Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 2 2012 *Article* 4

COMPUTATIONAL STATISTICAL METHODS FOR GENOMICS AND SYSTEMS BIOLOGY

Principal Components of Heritability for High Dimension Quantitative Traits and General Pedigrees

Karim Oualkacha, *University McGill*
Aurelie Labbe, *University McGill*
Antonio Ciampi, *University McGill*
Marc-Andre Roy, *Centre de recherche Université Laval Robert-Giffard*
Michel Maziade, *Centre de recherche Université Laval Robert-Giffard*

Recommended Citation:
Oualkacha, Karim; Labbe, Aurelie; Ciampi, Antonio; Roy, Marc-Andre; and Maziade, Michel (2012) "Principal Components of Heritability for High Dimension Quantitative Traits and General Pedigrees," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 2, Article 4.
DOI: 10.2202/1544-6115.1711

Principal Components of Heritability for High Dimension Quantitative Traits and General Pedigrees

Karim Oualkacha, Aurelie Labbe, Antonio Ciampi, Marc-Andre Roy, and Michel Maziade

Abstract

For many complex disorders, genetically relevant disease definition is still unclear. For this reason, researchers tend to collect large numbers of items related directly or indirectly to the disease diagnostic. Since the measured traits may not be all influenced by genetic factors, researchers are faced with the problem of choosing which traits or combinations of traits to consider in linkage analysis. To combine items, one can subject the data to a principal component analysis. However, when family date are collected, principal component analysis does not take family structure into account. In order to deal with these issues, Ott & Rabinowitz (1999) introduced the principal components of heritability (PCH), which capture the familial information across traits by calculating linear combinations of traits that maximize heritability. The calculation of the PCHs is based on the estimation of the genetic and the environmental components of variance. In the genetic context, the standard estimators of the variance components are Lange's maximum likelihood estimators, which require complex numerical calculations. The objectives of this paper are the following: i) to review some standard strategies available in the literature to estimate variance components for unbalanced data in mixed models; ii) to propose an ANOVA method for a genetic random effect model to estimate the variance components, which can be applied to general pedigrees and high dimensional family data within the PCH framework; iii) to elucidate the connection between PCH analysis and Linear Discriminant Analysis. We use computer simulations to show that the proposed method has similar asymptotic properties as Lange's method when the number of traits is small, and we study the efficiency of our method when the number of traits is large. A data analysis involving schizophrenia and bipolar quantitative traits is finally presented to illustrate the PCH methodology.

KEYWORDS: complex trait, heritability, linear discriminant analysis, principal component analysis, quantitative trait loci, variance components

1 Introduction

There is strong evidence that genetic mechanisms account for a large part of the etiology of many complex disorders, such as cardiovascular diseases, cancers, schizophrenia, autism and many others. Although genome scans have identified a number of candidate regions of interest for several complex diseases, replication of the results is often questionable. Foremost among the possible explanations for this is phenotype definition, a necessary prerequisite for establishing reliable genotype-phenotype relationships (Calum and Ramachandran, 2011). Since genetically relevant disease definition is still unclear for many complex disorders (Funalot et al., 2004), researchers tend to collect more and more items related directly or indirectly to the disease diagnosis. The number of items can range from a handful to a couple thousand, as in expression quantitative trait loci (eQTL) studies where gene expression measured on thousands of genes are used as quantitative traits in linkage analysis.

Multivariate variance-components linkage analysis using several correlated traits provides greater statistical power than single trait analysis to detect susceptibility genes in loci, as shown in Amos et al. (2001) and Almasy et al. (1998). However, since not all measured traits are necessarily influenced by genetic factors, researchers are faced with the problem of choosing which traits or combinations of traits to consider in linkage analysis. Typically, principal-component analysis (PCA) is used to combine traits into principal components and linkage analysis is based only on the first few of these. For example, Arya et al. (2003) used PCA to identify loci influencing the factors of insulin resistance syndrome (IRS)-related phenotypes using eight IRS-related phenotypes. When family data are collected, it is relevant to give larger weights to traits that have a larger degree of familial heritability in the combined traits, since they are more likely to be linked to genetic factors. Nevertheless, the data reduction achieved by principal-components analysis does not take into account family structure and heritability of traits.

In order to deal with the issues mentioned above, Ott & Rabinowitz (1999) introduced a new form of data reduction approach, which aims to capture the familial information carried by traits. They sought to calculate at a linear combination of traits that maximizes heritability, defined as the ratio of the family-specific (genetic) variation and the subject-specific (environmental) variation. Assuming that a trait is influenced by genetic factors, the idea is to give more weight to a trait showing similar values for subjects within the same family (smaller subject-specific variation). This linear combination was termed Principal Component of Heritability (PCH). Ott & Rabinowitz showed that using the first PCH as a quantitative trait in linkage analysis provided a substantial gain in power as compared to the use of first principal component from a standard PCA. However, when the number of traits is large, the PCH approach is not directly applicable and the components are unidentifiable and unstable. Wang et al. (2007a) proposed a penalized principal-components approach based on heritability (PCH_λ) that can be applied to high dimensional family data. By simulation studies, Wang et al. (2007a) showed that the traits combined by the PCH_λ approach have significant power gain in linkage analysis compared to the usual PCA and PCH.

To calculate the PCH and PCH_λ components, one needs to estimate the variance covariance matrices of the quantitative trait model, which can be decomposed as the subject-specific component of variance Σ_e and the family-specific component of variance Σ_g. Ott & Rabinowitz proposed the usual sample within-family variance-covariance matrix as estimate of Σ_e and a new estimate for Σ_g. Wang et al. (2007a) used the same estimate of Σ_e and the sample between-family variance-covariance matrix to estimate Σ_g. When dealing with family-members of the same type (e.g.: all the family members are siblings of each other), both approaches give efficient estimates of the variance components, and they validly take into account the family structure. However, this is not the case when more complex families are sampled in the study, since the correlation between two individuals is assumed to be the same for each pair in the family.

Wang et al. (2007a) stated without giving details that it is possible to extend the PCH approach to general pedigrees with more complex structures (see Figure 1 for an example of more complex structure). However, the estimation of the variance components in the general case (pedigrees of unequal sizes and describing complex family relationships) is by no means an easy task. While in the case considered by Ott & Rabinowitz, the data was balanced, making the estimation of variance components relatively straightforward, in the general case we deal with multivariate unbalanced data (unequal subclass numbers). The estimation of variance components from multivariate unbalanced data has received little attention in the literature. The ANOVA, the maximum likelihood (ML) and the Restricted Maximum Likelihood (REML, Calvin, 1993) estimators of variance components are three standard approaches for the estimation of the components Σ_g and Σ_e. The maximum likelihood approach in standard linear mixed models is discussed by Searle et al. (1992) and McCulloch & Searle (2002). The ML and REML estimators can be negative with positive probability in the unbalanced case. Similarly, the ANOVA estimators in the unbalanced case are limited since they can also lead to negative values and lose some of their properties, as illustrated in McCulloch & Searle (2002).

In multivariate genetic linkage analysis, the state-of-the art estimators of variance components are Lange's ML estimators implemented in Mendel software (Lange et al., 2001; 2006). These estimators are computed in most of the widely used computer programs performing quantitative trait genetic analyses based on variance components (eg: Solar (Almasy & Blangero, 1998)). In animal genetic studies, the REML estimators of the variance components for multivariate unbalanced genetic mixed linear models are usually computed using the specialized ASREML software (Gilmour et al., 1999); this approach can easily be adapted to the estimation of human genetic models as well. The desirable features of Lange's ML and ASREML's estimators are their excellent theoretical properties with respect to bias and mean squared error, provided that multivariate normality can be assumed. However, the equations defining these estimators do not have in general a closed form solution, but have to be solved by heavy numerical computations. Furthermore, both methods have problems dealing with a large number of traits.

