Original Paper

Human Heredity

Hum Hered 1999;49:106–111

Received: July 13, 1998
Revision received: September 14, 1998
Accepted: September 16, 1998

A Principal-Components Approach Based on Heritability for Combining Phenotype Information

Jürg Ott[a] Daniel Rabinowitz[b]

[a]Laboratory of Statistical Genetics, Rockefeller University, and [b]Department of Statistics, Columbia University, New York, N.Y., USA

Key Words

Complex traits · Factor analysis · Genetics · Multivariate · Segregation analysis · Variance components

Abstract

For many traits, genetically relevant disease definition is unclear. For this reason, researchers applying linkage analysis often obtain information on a variety of items. With a large number of items, however, the test statistic from a multivariate analysis may require a prohibitively expensive correction for the multiple comparisons. The researcher is faced, therefore, with the issue of choosing which variables or combinations of variables to use in the linkage analysis. One approach to combining items is to first subject the data to a principal components analysis, and then perform the linkage analysis of the first few principal components. However, principal-components analyses do not take family structure into account. Here, an approach is developed in which family structure is taken into account when combining the data. The essence of the approach is to define principal components of heritability as the scores with maximum heritability in the data set, subject to being uncorrelated with each other. The principal components of heritability may be calculated as the solutions to a generalized eigensystem problem. Four simulation experiments are used to compare the power of linkage analyses based on the principal components of heritability and the usual principal components. The first of the experiments corresponds to the null hypothesis of no linkage. The second corresponds to a setting where the two kinds of principal components coincide. The third corresponds to a setting in which they are quite different and where the first of the usual principal components is not expected to have any power beyond the type I error rate. The fourth set of experiments corresponds to a setting where the usual principal components and the principal components of heritability differ, but where the first of the usual principal components is not without power. The results of the simulation experiments indicate that the principal components of heritability can be substantially different from the standard principal components and that when they are different, substantial gains in power can result by using the principal components of heritability in place of the standard principal components in linkage analyses.

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
http://BioMedNet.com/karger

Daniel Rabinowitz
Department of Statistics, Mathematics Building
Columbia University
New York, NY 10027 (USA)
Tel. +1 (212) 854 3400, Fax +1 (212) 663 2454

Introduction

For many traits, while there may exist unequivocal diagnostic schemes for particular disease entities, genetically relevant disease definition is unclear. An underlying gene or set of genes may, for example, be related to a very specific trait such a schizophrenia or lead to susceptibility to a spectrum of diseases such as schizophrenia, schizotypical disorder, and bipolar disease. For this reason, researchers applying linkage analysis with complex traits often obtain from each study subject information on a variety of items relevant to the trait in question. The information on these items may be coded as quantitative variables or as scales that combine information from several items. For example, Basset et al. [1993] discuss the use of the Positive and Negative Syndrome Scale (PANSS) of Kay and Sandyk [1987] for characterizing schizophrenia in pedigree members.

Any of the variables or scales may be subjected to the desired genetic analysis and the results can then be compared heuristically. With a large number of items, however, such an exploratory approach may be unwieldy. Furthermore, the corrections needed to account for the multiple looks corresponding to the many items may be prohibitive. Alternatively, the items may used together as a multivariate outcome in a single analysis. See, for example, Schork [1993] and Amos et al. [1990]. With a large number of items, however, the power of a multivariate analysis to detect linkage can be substantially lower than the power of an analysis applied to a genetically relevant scale. This is because the increased degrees of freedom for the reference distribution of a multivariate analysis can result in too stringent a criterion for statistical significance.

One approach to choosing a scale is to subject the available items to a principal-components analysis. The first few principal components are then candidates for linkage analyses. Lindström and von Knorring [1993], for example, subject PANSS scores to a principal-components analysis. See also Lindenmayer et al. [1995]. An intermediate step may be to estimate the heritability of the first several principal components, and to use the components with highest heritability. See Hasstedt et al. [1994] for an example concerning sodium transport and Livshits et al. [1995] for an example concerning body size and shape. Farmer et al. [1987] examined heritability for several different diagnostic criteria for schizophrenia. However, principal components depend on the variance-covariance matrix of the data pooled from all pedigrees, and thus do not reflect family structure. By focusing only on scales obtained as principal components, other scales with higher heritability may be overlooked.

