Original Paper

Human Heredity

Hum Hered 2007;64:182–191
DOI: 10.1159/000102991

Received: October 11, 2006
Accepted after revision: March 19, 2007
Published online: May 25, 2007

A Ridge Penalized Principal-Components Approach Based on Heritability for High-Dimensional Data

Yuanjia Wang[a] Yixin Fang[b] Man Jin[c]

[a]Department of Biostatistics, Mailman School of Public Health, and Departments of [b]Psychiatry-Behavioral Medicine and [c]Statistics, Columbia University, New York, N.Y., USA

Key Words
Principal-components analysis · Complex traits · Linkage analysis · Cross-validation · Family data

Abstract
Objective: To develop a ridge penalized principal-components approach based on heritability that can be applied to high-dimensional family data. ***Methods:*** The first principal component of heritability for a trait constellation is defined as a linear combination of traits that maximizes the heritability, which is equivalent to maximize the family-specific variation relative to the subject-specific variation. To analyze high-dimensional data and prevent overfitting, we propose a penalized principal-components approach based on heritability by adding a ridge penalty to the subject-specific variation. We choose the optimal regularization parameter by cross-validation. ***Results:*** The principal-components approach based on heritability with and without ridge penalty was compared to the usual principal-components analysis in four settings. The penalized principal-components of heritability analysis had substantially larger coefficients for the traits with genetic effect than for the traits with no genetic effect, while the non-regularized analysis failed to identify the genetic traits. In addition, linkage analysis on the combined traits showed that the power of the proposed methods was higher than the usual principal-components analysis and the non-regularized principal-components of heritability analysis. ***Conclusions:*** The penalized principal-components approach based on heritability can effectively handle large number of traits with family structure and provide power gain for linkage analysis. The cross-validation procedure performs well in choosing optimal magnitude of penalty.

Introduction

In some genetic studies, a constellation of traits are measured because it is of interest to locate genes associated with multiple traits. Due to uncertainty of the underlying genetic model, not all the traits measured may be influenced by genetic factors. For example, Cheung et al. [1] measured gene expression levels on over 5,000 transcription factors of human lymphoblastoid cell lines, and found evidence of familial aggregation of a proportion of these expression levels. The follow-up study in Morley et al. [2] conducted genome-wide linkage scans on each of the 3,554 gene expression phenotypes using

KARGER

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

0001–5652/07/0643–0182$23.50/0

Accessible online at:
www.karger.com/hhe

Yuanjia Wang, PhD
Department of Biostatistics, Columbia University, Mailman School of Public Health
722 West 168th Street, Room 604
New York, NY 10032 (USA)
Tel. +1 212 342 3430, Fax +1 212 305 9408, E-Mail yw2016@columbia.edu

2,882 SNPs and found evidence of two 'master regulators' that were associated with co-expressions of 31 and 25 transcription factors, respectively. Bystrykh et al. [3] examined hematopoietic stem cell in recombinant inbred mouse strains and found 17 quantitative trait loci (QTLs) controlling 10 to 272 gene expression transcripts. Particularly, there was one QTL on chromosome 4 controlling 272 gene expression levels.

In real life, the number of traits controlled by a hotspot QTL varies from tens to hundreds. When the goal of a study is to discover genes related to multiple traits, single trait analysis can be used on each of the individual trait and the results may be compared as done in Morley et al. [2]. However, when the number of traits is large, such analysis is computationally intensive and corrections required to adjust for multiple comparisons may be severe. Multivariate linkage analysis was proposed to analyze all the traits simultaneously. See for example, Amos et al. [4], and Jiang and Zeng [5]. However, multivariate analysis is only applicable to moderate number of traits, and the power of such analysis may be compromised by the increased degrees of freedom of the distribution of the test statistic required when evaluating significance.