The objectives of this paper are the following: i) to review some standard strategies available in the literature to estimate variance components for unbalanced data in mixed models; ii) to propose an efficient ANOVA method for a genetic random effect model to estimate the variance components, which can be applied to general pedigrees and high dimensional family data within the PCH framework; iii) to elucidate the connection between PCH analysis and Linear Discriminant Analysis.

The paper is organized as follows: Section 2 reviews the genetic variance components model. Section 3 presents the principal component of heritability and studies the relationship between the linear discriminant analysis and the PCH approach. Section 4 proposes some ANOVA estimators of the variance components accounting for the family correlation structure and also reviews current maximum likelihood estimation strategies. In Section 5, we perform a simulation study to compare the properties of the estimators discussed here and to assess their performance in linkage analysis. Finally, our proposed method will be applied to a dataset on schizophrenia and bipolar disorder.

2 Variance component model

The variance component model for genetic quantitative traits aims to partition the phenotype variation into components attributable to shared genes and shared environment. As in Ott & Rabinowitz, these effects can be viewed as a family-specific component G and a subject-specific component E respectively. Let Y_{ij} be the vector of r traits for an individual j ($j = 1, \dots, n_i$) of family i ($i = 1, \dots, m$). We represent the model as

$$Y_{ij} = \mu + G_{ij} + E_{ij}, \tag{1}$$

where μ is a vector of dimension r representing the overall mean of all traits, with the random effects G and E assumed to be mutually independent and normally

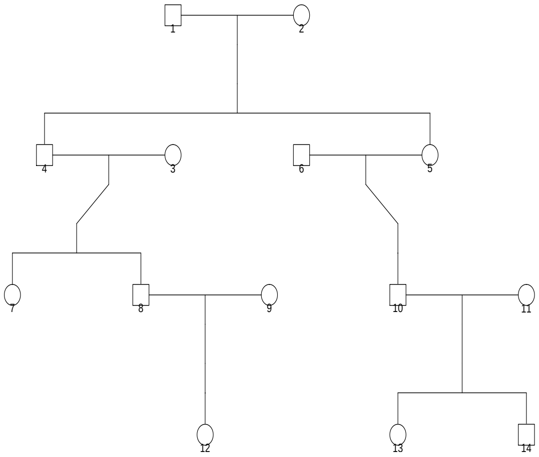

Figure 1: Example of a graphical representation of a general pedigree. Squares represent males, circles represent females.

distributed. We write this as follows:

$$G_{ij} \sim \mathcal{N}(0_r, \Sigma_g), \quad E_{ij} \sim \mathcal{N}(0_r, \Sigma_e),$$

where 0_r represents a vector of zeros of dimension r, Σ_g is an $r \times r$ nonnegative definite matrix and Σ_e is an $r \times r$ positive definite matrix. Thus, the variance of Y_{ij} is given by

$$Var(Y_{ij}) = \Sigma_g + \Sigma_e.$$

The contribution of the alleles shared identical by descent (IBD) in the covariance between individuals of the same family can be written as

$$Cov(Y_{ij}, Y_{ik}) = Cov(G_{ij}, G_{ik}) = 2\Phi_{jk}^{(i)}\Sigma_g, \quad j,k = 1,\ldots,n_i, j \neq k,$$

where $\Phi^{(i)}$ is the $n_i \times n_i$ matrix of kinship coefficients of family i (see Lange, 2002, Chap 5). The coefficient of kinship between two individuals j and k, Φ_{jk}, is the probability that two alleles sampled at random from each individual are identical by descent. For instance, $\Phi_{jj} = 1/2$ for all subjects j, $\Phi_{jk} = 1/4$ if j and k are siblings or if one of them is a parent of the other, $\Phi_{jk} = 1/8$ if one of them is a grand-parent of the other and $\Phi_{jk} = 1/16$ if j and k are cousins. Thus, the further apart individuals are in the family, the smaller the contribution of their alleles IBD in the covariance matrix. Note that the matrix $\Phi^{(i)}$ is known if the pedigree structure is known and can be directly computed using existing R libraries.

Assuming that families are independently sampled, the model (1) can be rewritten as

$$\begin{aligned} \mathbf{Y_i} &\sim \mathscr{MN}\left(1\!\!1_{n_i} \otimes \mu, 2\Phi^{(i)} \otimes \Sigma_g + I_{n_i} \otimes \Sigma_e\right), \\ \mathbf{Y} &\sim \mathscr{MN}\left(1\!\!1_n \otimes \mu, 2\Psi \otimes \Sigma_g + I_n \otimes \Sigma_e\right), \end{aligned} \tag{2}$$

where

$$\mathbf{Y}_i = (Y_{i1}^T, \ldots, Y_{in_i}^T)^T, \quad \mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_m^T)^T, \quad \Psi = diag\{\Phi^{(i)},\ i = 1, \ldots, m\}, \quad n = \sum_i n_i, \tag{3}$$

$\mathbf{Y}_i$ is the vector of the $n_i r$ measures for the individuals of family i and $\mathbf{Y}$ is the vector of the nr measures for all the individuals. The notations $1\!\!1_{n_i}$, I_{n_i} and $\otimes$ refer to the vector of 1's of dimension n_i, the $n_i \times n_i$ identity matrix and the Kronecker product respectively. Note that the model of equation (1) is a general version of the Quantitative Trait Locus (QTL) model. To detect linkage between QTLs and markers of the human genome, equation (1) can be specified so as to decompose the genetic effect G into an effect of a major locus or several loci, an effect of a residual polygenic, and possibly additional terms specifying interactions between genetic and environmental effects, see Almasy & Blangero (1998).

3 Principal components of heritability

3.1 General approach

Let Ω be a $r \times r$ matrix and $f(\Omega, .)$ defined by

$$f(\Omega, \beta) = \beta^T \Omega \beta, \quad \beta \in \Re^r, \quad ||\beta|| = 1.$$

The range of the function $f(\Omega, .)$, is the segment $[a, b] \in \Re$, where a and b are the smallest and largest eigenvalue of Ω, respectively. In what follows, we will discuss a number of data reduction approaches that are particular cases of the optimization problem of the function $f(\Omega, .)$ for various choices of Ω.

Suppose now we want to reduce the number of traits r. If we were not interested in the within-family correlations, we could apply Principal Component

Analysis (PCA) to the data matrix $\mathbf{Y}$. We recall that the first Principal Component (PC) is the linear combination of traits that maximizes the total variation, it is therefore the solution of a classic optimization problem:

$$PCA = \arg\max_{\beta} f(\Omega, \beta), \quad \Omega = \Sigma_g + \Sigma_e,$$

where β represents the weights associated to the r traits. Similarly, one would extract the second PC by repeating the maximization on the orthogonal complement of the first PC, and so on, until one obtains an orthogonal basis in $\Re^r$. Principal Components were used in the literature to build composite phenotypes for linkage analysis, in spite of the fact that their definition ignores an essential aspect of genetic data: the within family correlation structure. In an attempt to correct this problem, Ott & Rabinowitz introduced the notion of Principal Component of Heritability (PCH). Instead of looking at the linear combination of traits with maximum variance, the authors suggested to look for the linear combination of traits which maximizes its *heritability*, a quantity which explicitly accounts for intra family correlations. The heritability of a linear combination of traits is defined as

$$h(\beta) = \frac{\beta^T \Sigma_g \beta}{\beta^T (\Sigma_g + \Sigma_e) \beta}. \tag{4}$$

One can verify that the maximization of $h(\beta)$ is equivalent to the maximization of $f(\Omega, .)$, with $\Omega = \Sigma_e^{-1}\Sigma_g$, hence we can write

$$\text{PCH} = \arg\max_{\beta} f(\Sigma_e^{-1}\Sigma_g, \beta). \tag{5}$$

Thus, the β which maximizes equation (4) is the first eigenvector of the matrix $\Sigma_e^{-1}\Sigma_g$ (Mardia et al, 1979). As in PCA, one can extract from Ω a sequence of mutually orthogonal linear combinations, and retain only those corresponding to non-negligible eigenvalues.