Here, an approach is developed in which the available items are combined into scales, not on the basis of their variance-covariance structure, but instead, on the basis of heritability. The first scale has highest heritability, the second scale has highest heritability among all scales uncorrelated with the first, the third has highest heritability among those uncorrelated with the first and second, an so on. Heritability is the ratio of the variances of family-specific components and individual specific components of variance. Thus, a combination of the variance-covariance structure of both the between-family and within-family variance components are used to compute the scales.

Approaches similar in spirit have been contemplated previously. Zlotnik et al. [1983] describe an approach based on choosing linear combinations to maximize the likelihood under the hypothesis of single-gene segregation in a single pedigree. Principal-components analyses was used in the computations. Multivariate selection indices are computed in the context of plant and animal breeding. Selection indices combine the economic value or relative fitness of traits together with heritability, and the computation of selection indices has parallels with the computations suggested in the next section. See, for example, Humphreys [1995], Hazel [1943] or Kempthorne [1969]. Boomsma [1996] used simulation studies to examine the increase in power that can occur when using linear combinations based on factor scores to detect linkage in the presence of pleiotropy. Comuzzie et al. [1997] describe two similar approaches. The first involves 'conditioning each phenotype on the common or shared genetic effects with the others in the group to maximize the variance of each trait attributable to unique loci'. The goal of such an approach is to parcel out the effect of different loci, while the goal of the approach presented here is to combine information in order to have greater power in the presence of pleiotropy. The second of the approaches of Comuzzie et al. [1997] involves computing estimates of the principal components of the genetic portion of the phenotype data. The approach presented here differs in that it involves computing what might be thought of as the principal components of the genetic portion of the data relative to the residual variability in the phenotypes. In the next section, the approach to choosing scales is detailed. Also described in the next section are simulation experiments that investigate the relative efficiency of linkage analyses based on the approach. In the third section, the results of the simulation experiments are reported.

Table 1. Models for the simulation experiments

Model	PCH, heritability	PC, heritability
$Y = N\left(0, \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\right)$	Doesn't exist	$\begin{pmatrix} 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \end{pmatrix}, 0$
$Y = X\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + N\left(0, \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\right)$	$\begin{pmatrix} 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \end{pmatrix}, 0.556$	$\begin{pmatrix} 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \\ 0.447 \end{pmatrix}, 0.556$
$Y = X\begin{pmatrix} 1.5 \\ 1.5 \\ 0 \\ 0 \\ 0 \end{pmatrix} + N\left(0, \begin{pmatrix} 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 2.0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 2.0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 & 2.0 \end{pmatrix}\right)$	$\begin{pmatrix} 0.707 \\ 0.707 \\ 0 \\ 0 \\ 0 \end{pmatrix}, 0.667$	$\begin{pmatrix} 0 \\ 0 \\ 0.577 \\ 0.577 \\ 0.577 \end{pmatrix}, 0$
$Y = X\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + N\left(0, \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1.5 \end{pmatrix}\right)$	$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, 0.2$	$\begin{pmatrix} 0.309 \\ 0.541 \\ 0.541 \\ 0.541 \\ 0.190 \end{pmatrix}, 0.032$

The observations for an individual are denoted by Y, the number of transmitted first alleles at the trait locus is denoted by X. The notation $N(0, \Sigma)$ refers to a multivariate normal vector with expectation 0 and variance-covariance matrix Σ. The coefficients of the principal components of heritability (PCH) and principal components (PC), together with their heritabilities are also recorded. The models are listed in the same order as described in the text.

Methods

It is convenient to focus on the situation in which phenotype data in the form of a number of items have been obtained from the offspring in a sample of nuclear families. Let p denote the number of items, and let Y denote the p dimensional vector of items. It is natural to think of the vector of items as composed of a family-specific component, A, and an individual-specific component, E,

$$Y = A + E,$$

with the two components uncorrelated with each other. The family-specific component is induced by variation in the parental genes, and by shared environmental factors. The individual-specific component is induced by subject-specific environmental factors, and differences in transmission of parental alleles. Let Σ_A and Σ_E denote the variance-covariance matrices of A and of E, respectively. Then the heritability of a linear combination of the items c^TY may be expressed as

$$c^T\Sigma_A c|c^T (\Sigma_A + \Sigma_E)c,$$

where the column vector c is the coefficient of the linear combination, and c^T is its transpose.

