Revista da Estatística UFOP, Vol III(2), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS

MULTIVARIATE POLYGENIC MIXED MODEL IN ADMIXED POPULATION

Mariza de Andrade[13], Júlia Maria Pavan Soler [23]

Abstract: *In genome-wide association studies (GWAS) the Principal Component based Analysis (PCAs) provides a global ancestry value per subject, allowing corrections for population stratification. These coefficients are typically estimated assuming unrelated individuals and making use of dual-space properties to prevent high dimensional and sparse matrix problems. However, if family structure is present and is ignored, such sub-structure may induce artifactual PCAs. Considering the variable-space in high dimensional data set, extensions of the PCA have been proposed by Konishi and Rao (1992) taking into account only sibship relatedness and by Oualkacha et al. (2012) which can be applied to general pedigrees. Further, considering the subject-space, Blangero et al. (2013) obtained an Eigen simplification of the likelihood function from the univariate polygenic mixed model. In this work we propose to apply these methods to estimate the global individual ancestry using PCs extracted from different variance components matrix estimators and dual-space properties for subjects and variables. We use the GENOA sibship data consisting of European and African American subjects and the Baependi Heart Study consisting of 80 extended families collected from the highly admixture Brazilian population, both with SNPs data from Affymetrix 6.0 chip as applications. All the implementation are done using R package.*

1 Introduction

Studies of human complex diseases and traits associated with candidate genes are potentially vulnerable to bias (confounding) due to population stratification and inbreeding, especially in admixture population. In genome-wide association studies (GWAS) the Principal Components (PC) method provides a global ancestry value per subject, allowing corrections for population stratification (Price et al., 2006). However, these coefficients are

[1] Mayo Clinic - USA
[2] IME-USP. e-mail: pavan@ime.usp.br
[3] The authors thank Debashree Ray for your contribution to the computational implementation of the methodology as well as the Mayo Clinic (Rochester, MN, USA) and the Laboratory of Genetics and Molecular Cardiology - Heart Institute – University of Sao Paulo (SP, Brazil) for providing the data sets used in this study.

typically estimated assuming unrelated individuals and if family structure is present and is ignored, such sub-structure may induce artifactual PC.

Considering the variable-space in high dimensional data set, extensions of the PCA have been proposed by Konishi and Rao (1992) taking into account sibship relatedness. In order to combine a set of phenotypes in family-based studies, Ott and Rabinowitz (1999) introduced the principal components of heritability (PCH), which capture the familial information across phenotypes by calculating linear combinations of traits that maximize heritability. Within the PCH framework, Oualkacha et al. (2012) proposed an ANOVA estimate for the variance component matrices, which can be easily calculated to obtaining PCHs and are applied to general pedigrees and high dimensional family data. Further, considering the subject-space, Blangero et al. (2013) obtained an eigen simplification of the likelihood function from the univariate polygenic mixed model.

In this work, considering family-based studies and SNP (Single Nucleotide Polymorphism) data we apply these methods to estimate the global individual ancestry using PCs extracted from different variance components matrix estimators and dual-space properties for subjects and variables. For the application we use the GENOA sibship data consisting of European and African American subjects and the Baependi Heart Study consisting of 80 extended families collected from the highly admixture Brazilian population, both with SNPs data from Affymetrix 6.0 chip. All the implementations are done using R package.

2 Material and Methods

2.1. Multivariate family-based mixed model

Let y_{fj} be the ($n_f \times 1$) vector of responses for j-th variable evaluated on individuals of the f-th family ($f = 1, 2, ..., F;\ \ j = 1, 2, ..., p$). Consider the linear mixed model given by,

$$y_{fj} = \mu_{fj} + g_{fj} + e_{fj}, \tag{1}$$

where μ_{fj} is the ($n_f \times 1$) overall mean vector, g_{fj} is the ($n_f \times 1$) random effect vector defined as polygenic effect, and e_{fj} is the ($n_f \times 1$) error vector. The random components are

assumed to be uncorrelated with $E(u_{fj})=0$, $E(e_{fj})=0$ and covariance matrix $Cov(g_{fj})=2\Phi_f\sigma^2_{gj}$ and $Cov(e_{fj})=I_f\sigma^2_{ej}$, where σ^2_{gj} and σ^2_{ej} are the variance due the polygenic random effect and error components, respectively, associated to j-th variable. The ($n_f \times n_f$) matrix $2\Phi_f$ is the kinship matrix representing the expected identity by descend (index of relatedness) between members of the same family.