For high dimensional data, PCH analysis encounters the so-called small sample size problem. This problem arises, in familial data, whenever the number of families is smaller than the number of traits. For instance, in the microarray gene expression phenotypes used in genome-wide linkage analysis in Morley et al. (2004) to find evidence of linkage to specific chromosomal regions, there were $r = 3554$ phenotypes (traits) measured on only 14 families. Under these circumstances, many eigenvalues of Σ_e may be estimated erroneously as zero. Indeed, $f(\Sigma_e^{-1}\Sigma_g, \beta)$ is maximal for any β satisfying $\Sigma_e \beta = 0_r$, and so the maximum is not identifiable unless $\beta^T \Sigma_g \beta$ is also maximized. Due to this difficulty, Wang et al.

(2007a) proposed a ridge regularization of the PCH approach and defined the PCH as

$$\mathrm{PCH}_\lambda = \arg\max_\beta f(\Omega_\lambda, \beta), \quad \Omega_\lambda = (\Sigma_e + \lambda I_r)^{-1}\Sigma_g,$$

where λ is the regularization parameter. Note that when λ is equal to zero, one has PCH_λ =PCH, and when λ tends to infinity, the PCH_λ approaches the linear combination that maximizes the between-family variation. The regularization parameter can be determined by the data through cross-validation (Wang et al., 2007a).

3.2 Linear discriminant analysis (LDA) and relationship with the PCH analysis

In this section we will show that in the special case of siblings data, PCH analysis is equivalent to performing Fisher's Linear Discriminant Analysis (LDA). Discriminant analysis is used when subjects with multivariate measurements are partitioned into known classes and it is believed that the distribution of the multivariate measurements depends on class memberships. LDA was introduced by Fisher (1936) as an exploratory technique. The objective of LDA is to identify directions in variable space that best separate classes. In our context, variables are traits and classes are families. Specifically, LDA seeks a projection β that maximizes the ratio of between-class scatter S_b against within-class scatter S_w (Fisher's criterion):

$$\arg\max_{||\beta||=1} \frac{\beta^T S_b \beta}{\beta^T S_w \beta}, \tag{6}$$

where the $r \times r$ matrices S_w and S_b are defined as

$$S_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(Y_{ij} - \bar{Y}_i)^T, \quad S_b = \sum_{i=1}^{m} n_i(\bar{Y}_i - \bar{Y})(\bar{Y}_i - \bar{Y})^T, \tag{7}$$

with $\bar{Y}_i = \sum_j Y_{ij}/n_i$, $\bar{Y} = \sum_{i,j} Y_{ij}/n$, and $n = \sum_i n_i$. As is well known, LDA also reduces to a generalized eigenvalue-eigenvector problem, and it produces eigenvectors known as Canonical Discriminant Components (CANDISC) or discriminant axes. It is not difficult to see that in the genetic context one can explain LDA as follows: assume we consider traits that are influenced by familial (e.g. genetic) factors, then a trait which varies little within the same family (trait with small within-class scatter) should receive greater weight in a CANDISC than a trait with large within family scatter. Note that one can easily verify that the maximization in (6) is equivalent to maximizing $f(S_w^{-1}S_b, .)$. Moreover, if we let Σ_b and Σ_w be the population between and within variance-covariance matrices of the data' population respectively, then one can write

$$\text{CANDISC} = \arg\max_{\beta} f(\Sigma_w^{-1}\Sigma_b, \beta). \tag{8}$$

Therefore, the principal components of heritability are exactly the CANDISC when we assume that $\Sigma_g = \Sigma_b$ and $\Sigma_e = \Sigma_w$.

When working with multiple siblings data, the correlation between any two siblings of the same family is the same. In this case, the variance-covariance matrix of $\mathbf{Y}$ given in (3) can be written as

$$Var(\mathbf{Y}) \quad = \quad 1\!\!1_n 1\!\!1_n^T \otimes \frac{\Sigma_g}{2} + I_n \otimes (\Sigma_e + \frac{\Sigma_g}{2}).$$

Thus, the model given in (1) can be reduced to a standard one way random effect ANOVA model with $\tilde{\Sigma}_b = \Sigma_g/2$ and $\tilde{\Sigma}_w = \Sigma_e + \Sigma_g/2$. And so, one can easily verify that maximizing $h(\beta)$ given in (4) is equivalent to maximizing $f(\tilde{\Sigma}_w^{-1}\tilde{\Sigma}_b, \beta)$. To see how the PCHs can be used to discriminate families, one can notice that by maximizing heritability, the PCH framework also gives larger weights to heritable traits. Such traits are expected to be similar within families and therefore should have small within-class scatter. Thus, unless the total scatter matrix (sum of within- and between-class scatter) is negligible, the between-class scatter is large for these traits. Therefore, the PCH can be implicitly viewed as a search algorithm for projections discriminating classes (families).

4 Estimation of the variance components

When the data come from families with more complex structures than multiple siblings (as it is the case in most human genetic studies), the estimation of the variance components under a multivariate traits model is not an easy task. In this section, we see how ANOVA estimators can be obtained, by looking at the genetic model (1) as a multivariate unbalanced one way random effect model. Maximum likelihood methods are also reviewed.

4.1 ANOVA estimators of V.C. accounting for familial dependence

The ANOVA estimators of the VC under an unbalanced one way random effect ANOVA model are given in Searle (1971), Swallow & Monahan (1984) and Searle

(1992). However, these estimators do not take into account the familial structure in the data. When quantitative traits data are collected from general pedigrees, the expectations of the statistics S_b and S_w given in (7) are summarized in Proposition 1, which is proved in the Appendix.

Proposition 1 *Under the model (1), if we let S_b and S_w be defined as in (7) then one has*

$$I\!\!E(S_w) = (n-m)\Sigma_e + (\tau_a - \tau_c)\Sigma_g, \qquad I\!\!E(S_b) = (m-1)\Sigma_e + (\tau_c - \frac{\tau_b}{n})\Sigma_g, \quad (9)$$

where

$$\tau_a = \sum_{i=1}^{m} \tau_a^{(i)}, \quad \tau_b = \sum_{i=1}^{m} \tau_b^{(i)}, \quad \tau_c = \sum_{i=1}^{m} \frac{1}{n_i}\tau_b^{(i)}, \quad \tau_a^{(i)} = 2Trace\left[\Phi^{(i)}\right], \quad \tau_b^{(i)} = 2\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} \Phi_{jk}^{(i)}.$$

with $\Phi^{(i)}$ being the kinship matrix of the i th family.

Equating the two statistics S_b and S_w to their expectations given in Proposition 1 gives the following ANOVA estimates:

$$\hat{\Sigma}_g^A = \frac{S_b/(m-1) - S_w/(n-m)}{(\tau_c - \frac{\tau_b}{n})/(m-1) - (\tau_a - \tau_c)/(n-m)}, \qquad \hat{\Sigma}_e^A = \frac{1}{(n-m)}S_w - \frac{(\tau_a - \tau_c)}{(n-m)}\hat{\Sigma}_g, \ (10)$$

where τ_a, τ_b and τ_c are given in Proposition 1. Note that for the unbalanced one-way random effect ANOVA model the total variance's decomposition is orthogonal, (i.e. $S_t = S_b + S_w$), where S_t is the total scatter Sum of Squares matrix. This leads to independent sums of squares and competent and unique ANOVA estimators (see for instance Swallow & Monahan, 1984, for more details). Note also that the ANOVA estimators we obtained from Proposition 1 are linear combination of S_b and S_w. Similarly, in the case of siblings data, Ott & Rabinowiz (1999) proposed to use:

$$\hat{\Sigma}_g^{(O)} = S_t/(n-1) - S_w/(n-m), \qquad \hat{\Sigma}_e^{(O)} = S_w/(n-m). \qquad (11)$$

In fact, the next proposition shows that the principal components of heritability obtained from the ANOVA estimators and Ott's estimators are identical; moreover, they coincide with the LDA discriminant axes.