The standard principal components are defined as the scores with maximum variance, subject to being uncorrelated with each other. See, for example, Rao [1964] or Morrison [1976]. The principal components of heritability are defined not as the scores with maximum variance, but instead, as the scores with maximum heritability, subject to being uncorrelated with each other. That is, the principal components of heritability are scores c_1^TY, c_2^TY, ..., c_p^TY with the property that: c_1^TY has maximum heritability; c_2^TY has maximum heritability subject to c_2^TY being uncorrelated c_1^TY; c_3^TY has maximum heritability subject to c_3^TY being uncorrelated with c_1^TY and c_2^TY; and so on. It may be of interest to note that not only are the principal components of heritability uncorrelated, but that their family-specific components, the c_i^TA, are uncorrelated as well.

It is well known that c_1, c_2, ... and c_p are the solutions to the generalized eigensystem problem

$$\Sigma_A c = \lambda\Sigma_E c.$$

Press et al. [1988] discuss numerical methods for solving the generalized eigensystem problem based on an approach described by Wilkinson and Reinsch [1971]. Given estimates of the variance-covariance matrices, the principal components of heritability may be estimated by solving the generalized eigensystem problem for estimated matrices.

In order to examine the relative efficiency of using the principal components of heritability in linkage analyses, four simulation experiments were carried out. In all of the experiments, there were five items from each of two children in nuclear families. The genetic model had a single locus influencing the trait. The locus had two equally prevalent alleles. The effect of the alleles on the traits was additive: presence of a copy of the first of the two alleles added a constant to each of the items; the other allele had no effect. The constant was the same across families, but differed across items. The items were then generated by adding a multivariate normal vector to the effect of the

Table 2. Empirical probability, from simulation experiments with 10,000 replications, of detecting linkage when the Haseman-Elston procedure is applied to the first principal component (PC) and to the first principal component of heritability (PCH) in four different scenarios (described at the end of the Methods section)

α-Level	First		Second		Third		Fourth	
	PCH	PC	PCH	PC	PCH	PC	PCH	PC
0.100	0.103	0.105	0.706	0.806	0.998	0.144	0.890	0.543
0.050	0.051	0.051	0.577	0.694	0.996	0.085	0.784	0.349
0.010	0.011	0.010	0.303	0.405	0.969	0.031	0.438	0.082
0.005	0.005	0.006	0.211	0.301	0.928	0.023	0.291	0.039
0.001	0.002	0.001	0.082	0.128	0.744	0.011	0.094	0.006

gene. Thus, the family-specific component of variance was induced by the parental alleles at the gene, and the subject-specific variance was induced by the multivariate normal residual variance together with variation in the siblings' genetic makeup around the conditional expectation given the parental alleles.

The procedure of Haseman and Elston [1972] was used to test for linkage. A completely informative marker perfectly linked to the underlying trait locus was simulated. In all of the experiments, the procedure was applied to both the estimated first standard principal component and the estimated first principal component of heritability. That is, the squared differences in the first of each kind of principal component were separately regressed on the number of alleles shared identical by descent by the two siblings. The regression coefficient, normalized by its standard error, was used as the test statistic to assess significance. In each of the experiments, 10,000 replications were performed.

In order to compute the principal components, estimates of the family-specific and subject-specific variance-covariance matrices are required as input to the generalized eigensystem problem. Let Y_{ij} denote the column vector of items from the j^{th} sibling in the i^{th} family. Then, the estimates of the variance covariance matrices of the subject-specific component of variance and the family-specific component of variance used in the simulations may be expressed as

$$\hat{\Sigma}_E = \frac{1}{\Sigma_{i=1}^{n}(m_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_{i\cdot})^{\otimes 2},$$

and

$$\hat{\Sigma}_A = \frac{1}{\Sigma_{i=1}^{n} m_i - 1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^{\otimes 2} - \hat{\Sigma}_E,$$

where $\bar{Y}_{\cdot\cdot}$ is the average over all siblings in the study, and $\bar{Y}_{i\cdot}$ is the average over the m_i siblings in the i^{th} family. The notation $V^{\otimes 2}$ denotes the matrix multiplication of the column vector V by its transpose, $V^{\otimes 2} = VV^T$. All of the calculations, including random number generation, were carried out in FORTRAN77 using the IMSL [1994] subroutines library.

The models for the simulation experiments are described in the first column of table 1. In the first experiment, there was no genetic effect at all, and the items were independent and identically distributed. Each replication had 250 families. In the second simulation, the vector added to the phenotype through presence of one of the alleles had all of its components equal to 1. The multivariate normal component was independent with identically distributed variables with expectation equal to 0 and variance equal to 1. The number of families in the second experiment was 600. In the third experiment, the vector added had 1.5 in its first two components, and 0 in the other three components. The variance covariance matrix of the multivariate normal observation had 2.0 in the last three components of the diagonal, and 0.25 in the first two. The covariance between the last three items was 0.5, while the first two items were assumed uncorrelated with all the others. The number of families was 150. In the fourth experiment, the first and last components of the multivariate normal contribution to the items had covariance 1.0. All other covariances were zero. The last component had variance 1.5, while all the others had variance 1. The vector added by presence of the allele was 1.0 in the first component, and zero in the others. The number of families was 80.