Let $Y_f=(y'_{f1}, y'_{f2},..., y'_{fp})'$ be a ($n_f p\times 1$) vector for all p variables and all members of the f-th family, with $E(Y_f)=1_f\otimes\mu_f$ and $Cov(Y_f)=2\Phi_f\otimes\Sigma_g+I_f\otimes\Sigma_e$, where 1_f is a ($n_f p\times 1$) unity vector, and Σ_g Σ_e are ($p\times p$) covariance matrix for p variables associated with polygenic and error component, respectively. For all F families we have $Y=(Y'_1, Y'_2..., Y'_F)'$ a ($Np\times 1$) vector containing all p variables for all individuals, $N=\sum_1^F n_f$ with $E(Y)=1_N\otimes\mu_f$ and $Cov(Y)=Diag(2\Phi_f)\otimes\Sigma_g+I_N\otimes\Sigma_e$, .f = 1,..,F. This framework represents the multivariate family-based model. In our application, the p variables represent the set of SNPs selected from the whole genome to estimate the global ancestry coefficients for individuals in family structures. Our challenge is to determine which covariance matrix, Σ_g or Σ_e, to choose to extract the PCs, to obtain the appropriate estimators for these covariance matrices, and how to handle the large number of variables (SNPs).

2.2. Principal component analysis for unrelated individuals

The methodology proposed by Price et al. (2006) to obtain the global ancestry coefficients extracts the PC from the eigenvalue decomposition of the ($N\times p$) matrix $X^P=(X_1^P, X_2^P,..., X_F^P)'$, where X_f^P is a ($n_f\times p$) matrix with the standardized values of SNPs genotype for f-th family. The method is based on classical PC optimization problem and is equivalent to obtain the PC of the spectral decomposition of correlation matrix obtained from the model in equation (1) but assuming independence between the subjects, i.e.,

$$y_{fj}=\mu_f+e_f, \tag{2}$$

such that, for $Y_{Np\times 1} = \left(y'_{11},...,y'_{F1},...,y'_{Fp}\right)'$, we have $E(Y) = 1_N \otimes \mu$ and $Cov(Y) = I_N \otimes \Sigma^P$. The covariance matrix Σ^P is estimated as

$$\hat{\Sigma}^P = \frac{1}{N-1}\sum_f \sum_i \left(y_{if} - \bar{y}\right)\left(y_{if} - \bar{y}\right)' = \frac{1}{N-1} S \tag{3}$$

with $S = \sum_f \sum_i \left(y_{if} - \bar{y}\right)\left(y_{if} - \bar{y}\right)'$, and $\bar{y} = \hat{\mu} = \frac{1}{N} Y'_{N\times p} 1_N$ as the overall mean. The upper symbol P represents the Price method. The global ancestry coefficients are obtained from the spectral decomposition of the correlation matrix $\hat{R}^P = D^P \hat{\Sigma}^P D^P$, with D^P as the diagonal matrix with elements $\sqrt{s_{jj}} = \sqrt{p_j\left(1-p_j\right)}$, $p_j = \frac{1+N\,\bar{y}_j}{2+2N}$, $j = 1,2,...,p$. The global ancestry coefficients can also be obtained from the singular value decomposition of the standardized X^P matrix, having the j–th column given by $X_j^P = \left(Y_j - \bar{y}_j\right)D^P$.

2.3. Principal component analysis for sibship data

Konishi and Rao (1992) extended the PCA to take into account the sibship relationship (equation (1)) and proposed ANOVA based-estimators for the covariance matrices with the PCs extracted from those matrices. For sibship of any size the kinship matrix is given by $2\Phi_f = 1_{n_f} 1'_{n_f}$ and using the model described in section 2.1, the estimators for the covariance matrices can be directly obtained from the classical ANOVA results, as given below