Proposition 2 *Let the principal components of heritability be defined by (5). If the estimators of Σ_g and Σ_e are linear combinations of S_b and S_w then the corresponding PCH are exactly the discriminant axes.*

Proof: To prove Proposition 2, let $\hat{\Sigma}_g$ and $\hat{\Sigma}_e$ be defined as

$$\hat{\Sigma}_g = \alpha_g S_b + \delta_g S_w, \quad \hat{\Sigma}_e = \alpha_e S_b + \delta_e S_w.$$

Let β be a discriminant axis obtained from (6) and let η be the eigenvalue of $S_w^{-1} S_b$ associated to β, i.e. $S_w^{-1} S_b \beta = \eta \beta$. It follows that:

$$\begin{aligned}
\hat{\Sigma}_e^{-1} \hat{\Sigma}_g \beta &= (\alpha_e S_b + \delta_e S_w)^{-1} (\alpha_g S_b + \delta_g S_w) \beta \\
&= (\alpha_e S_w^{-1} S_b + \delta_e I_r)^{-1} S_w^{-1} S_w (\alpha_g S_w^{-1} S_b + \delta_g I_r) \beta \\
&= (\alpha_g \eta + \delta_g) \left(\alpha_e S_w^{-1} S_b + \delta_e I_r\right)^{-1} \beta \\
&= \frac{\alpha_g \eta + \delta_g}{\alpha_e \eta + \delta_e} \beta.
\end{aligned}$$

Hence $\hat{\Sigma}_e^{-1} \hat{\Sigma}_g$ has β as an eigenvector. Note that the PCH are obtained as the eigenvectors of $\Sigma_e^{-1} \Sigma_g$, and so, the PCH obtained from their estimators of Proposition 2 are exactly the discriminant axes.

As we mentioned, both our ANOVA estimators and Ott's estimators are linear combinations of S_b and S_w. In view of Proposition 2, this implies that their associated PCH are equal to the discriminant axes. However, the order of these PCH is not necessarily the same since their associated eigenvalues are not equal to those associated with the discriminant axes. Note that the order of the PCH's leads to important consequences when using only the first ones in linkage analysis.

4.2 ML and REML estimators of V.C. under the quantitative genetic model

Consider now the multivariate variance component genetic model given by Lange (2002, Chap 8). If $\mathbf{Y}$ is defined as in (3) and $\tilde{\mu} = 1\!\!1_n \otimes \mu$, then the ML and REML estimates of μ, Σ_g and Σ_e can be obtained using the scoring algorithms, which are implemented in the Mendel and ASREML software respectively, (Lange et al., 2001; 2006, Gilmour et al., 1999). As for the ANOVA method, the REML algorithm searches for translation invariant maximum likelihood VC estimators (i.e. estimators which don't involve μ). By default, ASREML estimators are calculated for unbalanced animal genetic mixed linear models. However, we could adapt this approach to human genetic models as well. Although REML estimators are arguably preferable on a theoretical basis to ML estimators, the Lange's ML estimators are still considered state-of-the art in multivariate human linkage analysis . Both ML and REML estimators share two highly desirable asymptotic properties : under the

DOI: 10.2202/1544-6115.1711

assumption of multivariate normality, they are consistent and asymptotically efficient. However, they have the disadvantage of requiring heavy computer resources. For example, when r is large ($r > 30$), Mendel software runs into severe memory problems and fails to complete the calculations of the ML estimators. Similarly ASREML has a limitation of $r \leq 20$.

5 Simulations

In this section we describe and report the results of some simulation studies we have carried out with the aim to evaluate and compare several estimators of the variance components. In practice, ANOVA's, MLE, REML's and Ott's estimators may be negative with non-zero probability. In view of this, we have adopted the procedure used by Amemiya (1985) to make these estimators nonnegative definite (n.n.d). This is achieved by replacing in the corresponding spectral decomposition any negative eigenvalue by zero. This approach was also used by Mathew et al. (1994) and Srivastava & Kubokawa (1999) to make their V.C. estimators n.n.d. Furthermore, to compute the PCH_λ, the regularization parameter λ was chosen by the cross-validation optimal criterion given in Wang et al. (2007a).

In all simulations, the basic genetic model has a single disease susceptibility locus with two alleles denoted by d and D with frequencies p and q respectively (d is the susceptible allele). The following model, which is used to generate traits for general pedigrees, was also used by Ott & Rabinowitz (1999) and Wang et al. (2007a) to simulate quantitative traits data under a PCH framework. It is given by

$$Y_{ij} = X_{ij}\mu + E_{ij}, \tag{12}$$

where X_{ij} is the number of the disease susceptible alleles carried by the $j-$th subject in the $i-$th family, $\mu \in \Re^r$ is the effect of the susceptible allele and E_{ij} represents the environmental component which is normal with mean zero and variance-covariance Σ. Thus, the effect of allele d on the traits is assumed to be additive (i.e. carrying one copy of d adds μ to the mean of the traits). The effect μ is the same for each family, but is different across traits.

5.1 Efficiency of the proposed ANOVA estimator for variance components

We relied on the simulation setting used by Ott & Rabinowitz (1999) and Wang et al. (2007a) with $r = 5$ traits. This setting assumes that the genetic effect on the first two traits is smaller than the last three, but the subject-specific (i.e. non genetic) effect on the first two traits is larger. Specifically, this leads to:

$$\mu = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{pmatrix}, \quad \Sigma_e = \begin{pmatrix} 3.0 & 0.5 & 0 & 0 & 0 \\ 0.5 & 3.0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 1 \end{pmatrix}.$$

To simulate data with an increasing number of traits ($r = 10, 15, 50, 100$), we increased the size of the variance-covariance matrix Σ_e and the genetic effect μ. The expanded effect vector of the genetic effect for r traits ($r > 5$) is $\mu = (1,1,2,2,2,0,\ldots,0)_r^T$, which implies that the first five traits are influenced by a single genetic locus and the other components are random noise with no genetic effect. Similarly, the expanded variance-covariance matrix of the multivariate gaussian random vector is

$$\begin{pmatrix} \Sigma_e & 0 & 0 \\ 0 & \Sigma_s & 0 \\ 0 & 0 & I_{(r-10)} \end{pmatrix},$$

where Σ_s is a 5×5 matrix with 1 in diagonal and 0.1 out of diagonal. Thus, it means that the first five non-genetic traits are correlated and the remaining $(r-10)$ non-genetic traits are independent. We generated 100 families: 55 families with two parents and one child, 35 families with two parents and two children and 10 families with two generations (8 subjects per family). The number of replications for the simulation was $B = 200$.

In order to evaluate the efficiency of our proposed ANOVA estimator, we used risks and squared bias as evaluation criteria. Following Sun et al. (2003), the risk and the squared bias of an estimator $\hat{\Sigma}$ are defined respectively as

$$R(\hat{\Sigma}, \Sigma) = tr\left\{\mathbb{E}\left[(\hat{\Sigma} - \Sigma)^2\right]\right\}, \qquad sb(\hat{\Sigma}, \Sigma) = tr\left\{\left[\mathbb{E}(\hat{\Sigma}) - \Sigma\right]^2\right\}. \tag{13}$$

The true values of the VC component can be computed under the model (12) (see supplementary material, web appendix A): the family-specific component Σ_g and the subject-specific component Σ_e are given by

$$\Sigma_g = 2pq\mu\mu^T, \quad \Sigma_e = \Sigma. \tag{14}$$

Furthermore, in order to compute the risk and bias of the PCH using the different VC estimators, one needs to know the true value of the PCH. In fact, assuming a frequency of the susceptible allele $p = 0.5$ and using the true values of the variance components given in (14), one can calculate the first true PCH (PCH_1) and its associated heritability (h_1). These are given, respectively, by

$$\text{PCH}_1 = (0.161, 0.161, 0.562, 0.562, 0.562, 0, \ldots, 0)_r^T, \quad h_1 = 0.380.$$

The risk and the squared bias of each VC and PCH are calculated from (13). The bias and the mean square error of heritability for the first PCH were respectively:

$$Bias(h) = \frac{1}{B}\sum_{b=1}^{B}(\hat{h}_b - h_1), \quad \text{and} \quad MSE(h) = \frac{1}{B}\sum_{b=1}^{B}(\hat{h}_b - h_1)^2,$$

where h_1 is the true heritability and $\hat{h}_b$ was the estimate of h_0 in the b th replication.

