doi: 10.1038/srep26243 (2016).