$$\hat{\Sigma}_e^K = \frac{S_w}{N-F}, \tag{4}$$

$$\hat{\Sigma}_g^K = N_0^{-1}\left(\frac{S_b}{F-1} - \frac{S_w}{N-F}\right) = \frac{S_b/(F-1) - S_w/(N-F)}{\left(N - \sum_f n_f^2/N\right)/(F-1)}, \tag{5}$$

with $N_0 = \frac{N - \left(\sum_f n_f^2/N\right)}{F-1}$, S_w and S_b as the sum of square and products matrices within and between families, respectively, with $S = S_w + S_b$, and $S = S^P$, the total sum of square and products matrix. These matrices can be written as

$$S_w = \sum_f \sum_i \left(y_{if} - \bar{y}_f\right)\left(y_{if} - \bar{y}_f\right)', \tag{6}$$

$$S_b = \sum_f n_f \left(\bar{y}_f - \bar{y}\right)\left(\bar{y}_f - \bar{y}\right)'. \tag{7}$$

By considering $\Sigma^K = \Sigma_g^K + \Sigma_e^K$ and using (4) and (5), their estimators are given by $\hat{\Sigma}^K = \hat{\Sigma}_g^K + \hat{\Sigma}_e^K$. One should keep in mind that the estimator of Σ^K does not correspond to the estimator $\hat{\Sigma}^P$. The PCs may also be obtained using the spectral decomposition of the matrices, $\hat{\Sigma}^K$, $\hat{\Sigma}_g^K$ and $\hat{\Sigma}_e^K$.

To obtain the PCs from the standardized data, one can use $\hat{R}_g^K = D^K \hat{\Sigma}_g^K D^K$ and $\hat{R}_e^K = D^K \hat{\Sigma}_e^K D^K$, with D^K as the diagonal matrix with elements $\left(s_{jj}^K\right)^{-1/2}$, s_{jj}^K being the diagonal elements of $\hat{\Sigma}^K$, $j = 1,2,...,p$. Similarly, these PCs can be obtained from the singular value decomposition of the standardized X^A matrix, with the j–th column represented by $X_j^A = \left(Y_j - \bar{y}_j\right)D^A$.

2.4. Principal component analysis for extended pedigrees data

When the multivariate family-based model described by (1) is extended for general pedigrees, Oualkacha et al. (2012) proposed to use ANOVA estimators for the variance component matrices. By using the matrices S_w e S_b given by equations (6) and (7), respectively, the estimators of the covariance matrices are written as

$$\hat{\Sigma}_g = \frac{S_b/(F-1) - S_w/(N-F)}{\left(\tau_c - \frac{\sigma_b}{N}\right)/(F-1) - (\tau_a - \tau_c)/(N-F)}, \quad (8)$$

$$\hat{\Sigma}_e^A = \frac{1}{N-F} S_w - \frac{(\tau_a - \tau_c)}{N-F} \hat{\Sigma}_g^A, \quad (9)$$

where $\tau_a = \sum_{f=1}^{F} \tau_a^{(f)}, \tau_b = \sum_{f=1}^{F} \tau_b^{(f)}, \tau_c = \sum_{f=1}^{F} \frac{1}{n_f} \tau_b^{(f)}, \tau_a^{(f)} = Tr\left(2\Phi^{(f)}\right), \tau_b^{(f)} = \sum_{j=1}^{nf} \sum_{k=1}^{nf} \Phi_{jk}^{(f)}$.

The upper symbol A indicates the methodology proposed by Oualkacha et al. (2012). When $2\Phi_f = 1_{n_f} 1'_{n_f}$, the estimators in (8) and (9) are the same estimators in (4) and (5),

proposed by Konishi and Rao (1992). The PCs can be obtained by spectral decomposition of the matrices, $\hat{\Sigma}^A$, $\hat{\Sigma}_g^A$ and $\hat{\Sigma}_e^A$, with $\hat{\Sigma}^A = \hat{\Sigma}_g^A + \hat{\Sigma}_e^A$.

Furthermore, we can also use the correspondent correlation matrices for (8) and (9) to calculate the PCs, where $\hat{R}_g^A = D^A \hat{\Sigma}_g^A D^A$ and $\hat{R}_e^A = D^A \hat{\Sigma}_e^A D^A$ are the decomposition of the standardized X^A matrix, with the j–th column given by $X_j^A = \left(Y_j - \bar{y}_j\right) D^A$, and D^A is a diagonal matrix with elements $\left(s_{jj}^A\right)^{-1/2}$, s_{jj}^A are the diagonal elements of $\hat{\Sigma}^A$, $j = 1,2,...,p$. Then the PCs can be obtained from the singular value $X_j^A = \left(Y_j - \bar{y}_j\right) D^A$.