Table 1: Rows 1-4 represent the squared bias (sb) of the estimators of Σ_g, Σ_e, PCH1$_\lambda$ and all PCH$_\lambda$ (mean squared bias of each PCHs) respectively. Row 5 represents the bias of the heritability of the first PCH. Rows 6-9 represent the risk (R) of the estimators of Σ_g, Σ_e, PCH1$_\lambda$ and all PCH$_\lambda$ (mean risk of each PCHs) respectively. Row 10 represents the mean squared error of the heritability of the first PCH.

	5 traits				10 traits				15 traits			
	MLE	*Anov*	*REML*	*Ott*	*MLE*	*Anov*	*REML*	*Ott*	*MLE*	*Anov*	*REML*	*Ott*
$sb(\hat{\Sigma}_{g+})$	0.76	0.11	0.07	22.1	0.45	0.18	0.31	14.93	0.95	0.56	0.36	22.6
$sb(\hat{\Sigma}_{e+})$	0.67	0.02	0.01	21.3	0.67	0.05	0.05	21.8	1.13	0.09	0.03	22.0
sb(PCH1)	0.02	0.02	0.02	0.02	0.06	0.06	0.08	0.06	0.95	0.35	0.19	0.96
sb_{all}(PCH)	1.32	1.99	2.18	1.96	1.64	1.87	1.54	1.90	2.15	1.93	1.89	1.95
$Bias(h)$	-0.08	-0.00	0.02	-0.50	-0.06	0.01	0.04	-0.50	-0.04	0.01	0.08	-0.50
$R(\hat{\Sigma}_{g+})$	2.69	4.43	1.96	22.6	5.75	7.03	4.90	15.8	8.68	7.67	4.46	23.4
$R(\hat{\Sigma}_{e+})$	1.96	3.45	1.51	22.2	2.91	5.60	2.99	23.1	5.70	8.10	4.94	24.2
R(PCH1)	0.03	0.04	0.03	0.04	0.30	0.34	0.40	0.36	1.95	1.96	1.85	1.96
R_{all}(PCH)	3.17	3.51	3.57	3.6	2.74	2.94	2.53	2.94	2.90	2.78	2.87	3.15
$MSE(h)$	0.01	0.02	0.01	0.26	0.01	0.02	0.01	0.25	0.01	0.01	0.01	0.25

We compared our estimating approach (noted *Anov* in this section) with the MLE and REML estimators as well as with the variance components proposed in the original PCH paper from Ott & Rabinowitz (1999) (noted *Ott* in this section). The squared bias and the empirical mean of the risk of $\hat{\Sigma}_g$, $\hat{\Sigma}_e$ are presented in Table 1 for $r = 5, 10, 15$ traits. This table also presents the risk and squared bias of the matrix that has all PCH as columns as well as of the first PCH and its associated heritability. We can see that our proposed estimators and the REML estimators provide substantial reduction in squared bias over the MLE VC estimators (rows 1-4). When comparing the risks (rows 6-9), we see that ANOVA's, MLE and REML's methods are similar in terms of the efficiency except for the risk of Σ_g and Σ_e. This is consistent with the Monte Carlo simulation study conducted by Swallow and Monahan (1984). The significance of the difference between ANOVA's and MLE's risks decreases when the number of traits, r, increases, while REML's risks maintain good performance even as r increases. Furthermore, as proven in Proposition 2, Table 5.1 shows that the estimation method of Ott gives the same PCH as the ANOVA method. However, as one can see, Ott's estimators do not lead to good estimators of variance components and heritability.

Table 2: Empirical mean and mean-squared-error of the first ten components of PCH1$_\lambda$ calculated from the ANOVA method for 50 and 100 traits respectively:

PCH1	$\widehat{\text{PCH1}}_\lambda$(50 traits)				$\widehat{\text{PCH1}}_\lambda$(100 traits)			
	ANOVA		MLE		ANOVA		MLE	
	mean	MSE	mean	MSE	mean	MSE	mean	MSE
0.161	0.230	0.015	—	—	0.231	0.012	—	—
0.161	0.208	0.012	—	—	0.221	0.011	—	—
0.562	0.507	0.008	—	—	0.524	0.006	—	—
0.562	0.489	0.010	—	—	0.512	0.008	—	—
0.562	0.541	0.003	—	—	0.470	0.011	—	—
0	0.001	0.003	—	—	0.003	0.002	—	—
0	0.001	0.004	—	—	0.0004	0.002	—	—
0	0.001	0.003	—	—	0.004	0.002	—	—
0	0.001	0.004	—	—	0.003	0.002	—	—
0	0.001	0.004	—	—	0.0003	0.002	—	—

Table 2 summarizes the performance of the ANOVA estimator for $r = 50$ and 100 traits while keeping the same simulation design as in Table 5.1. Note that Mendel and ASREML Softwares could not provide estimates for such a high number of traits. Using Mendel, the extreme demands on working memory led to a crash in the machine system; also, ASREML gives automatically an error message mentioning its incapability to deal with more than 20 traits. Table 2 provides the weights of the first PCH for the first 10 traits (5 genetic trait components and 5 non-genetic trait components). The other weights are not shown here. For $r = 50$, the minimal value of these weights was 0.0001 and the maximal was 0.001. For the augmented case $r = 100$ traits, PCH$_\lambda$ has a similar behavior compared to the case $r = 50$. Note that the minimal value of the last 90 weights was 0.0001 and the maximal was 0.005. A summary of the above analyses is presented in Table 3.

Table 3: Conclusion of the ANOVA's, MLE and REML's method comparison. + denotes significant difference, ++ denotes higher significant difference and 0 denotes no evidence for difference:

Method	$R(\hat{\Sigma}_+)$	$R(\text{PCH1}_\lambda)$	square bias	MSE(h)	Bias(h)	risk of all PCH	high dimension	CPU time
MLE	+	0		0	0	0		
ANOVA		0	++	0	0	0	++	++
REML	++	0	++	0	0	0		+

5.2 Comparing the use of PCH versus PCA in linkage analysis

The previous set of simulation compared the efficiency of our proposed VC estimators with respect to the MLE and REML estimators. We also conducted two other sets of simulations to investigate linkage analysis using the first PCA (PCA1) and first PCH (PCH1) as phenotypes. This was done by performing a univariate multipoint variance-component linkage analysis and test for linkage using the Mendel software (Lange et al., 1976, 1983, 2001, 2006 and Bauman et al., 2005). The two settings we used for linkage analysis were identical to Wang et al. (2007a). Specifically, parameters for the simulation were:

$$\mu = \begin{pmatrix} 1.5_5 \\ 0_5 \\ 0_{40} \end{pmatrix}, \quad \Sigma_e = \begin{pmatrix} diag(0.25_{5\times1}) & 0_{5\times5} & 0_{5\times40} \\ 0_{5\times5} & diag(2.5_{5\times1}) + 0.5_{5\times5} & 0_{5\times40} \\ 0_{40\times5} & 0_{40\times5} & diag(0.5_{40\times1}) \end{pmatrix} \qquad (15)$$

$$\mu = \begin{pmatrix} 1_5 \\ 2_5 \\ 0_{40} \end{pmatrix}, \quad \Sigma_e = \begin{pmatrix} diag(2.9_{5\times1}) + 0.1_{5\times5} & 0.1_{5\times5} & 0_{5\times40} \\ 0.1_{5\times5} & diag(0.9_{5\times1}) + 0.1_{5\times5} & 0_{5\times40} \\ 0_{40\times5} & 0_{40\times5} & diag(2_{40\times1}) \end{pmatrix} \qquad (16)$$

We generated the same 100 families scenario as before. For each subject, five markers were simulated 20cM appart from each other, with two alleles each using the SIMULATE program (Terwilliger & Ott, 1993). Except marker 2, which had allele frequencies of 0.8 and 0.2, all markers had equal allele frequencies of

0.5. We used the model given in (12) to generate traits for general pedigrees where X represents the number of minor allele at marker 2. So, marker 2 is the QTL in this case.