The first principal component and the first principal component of heritability for the models corresponding to the simulation experiments are listed in the second and third columns of table 1. The matrices Σ_A and Σ_E are computed as $\mu\mu^T/4$ and $\mu\mu^T/4 + \Sigma$, where Σ is the variance covariance matrix of the multivariate normal component. From the heritabilities, it is expected that linkage analyses based on the principal components of heritability should improve on analyses based on the usual principal components in the third and fourth experiments, while it is hoped that the power of both approaches will be comparable in the second experiment. Neither approach is expected to have power in the first experiment.

Results

The results of the simulation experiments are recorded in table 2. The rows of the table correspond to different significance levels, and the entries in the table are the empirical probability of detecting linkage. The columns of the table correspond to the results for the principal components of heritability and the standard principal components for the four scenarios described in the previous section.

The first set of experiments had no genetic effect. The power is as expected for both scales: it is the nominal α-level.

The second set of experiments corresponds to a situation in which the first principal component and the first principal component of heritability are the same. The standard principal-component analysis has somewhat greater power than the principal component of heritability analysis, however. An examination of the estimated coefficients of the principal components provides an explanation: the coefficients of the principal components of heritability are estimated with somewhat less accuracy than the principal components.

The third simulation experiment corresponds to a situation in which the principal components of heritability are very difficult from the standard principal components. In this setting, the variability due to genetic causes is in the first two items. The first of the standard principal components, however, is a linear combination of the last three components. It is not surprising that the analysis based on the principal components of heritability improves, in this setting, on the analysis based on the standard principal components.

The fourth simulation experiment is intermediate between the second and the third. The first of the standard principal components puts weight on the first item, while the first of the principal components of heritability places weight on only the first item. As expected, the power of the analysis based on the principal component of heritability is greater than that of the analysis based on the standard principal component. However, the analysis based on the standard principal component is not without power.

Discussion

In many applications, there may be additional environmental or subject-specific covariates that should be included in the analyses. When this is the case, it would be reasonable to first regress out the covariates before estimating the covariance structure. Perhaps preferably, a mixed-effects regression model could be used to estimate the regression coefficients and the variance components simultaneously. See, for example, Laird and Ware [1983].

Here, attention has focussed on sib-pair analyses and additive genetic models. With general pedigree data, modeling and estimation of the family and individual specific components of variance are more complicated. However, in principle, the approach proposed here is applicable: after estimation of the components of variance, the principal components of heritability may be estimated by solving the generalized eigensystem problem. In any case, crude estimates of the variance-covariance structures may be obtained by treating all individuals within a pedigree the same. A variety of more sophisticated techniques for computing variance components in pedigrees have been devised. See, for example, Blangero and Konigsberg [1991].

Issues related to ascertainment have not been discussed. For a monogenic group of traits, if pedigrees are ascertained through affected individuals, the family-specific component of variance can be, in the data set, roughly constant. In such settings, especially if variability due to genetic variability within a pedigree are a large portion of the subject-specific component of variance, then the usual principal components may be preferable to the principal components of heritability.

With only pedigree structure and phenotype information, but without genotype information, it is difficult to disentangle the effects of shared environmental factors from the effect of shared genetic material, and it can be impossible to disentangle the effects of shared genetic material at different loci. Therefore, obtaining scales on the basis of heritability expose one to the risk that the scales are not relevant to any particular locus, but rather to some combination of loci and environmental factors. In fact, it is possible that in some situations, environmental factors may induce family-specific components of variance that lead to principal components of heritability that have no power for detecting linkage, while the standard principal components would have at least some power. See, for example, Jiang and Zeng [1995] for a discussion of this issue. Nevertheless, it seems natural that, if scales are to be formed without reference to genotype information, then the scales are most likely best developed with reference to heritability. A possible strategy when faced with more than a few phenotype variables might be to first apply a linkage analysis to the first or the first few principal components of heritability. If linkage is not detected with the principal component, then a multivariate linkage analysis might be performed, even if the large degrees of freedom associated with a multivariate analysis with many phenotype variables would most likely preclude a statistically significant result.