Other way to determine the PCs for extended pedigrees it is to apply the Principal Component of heritability (PCH) proposed by Ott and Rabinowitz (1999) and also used by Oualkacha et al. (2012). The motivation behind this method is that instead of looking for the linear combinations of phenotypes with maximum variance, one should look for the linear combination of traits that maximizes its heritability, a measure which accounts for intra family correlations. In our case, it is to find combinations of the phenotypes (in our case, SNPs) that maximize the trace of the heritability matrix, $H_g^2 = \dfrac{\Sigma_g}{\Sigma_g + \Sigma_e}$. Thus, the maximization of $H_g^2(b) = \dfrac{b'\Sigma_g b}{b'(\Sigma_g + \Sigma_e)b}$ is equivalent to

$$\arg\max_{\|b\|=1} {}_b \frac{b'\hat{\Sigma}_g^A b}{b'\left(\hat{\Sigma}_g^A + \hat{\Sigma}_e^A\right)b}. \qquad (10)$$

To maximize (10) is equivalent to obtain the eigenvectors b such that $b'\hat{\Sigma}_e^A b = 1$, and $\max_b \dfrac{b'\hat{\Sigma}_g^A b}{b'\hat{\Sigma}_e^A b}$. To obtain the PCs in this case, we can use the eigenvectors of the generalized eigen system (Mardia et al., 1979). Since these calculations use high-dimensional and sparse matrices ($p >> N$), Wang et al. (2007) proposed a ridge penalized principal components approach to obtaining the PCs of heritability to accommodate large number of phenotypes. In this situation, the leading PC is defined as $PCH_\lambda = \arg\max_{\|\beta\|=1} \dfrac{b'\Sigma_g^A b}{b'\left(\Sigma_e^A\right)b + \lambda\|\beta\|^2}$, with λ is the regularization parameter to be specified. When λ = 0, the PCHλ is the original non-penalized leading PC (PCH). When $\lambda \to \infty$, the second term of the denominator of PCHλ dominates and the PCHλ approaches the linear combination that maximizes the between family

variation, Σ_g^A, i.e., $PCB = \arg\max_{\|\beta\|=1} b' \sum_b^A b$. When λ is between zero and infinity, the PCHλ changes between the PCH and the PCB.

2.5. GENOA Study

GENOA sibship data consists of European (Rochester, R) and African American (Jackson, J) subjects with SNPs data from Affymetrix 6.0 chip. For Rochester (Jackson), the screened data have 534 (548) families with 1,386 (1,263) individuals and data on 83,568 (50,510) SNPs. The two screened datasets include 9,224 common SNPs and 2,383 individuals (J:1079, R:1304) from 816 sibships (J:364, R:452). Detailed description of the sibship size and number of families are described in Table 1.

Table 1: Distribution of family size for the 2 populations

N_f	2	3	4	5	6	7	8	9	10	11	12	14	17
Jackson	191	92	38	15	17	4	4	2	–	–	–	1	–
Rochester	264	101	37	24	9	6	4	3	1	1	1	–	1

2.6. Baependi Heart Study

Baependi Heart Study consists of 119 extended Brazilian families with 1,712 individuals and SNPs data from Affymetrix 6.0 SNP chip. The screened dataset includes data from 80 families (1109 individuals) and 8,764 SNPs. Families with one individual or unrelated individuals with genotype data were excluded due to lack of information. Detailed description of the sibship size and number of families are described in Table 2.

Table 2: Distribution of family size for the Baependi Heart Study

N_f	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequency	2	8	6	8	5	1	4	6	5	6	1	5	3
N_f	16	18	19	21	24	27	32	46	48	60	61	68	93
Frequency	4	3	1	3	1	1	1	1	1	1	1	1	1

3 Results and Discussion

For sibship data (GENOA), Price et al. (2006) methodology showed very sensitive for data standardization (results not shown). The PCs unstandardized results had power to discriminate the two racial groups, but they were not sensitive to detect outlier families.