Results from the set of linkage analysis are summarized in Figure 2. Using the variance-component model, we can test the null hypothesis that the additive genetic variance due to the QTL equals zero (no linkage) by comparing the likelihood of this restricted model with that of a model in which the variance due to the QTL is estimated. The difference between the two log_{10} likelihoods produces a LOD score. A LOD score above 2.0 is suggestive and above 3.0 shows evidence in favor of a QTL near the given map position. LOD scores were computed using both the ANOVA PCH1$_\lambda$ and PCA1 as phenotypes. As we can see, using the PCH as a phenotype leads to much higher LOD scores than using the first regular principal component. In fact, such results are not surprising because classical PCA ignores the correlations between family members. Since the variance for the non-genetic traits were larger than the variance for the genetic traits, classical PCA did not capture most of the genetic variability and gave more weights to the non-genetic traits. Note that the higher values of PCH1's LOD score for the first setting is due to the fact that the genetic traits were not influenced by non genetic factors (i.e. they are traits with small environmental variations). Also, in the second setting, PCA1 captures relatively large signal since traits 1-5 are genetic traits and had large total variation.

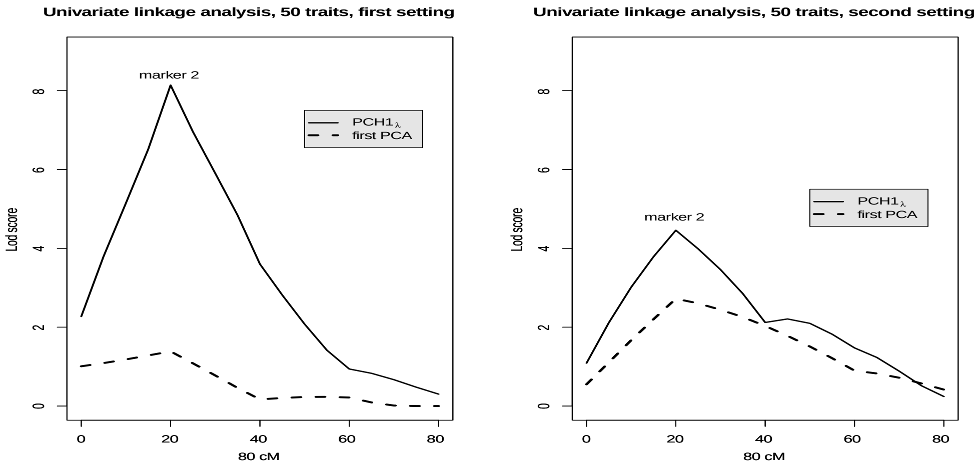

Figure 2: Univariate multipoint variance-component linkage analysis using both the ANOVA PCH1$_\lambda$ and PCA1 traits as phenotypes, $\lambda_{optim} = 60$.

In both settings we assumed no correlation between the last $(r-10)$ non-genetic traits. In microarray data, most gene-expression levels are highly correlated with each other; this is due on the one hand to underlying environmental factors, and on the other to some common genetic regulation. To reflect real situations, we added to both settings above a few scenarios corresponding to varying environmental correlations between the last $r-10$ traits ($\rho = 0.2, 0.5, 0.7$) for any pair of traits. Figure 3 shows results for LOD scores computed using the ANOVA PCH1$_\lambda$ as phenotype. One can notice that the performance of the ANOVA PCH1 estimators remained good as the correlation increased. Note that the same analysis using the first ANOVA PCA1 is not reported here as it did not detect linkage in any of the settings (LOD scores equal to zero).

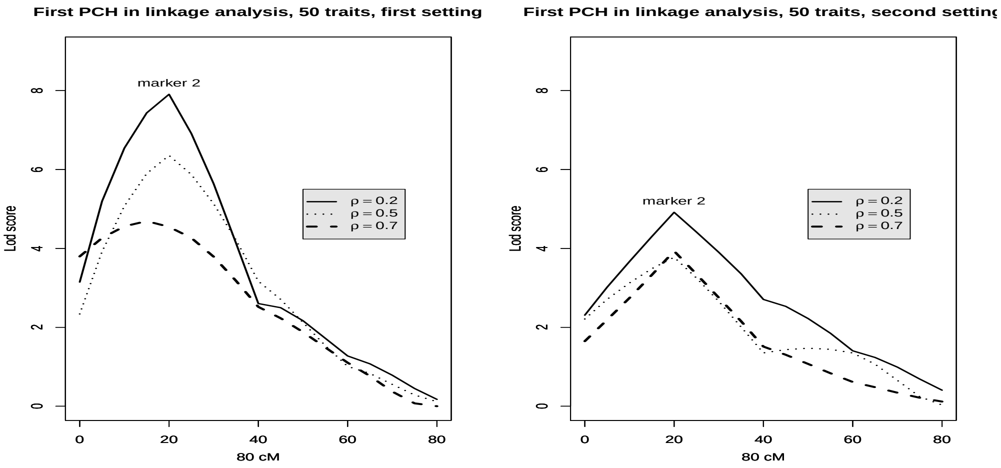

Figure 3: Univariate multipoint variance-component linkage analysis using the ANOVA PCH1$_\lambda$ trait as phenotype, $\lambda_{optim} = 60$.

6 Data analysis

We applied the PCH framework using our proposed ANOVA estimators to a unique schizophrenia (SZ) and bipolar disorder (BP) sample from the Eastern Quebec population. This sample of 48 multigenerational families comprises a total of 1,278 individuals, 365 of whom are affected by SZ or BP spectrum disorder. The lifetime presence of symptoms of psychosis, manic and depressive symptoms in a total of 439 subjects were evaluated on a six-point rating scale on each of the 82 items of the Comprehensive Assessment of Symptoms and History (CASH) instrument (Andreasen et al., 1992). These 439 subjects contained the 365 subjects affected by SZ or BP spectrum disorder plus another set of 74 subjects unaffected by these disorders but showing the presence of some of these symptoms. The procedure of assessment is described in detail in Maziade et al. (1995). The 82 items covered the following 11 dimensions: Delusion (15 items), Hallucination (6 items), Bizarre behavior (5 items), Catatonia (6 items), Thought disorder (7 items), Alogia (4 items), Anhedonia (5 items), Apathy (4 items), Affective blunting (8 items), Mania (8 items) and Depression (14 items). In the current paper, these 11 -acute episode- dimensions scores, computed as the average of their items, were used for subsequent analyses.