Conclusions

The results of the simulation experiments indicate that in some situations, the principal components of heritability can be substantially different from the standard principal components. Furthermore, when the principal components of heritability are used in the place of the standard principal components, substantial gains in power can result. The third of the simulation experiments is an example in which the standard principal component would be a much poorer choice than the principal component of heritability.

However, the results of the simulation experiments also indicate that some care must be taken in settings where the estimate of the subject-specific components of variance is unstable. In such cases, the instability in the estimate of the principal components of heritability may outweigh their potentially greater power. This will especially be the case in settings where the principal components of heritability and the standard principal components are not very different.

Acknowledgments

This work was supported by grant GM55978 (to D.R.) from the National Institutes of General Medical Sciences, and by the US Army Medical Research Acquisition Activity under award No. DAMD17-94-J-4406 (to J.O.). Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the US Army Medical Research Acquisition Activity. The authors thank Dr. Simon Heath for pointing out the relationship to selection indices.

References

Amos CI, Elston RC, Bonney GE, Keats BJB, Berenson GS: A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. Am J Hum Genet 1990;47:247–254.

Basset AS, Collins EJ, Nuttall SE, Honer WG: Positive and negative symptoms in families with schizophrenia. Schizophrenia Res 1993;11:9–19.

Bell MD, Lysaker PH, Beam-Goulet JL, Milstein RM, Lindenmayer JP: Five-component model of schizophrenia: Assessing the factorial invariance of the positive and negative syndrome scale. Psychiatry Res 1994;52:295–303.

Blangero J, Konigsberg LW: Multivariate segregation analysis using the mixed model. Gen Epidemiol 1991;8:299–316.

Boomsma DI: Using multivariate genetic modeling to detect pleiotropic quantitative trait loci. Behav Genet 1996;26:161–166.

Comuzzie AG, Mahaney MC, Almasy L, Dyer TD, Blangero J: Exploiting pleiotropy to map genes for oligogenic phenotypes using extended pedigree data. Gen Epidemiol 1997;14:975–980.

Farmer AE, McGuffin P, Gottesman II: Twin concordance for DSM-III schizophrenia. Scrutinizing the validity of the definition. Arch Gen Psychiatry 1987;44:634–641.

Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1972;2:3–19.

Hasstedt SJ, Hunt SC, Wu LL, Williams RR: Evidence for multiple genes determining sodium transport. Gen Epidemiol 1994;11:553–568.

Hazel LN: The genetic basis for constructing selection indexes. Genetics 1943;28:476–490.

Humphreys MO: Multitrait response to selection in *Lolium perenne* L. (perennial ryegrass) populations. Heredity 1995;74:510–517.

IMSL: FORTRAN subroutines for statistical applications and FORTRAN subroutines for mathematical applications. Houston, Visual Numerics, 1994.

Jiang C, Zeng ZB: Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 1995;140:1111–1127.

Kay SR, Sandyk R: Experimental models of schizophrenia. Int J Neurosci 1987;58:69–82.

Kempthorne O: An introduction to genetic statistics. Ames, Iowa State University Press, 1969.

Laird N, Ware J: Random-effects models for longitudinal data. Biometrics 1983;38:963–974.

Lindenmayer JP, Bernstein-Hyman R, Grochowski S, Bark N: Psychopathology of schizophrenia: Initial validation of a 5-factor model. Psychopathology 1995;28:22–31.

Lindström E, von Knorring L: Symptoms of schizophrenic syndromes in relation to age, sex, duration of illness and number of previous hospitalizations. Acta Psychiatr Scand 89:274–278.

Livshits G, Otremski I, Kobyliansky E: Genetics of human body size and shape: Complex segregation analysis. Ann Hum Biol 1995;22:13–27.

Morrison DF: Multivariate Statistical Methods, ed 2. New York, McGraw-Hill, 1976.

Press WH, Fannery BP, Teukolsky SA, Vetterling WT: Numerical Recipes in C. Cambridge, Cambridge University Press, 1988.

Rao CR: The use and interpretation of principal component analysis in applied research. Sankhya A 1964;26:329–358.

Schork NH: Extended multipoint identity-by-descent analysis of human quantitative traits; efficiency, power and modelling considerations. Am J Hum Genet 1993;53:1306–1319.

Wilkinson JH, Reinsch C: Linear Algebra, Handbook for Automatic Computation. New York, Springer, 1971, vol 2.

Zlotnik LH, Elston RC, Namboodiri KK: Pedigree discriminant analysis: A method to identify monogenic segregation. Am J Med Genet 1983; 15:307–313.