Applying Konishi and Rao (1992) methodology the decomposition of the matrix $\hat{\Sigma}_g^K$ was more powerful and robust comparing with the residual covariance matrix decomposition, $\hat{\Sigma}_g^K$, and provided similar results for $\hat{\Sigma}_t^K$ (Figure 1). In our data, this was due to the fact that the covariance matrix $\hat{\Sigma}_e^K$ is close to null matrix, $\hat{\Sigma}_g^K \cong \hat{\Sigma}^K$. When the PC of heritability (PCH) was used, we applied the penalization procedure to find the PCs. The results showed that the PCHλ (PC of heritability) had power to discriminate the groups (result not shown). The penalization parameter (λ) estimated using cross validation was equal to 100.

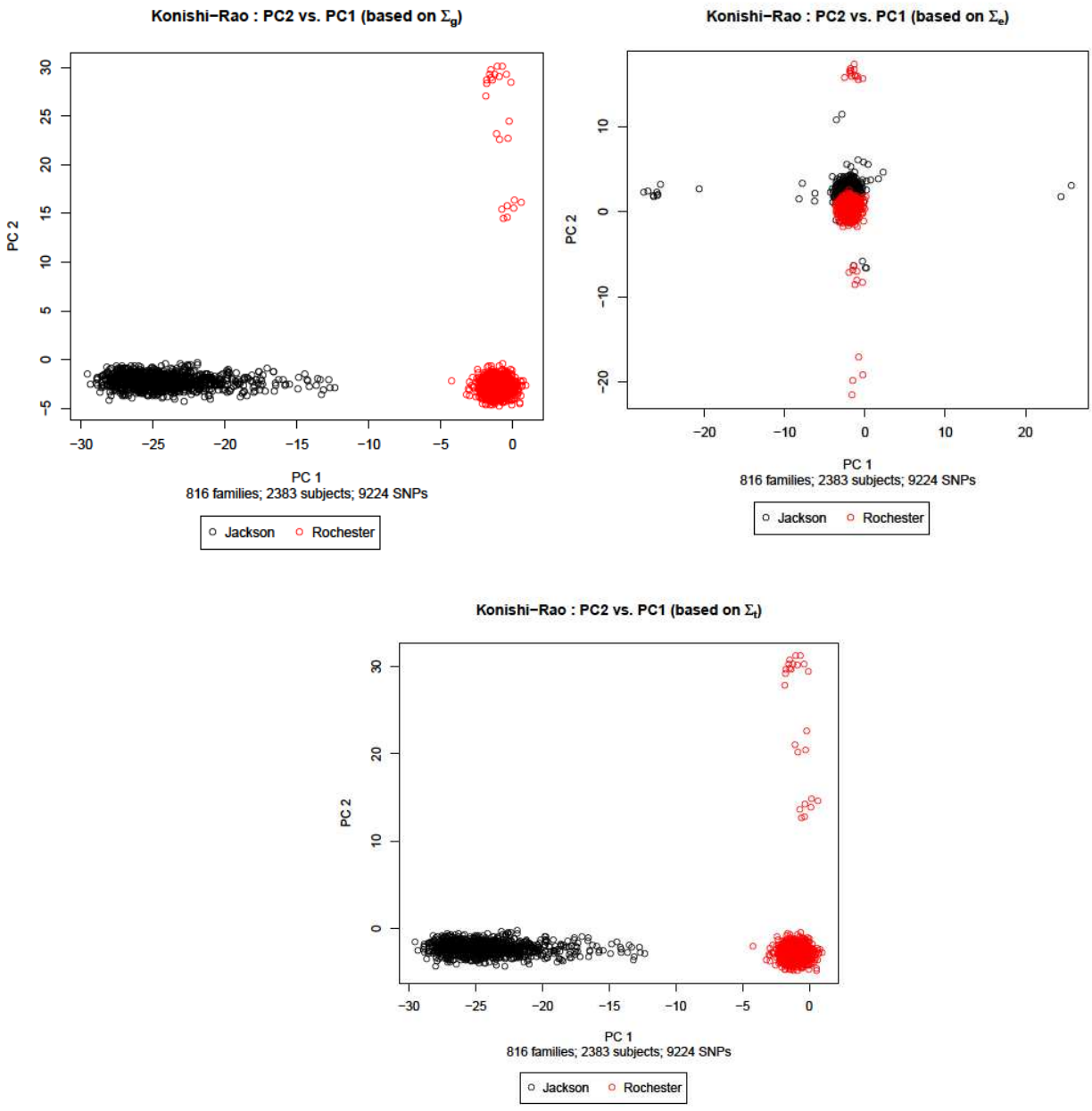

Figure 1: Konishi and Rao method for unstandardized data

Considering Baependi data and using Oualkacha method the population stratification was observed only for decomposition of $\sum_g$ but not for $\sum_e$ indicating the importance of the family relatedness (Figure 2). For standardized data we observed similar pattern when using Price and Oualkacha's method. We also observed that larger the family more admixture is present. For example, family 15 (family size = 60 – red color) is spread out over the 3 axes indicating the there are three potential racial admixture in this family. On the other hand, family 46 (family size = 4, yellow color) has three individuals from one race and the other classified as mixed race. One interesting point is that Price's method has more compressed axes than Oualkacha due to the inclusion of the family structure in the latter.