Table 4: Estimates of $PCH1_\lambda$ and associated heritability using the ANOVA (A) and MLE methods respectively (first two columns). The remaining columns represent the estimates of the next three PCHs and their corresponding heritability using the ANOVA method

	$PCH1_\lambda^{(A)}$	$PCH1_\lambda^{(MLE)}$	$PCH2_\lambda^{(A)}$	$PCH3_\lambda^{(A)}$	$PCH4_\lambda^{(A)}$
delusion	0.160	0.093	0.215	0.288	-0.072
hallucination	0.276	0.200	0.421	0.292	0.522
bizarre behavior	0.209	0.178	0.345	-0.146	-0.462
anhedonia	0.531	0.615	-0.228	-0.525	0.042
apathy	0.419	0.459	-0.020	-0.317	-0.106
catatonia	0.077	0.065	0.123	-0.013	-0.100
affective blunting	0.388	0.374	-0.045	0.190	-0.004
thought disorder	0.141	0.124	0.410	-0.094	0.313
alogia	0.221	0.258	0.203	0.063	0.079
mania	-0.348	-0.288	0.608	-0.426	-0.186
depression	-0.223	-0.148	-0.090	-0.450	0.589
heritability	0.925	0.781	0.454	0.411	0.296

First, we estimated the variance components of the genetic model (1) using both ANOVA and maximum likelihood methods. Using these estimates, we computed the corresponding principal components of heritability. The estimates of $PCH1_\lambda$ from these two methods and their associate heritabilities are given in Table 4. As we can see in this Table (first two columns), the estimates of the first PCH using the ANOVA and MLE methods are relatively similar. This is also true for the remaining PCHs (results not shown). However, we can see that the ANOVA's estimate produces a larger heritability, and therefore can be more informative for a future genetic linkage analysis. The principal components of heritability can be interpreted by looking at the weights for each trait component. Traits with large weights (in absolute value) tend to have larger genetic effects. As we can see in Table 4 (first column), the first principal component of heritability discriminates subjects with mania and depression symptoms from other subjects. This is interesting, as mania and depression symptoms characterize bipolar disorder. In other words, the first PCH discriminates BP subjects from other subjects. This was somehow expected, as it is known that the disease diagnostic is a heritable trait. However, using several PCHs may enable us to refine the diagnostic criterion. For example, by looking at the largest weights, the second PCH can be seen as a contrast between subjects showing an agitated and disorganized psychotic state (mania, hallucination, thought disorder, bizarre behavior) and other subjects.

Table 5: Latent class analysis using the first four PCH's of the 439 non-missing data subjects. Four classes:

Diagnosis	Class 1	Class 2	Class 3	Class 4	Total
SZ	13	103	0	11	127
BP	105	19	26	49	199
other	3	8	20	45	74
SZ affective	15	21	0	3	39

In order to evaluate the concordance between the PCHs and the diagnoses, we applied a latent class model for pedigree data developed by Labbe et al. (2009) based on the first four PCHs. The probability of class membership returned by the model allowed us to classify subjects in their most probable class (or cluster) according to their PCH patterns. In order to compare the classification obtained from the PCHs with the four diagnoses categories (SZ, BP, SZ affective and other), we specified a model with 4 classes (or clusters). Table 5 shows how the four PCH clusters intersect with these four diagnostic categories. As we can see, cluster 1 contains a majority of SZ subjects and cluster 2 contains a majority of BP subjects. However, the PCHs seem to capture the genetic heterogeneity of these two diseases, as BP subjects tend to be splitted mainly into 2 clusters. Interestingly, subjects with other diagnoses seem to be assigned in a cluster with BP and SZ subjects. The fact that we didn't obtain a complete concordance with the four diagnostic categories (SZ, BP, SZ affective and other) indicates that PCHs may represent new phenotype definitions.

7 Discussion

We proposed some modified ANOVA estimators of the variance components of a one way multivariate genetic random effect ANOVA model for general pedigrees and applied them to the PCH context. The main contribution of our work is that these ANOVA estimators allow to estimate PCH in families of any structure (and not only siblings) as well as in situations where the number of traits is large. In human studies, family data are typically collected over at least two generations. Furthermore, in fields such as mental health where phenotype definition is a challenge, researchers collect more and more traits in order to better characterize the phenotype-genotype relationships. An extreme case of large number of phenotypes is also illustrated in eQTL (expression-QTL) studies where thousands of gene expressions are used as phenotypes to better characterize the genetic contribution to the variation of gene expressions.

As illustrated in our simulation study, our proposed ANOVA VC estimators significantly reduce the bias over the corresponding maximum likelihood estimators, at the price of a slight increase in risk. Since it is beneficial to have estimators close to the true value (small bias) with modest risk than the opposite, we would recommend in practice the use of ANOVA's estimators if the risk is not a serious concern. Otherwise, we recommend the REML estimators when applicable (i.e. small number of traits). Furthermore, these estimators are extremely fast to compute and are based on explicit formulas. They can also handle a large number of traits, which is not the case for the maximum likelihood estimator. In fact, what could be seen first as an outdated approach is actually a fast and powerful way to estimate variance components. As shown in Swallow & Monahan (1984), the ANOVA VC estimators we propose are unique and efficient, as it is the case with one-way random effect models. Note that one can also include covariates in the model. In this case, a mixed-effects regression model could be used to estimate the regression coefficients and the variance components simultaneously (see for example Baltagi & Chang, 1994, for a discussion of the ANOVA estimates in such a model).

Principal component of heritability is a great tool that allows the selection of the most heritable linear combination of phenotypes, measured in families. In this paper, we established a link between PCHs and discriminant axes from a traditional linear discriminant analysis. In particular, we showed that the PCHs computed using the variance components of Ott & Rabinowitz and Wang et al. (2007a) for simple pedigrees are exactly equal to the discriminant axes. As in standard principal component analysis, selection of the PCH is an issue. In PCA, PCs are chosen according to the percentage of the variance explained. In a similar spirit, one may choose the PCHs that have a higher heritability than each of the traits separately.

Note that in order to obtain a better interpretation of the PCHs, it would also be possible to apply a rotation R to the matrix A of genetic correlations between the traits and the PCHs, as well as to the matrix of PCHs. The rotation would be chosen such that the genetic correlations obtained in the matrix AR are either close to 0 or 1.

Using the first PCH as a quantitative phenotype, or several PCH's as multivariate quantitative phenotypes in linkage analysis provides greater power to detect disease susceptibility genes. Our simulations show that using $PCH1_\lambda$ as a phenotype in linkage analysis leads to higher LOD score, while using the standard PCA results in lower of LOD scores. This is coherent with the simulation studies conducted by Ott & Rabinowitz and Wang et al. (2007a). Wang et al. (2009) also compared PCA and Factor Analysis (FA) to uncover genetic factors that contribute to complex disease phenotypes. They found that FA generally produced factors that had stronger correlations with the genetic traits. Faced with high dimensional traits as in Morley et al. (2004), where 3,354 gene expression traits were measured, one could construct clusters that combine similar traits among family members and then used the ML estimates of the PCH to combine the phenotypes in each cluster. However, the trait clusters can still be relatively large as it is the case with microarray gene expression data. Wang et al. (2007b) proposed a clustering approach that takes into account the family structure information. They applied this approach to the gene expression data used in Morley et al. (2004) and then used the PCH approach to combine the phenotypes in each cluster. Nevertheless, they used the sample between-family sum-of-squares to estimate Σ_g, and used the sample within-family sum-of-squares to estimate Σ_e for the 14 large families used in this data. Using the proposed ANOVA estimates of VC could greatly improve the estimation of the PCHs for these data.

Finally, we point out some limitations of the ANOVA approach: first, the risk of the VC estimators deteriorates when one deals with extremely large pedigrees (results not shown). In practice, this is not of great importance since the size of a pedigree rarely exceeds 30 subjects. Second,we focused on simple variance-covariance structure of the genetic model. With a more general variance structure, the model will be more complex. For instance, one can assume that the phenotype is influenced by l loci. However, if we are focusing on the analysis of one locus specifically, we can absorb the effects of all the remaining QTLs in a residual genetic component and the variance-covariance between relatives will be decomposed into three components: the major locus component (specific QTL), the polygenic component (residual genetic) and the environmental component. Thus, a system of three equations, rather than the two independent equations in (9), can solve this estimation problem. It would be then interesting to compute the heritability based on the major locus. This is the object of a future work. Finally, an R package computing the estimators of Σ_g, Σ_e and PCH_λ is available from the authors upon request.

Supplementary Materials

Web Appendix A, referenced in Section 5.1, is available under the Paper Information link at the Berkeley electronic press website

A Appendix

A.1 A.1. Proof of Proposition 1

The expectations of Proposition 1 come from the decomposition of S_w and S_b as follows:

$$S_w = \sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{ij}Y_{ij}^T - \sum_{i=1}^{m} n_i\bar{Y}_i\bar{Y}_i^T, \quad S_b = \sum_{i=1}^{m} n_i\bar{Y}_i\bar{Y}_i^T - n\bar{Y}\bar{Y}^T. \tag{17}$$

Note that the mean of Y_{ij} given in (12) is $2p\mu$, however, to facilitate the calculations, we assume here that it is centered. Thus, to prove Proposition 1, one needs to evaluate the expectations $\mathbb{E}(Y_{ij}Y_{ij}^T)$, $\mathbb{E}(\bar{Y}_i\bar{Y}_i^T)$, $\mathbb{E}(\bar{Y}\bar{Y}^T)$. Following the model (1) one can write:

$$\mathbb{E}(Y_{ij}Y_{ij}^T) = 2\Phi_{jj}^{(i)}\Sigma_g + \Sigma_e. \tag{18}$$

$$\begin{aligned}\mathbb{E}(\bar{Y}_i\bar{Y}_i^T) &= \frac{1}{n_i^2}\left[\sum_{j=1}^{n_i}\mathbb{E}(Y_{ij}Y_{ij}^T) + \sum_{j\neq k}^{n_i}\mathbb{E}(Y_{ij}Y_{ik}^T)\right] \\ &= \frac{1}{n_i^2}\left\{\sum_{j=1}^{n_i}\left[2\Phi_{jj}^{(i)}\Sigma_g + \Sigma_e\right] + \sum_{j\neq k}^{n_i} 2\Phi_{jk}^{(i)}\Sigma_g\right\} \\ &= \frac{1}{n_i}\Sigma_e + \frac{\tau_b^{(i)}}{n_i^2}\Sigma_g, \end{aligned} \tag{19}$$

DOI: 10.2202/1544-6115.1711

where $\tau_b^{(i)}$ is defined in Proposition 1. The expectation of $\bar{Y}\bar{Y}^T$ can be written as

$$\begin{aligned}
\mathbb{E}(\bar{Y}\bar{Y}^T) &= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{m} n_i\bar{Y}_i \sum_{l=1}^{m} n_l\bar{Y}_l^T\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{m} n_i^2\mathbb{E}(\bar{Y}_i\bar{Y}_i^T) + \sum_{i\neq l}^{m} n_i n_l \mathbb{E}(\bar{Y}_i\bar{Y}_l^T)\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{m} n_i^2\left(\frac{1}{n_i}\Sigma_e + \frac{\tau_b^{(i)}\Sigma_g}{n_i^2}\right)\right] \\
&= \frac{1}{n}\Sigma_e + \frac{\tau_b}{n^2}\Sigma_g, \qquad (20)
\end{aligned}$$

where τ_b is defined in Proposition 1. Using (17), (18) and (19), one has

$$\begin{aligned}
\mathbb{E}(S_w) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(2\Phi_{jj}^{(i)}\Sigma_g + \Sigma_e\right) - \sum_{i=1}^{m} n_i\left(\frac{1}{n_i}\Sigma_e + \frac{\tau_b^{(i)}}{n_i^2}\Sigma_g\right), \\
&= (n-m)\Sigma_e + (\tau_a - \tau_c)\Sigma_g.
\end{aligned}$$

The expectation of S_b is deduced in a similar way.

References

[1] Almasy L, Dyer TH, Blangero J (1998). Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. *Genet. Epidemiol.*, **14**: 953-958.

[2] Almasy L, Blangero J (1998). Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, **62**:1198-1211.

[3] Amemiya Y (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definit? *Amer. Stat.*, **39**, 112-117.

[4] Amos CI, de Andrade M, Zhu DK (2001). Comparison of multivariate tests for genetic linkage. *Hum Hered*, **51**, 133-144.

[5] Andreasen NC, Flaum M, Arndt S (1992). The Comprehensive Assessment of Symptoms and History (CASH) an instrument for assessing diagnosis and psychopathology. *Arch Gen Psychiatry*, **49**(8), 615-623.

[6] Arya R, Blangero J, Williams K, Almasy L, Dyer TD, Leach RJ, O'Connell P, Stern MP, Duggirala R (2003). Factors of insulin resistance syndrome- related phenotypes are linked to genetic locations on chromosomes 6 and 7 in nondiabetic Mexican-Americans. *Diabetes*, **51**, 841-847.

[7] Baltagi BH, Chang YJ (1994). Incomplete panels: A comparative study of alternative estimators for the unbalanced one-way error component regression model. *Journal of Econometrics*, **62**: 67–89.

[8] Bauman L, Almasy L, Blangero J, Duggirala R, Sinsheimer JS, Lange K (2005). Fishing for pleiotropic QTLs in a polygenic sea. *Ann Hum Genet*, **69**: 590-611.

[9] Calum AM, Ramachandran SV (2011). Next-generation genome-wide association studies: Time to focus on phenotype? *Circ Cardiovasc Genet.*, **4**: 334-336.

[10] Fisher RA (1936). The use of multiple measures in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

[11] Funalot B, Varenne O, Mas JL (2004). A call for accurate phenotype definition in the study of complex disorders.. *Nature Genetics*, **36**:3.

[12] Gilmour AR, Cullis BR, Welham SJ, Thompson R (1999). *ASREML, reference manual*. Biometric bulletin, no 3, NSW Agriculture, Orange Agricultural Institute, Forest Road, Orange 2800 NSW Australia.

[13] Labbe A, Bureau A, Merette C (2009). Integration of genetic familial structurein latent class models. *The International journal of Biostatistics*, **5**(1).

[14] Lange K, Westlake J, Spence MA (1976). Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann Hum Genet*, **39**, 485-491.

[15] Lange K, Boehnke M (l983). Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. *Amer J Med Genet*, **14**, 513-524.

[16] Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E (2001). MENDEL version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Amer J Hum Genet*, **69** (Suppl):504.

[17] Lange K (2002). *Mathematical and Statistical Methods for Genetic Analysis*. Springer Verlag, New York.

[18] Lange K, Sobel E (2006). Variance component models for X-linked QTLs. *Genet Epidemiology*, **30**, 380-383.

[19] Mardia KV, Kent JT, Bibby JM (1979). *Multivariate Analysis*. Academic Press, London.

[20] Mathew T, Niyogi A, Sinha BK (1994). Improved nonnegative estimation of variance components in balanced multivariate mixed models. *J Multivariate Anal*, **51**, 83-101.

[21] Maziade M, Roy MA, Martinez M, Cliche D, Fournier JP, Garneau Y, et al. (1995). Negative, psychoticism, and disorganized dimensions in patients with familial schizophrenia or bipolar disorder: continuity and discontinuity between the major psychoses. *Am J Psychiatry*, **152**(10), 1458-63.

[22] McCulloch CE, Searle SR (2002). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.

[23] Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung BG (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430 (7001)**, 743-747.

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

[25] Searle SR, Casella G, McCulloch CE (1992). *Variance Components*. John Wiley & Sons, New York.

[26] Srivastava MS, Kubokawa T (1999). Improved nonnegative estimation of multivariate components of variance. *The Annals of Statistics*, **27**, No. 6, 2008-2032.

[27] Sun YJ, Sinha BK, von Rosen D, Meng QY (2003). Nonnegative estimation of variance components in multivariate unbalanced mixed linear models with two variance components. *J Stat Plann and Inf*, **115**, 215234.

[28] Swallow WH ,Monahan JF (1984). Monte carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, **26**, 47-57.

[29] Terwilliger JD, Speer M, Ott J (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet Epidemiol*, **10**, 217-224.

[30] Wang Y, Fang Y, Wang S (2007). Clustering and principal component analysis for mapping co-regulated genome-wide variation using family data. *BMC Genet*, **1**(Suppl 1):S121.

[31] Wang Y, Fang Y, Jin M (2007). A ridge penalized principal-components approach based on heritability for high-dimensional data. *Hum Hered*, **64**, 182-191.

[32] Wang X, Kammerer CM, Anderson S, Lu J, Feingold E (2009). A comparison of principal component analysis and factor analysis strategies for uncovering pleiotropic factors. *Genet Epidemiol*, **33**(4), 325-331.