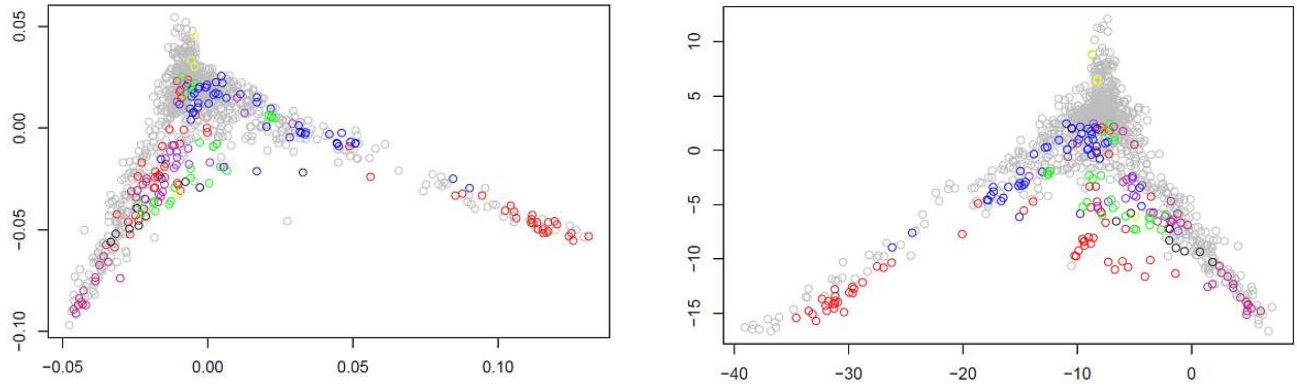

Figure 2: Distribution of the Brazilian families using Price (left) and Oualkacha's (right) methods.

4 Conclusion

Konishi and Rao approach showed that the discrimination of racial groups was independent of the standardization, imputation procedure and was not affected by the outlier families. For standardized data we observed similar pattern when using Price and Oualkacha's methods, but the former has more compressed axes. The admixture is better characterized by PCs from Oualkacha, which include family structure. We also observed that larger the family more admixture is present.

References

[1] Konishi, S and Rao, CR. Principal component analysis for multivariate familial data. *Biometrika* **79**: 631-641, 1992.

[2] Blangero J et al. A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees. *Adv. in Genetics* **8**, 2013.

[3] Ott, J, Rabinowitz D. A principal-components approach based on heritability for combining phenotype information. *Hum Hered* **49**: 106-111, 1999.

[4] Oualkacha, K, Labbe, A, Ciampi, A, Roy, MA and Maziade, M. Principal components of heritability for high dimension quantitative traits and general pedigrees. *Journal of Statistical Applications in Genetics and Molecular Biology* 11(**2**), Article 4, 2012.

[5] Price, AL et al. Principal Component analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(**8**): 904-909, 2006.
[6] Wang Y, Fang Y, Jin M. A ridge penalized principal-components approach based on heritability for high-dimensional data. *Hum Hered* **64**: 182-191, 2007.