Principal-components analysis (PCA) is used to combine phenotypes and provide scores for subsequent genetic analysis. For example, Dick et al. [6] used PCA to combine quantitative alcohol-related phenotypes and map genes influencing the combined traits in the Collaborative Study on the Genetics of Alcoholism (COGA). However, the usual principal-components analysis ignores the familial aggregation patterns and heritability information in the phenotypes. Since not all of the measured traits may be influenced by genetic factors, it is desirable to provide combined traits with larger weights on those traits that have larger degree of familial aggregation or heritability, because they are more likely to be linked to genetic factors. Principal-components approach based on heritability (PCH) was proposed by Ott and Rabinowitz [7] to exploit the familial information in traits by computing linear combinations of traits that have maximal heritability. Simply put, the PCH approach maximizes the ratio of the relevant family-specific variation to the subject-specific variation instead of maximizing the total variation, so it captures the genetic information across traits. By simulations stuides, Ott and Rabinowitz [7] showed that the PCH approach provides substantial gain of power compared to the usual PCA in some situations.

When the number of traits is very large such as in Morley et al. [2], the PCH method developed in [7] is not directly applicable. Zou and Hastie [7] proposed a sparse principal-components analysis that can incorporate large numbers of traits. However, their approach is the sparse version of the usual PCA and is not appropriate for family data. Here is proposed a penalized principal-components approach based on heritability (PCH_λ) that can be applied to high-dimensional family data. The method stabilizes the PCH estimates by adding a ridge penalty to the subject specific variation. Adding this penalty prevents the problem of over-fitting and regularizes the principal-components of heritability towards the linear combination that maximizes the family-specific variation. The regularization parameter in the penalty term controls the model complexity. A hypothetical example examining the function of the ridge penalty was given. A cross-validation procedure to choose the optimal regularization parameter was also developed. An extension of this approach to provide sparse loadings was briefly discussed. The methods were illustrated through simulation studies. It was shown that the penalized principal-components based on heritability approach can provide scores with large weights on traits with larger degree of familial aggregation for high-dimensional data. Linkage analysis using traits combined by PCH_λ approach showed significant power gain compared to the usual PCA or PCH in [7].

Methods

As in [7], traits Y can be decomposed into a family-specific component B and a subject-specific component W as

$$Y = B + W.$$

When the data consists of nuclear families with multiple siblings, the variation of the family-specific component, Σ_B, can be estimated by the sample between-family variance-covariance matrix and the variation of the subject-specific component, Σ_W, can be estimated by the sample within-family variance-covariance matrix.

The usual principal-components analysis (PCA) searches for linear combination of traits that maximizes the total variation. The leading principal component can be defined as

$$\mathrm{PCA} = \arg \max_{\|\beta\|=1} \beta^T \left(\Sigma_B + \Sigma_W\right)\beta, \quad (1)$$

where β is the score for the leading principal component, and has the same dimension as the vector Y. However, the PCA does not take into account of the family structure information. When family data is available, the PCA can only be applied to traits from independent subjects (founders).

To incorporate family structure information in the traits, a principal-components approach based on heritability was proposed in [7] to find a linear combination of traits that maximizes the family-specific variation relative to the subject-specific varia-

tion. To be precise, the leading principal component of heritability is defined as

$$\mathrm{PCH} = \arg\max_{\|\beta\|=1} \frac{\beta^T \Sigma_B \beta}{\beta^T \Sigma_W \beta}. \quad (1)$$

Note that this maximization is equivalent to maximizing

$$\frac{\beta^T \Sigma_B \beta}{\beta^T \left(\Sigma_B + \Sigma_W\right) \beta},$$

which is the definition of heritability. In this sense, the PCH is also the linear combination that maximizes the heritability. It is well known that the solution to (1) is the first eigenvector of the matrix $\Sigma_W^{-1}\Sigma_B$ (Mardia et al. [9]).

Unfortunately, using eigen-analysis of $\Sigma_W^{-1}\Sigma_B$ to solve for the PCH is not directly applicable to high-dimensional data. To see this, consider the situations where the number of traits is much larger than the number of families. For example, in [2], there were 3,554 trait components and 14 families. Facing these cases, PCH analysis will always find a linear combination that over-fits the observed data: there always exists a linear combination such that the within-family variance of that combination is zero, because Σ_W has eigenvalues of zeros. Subsequently, the denominator of (1) will be zero, so that the PCH will be unidentifiable and extremely unstable. To be precise, any β that satisfies $\beta^T Y_{ij}$ having zero within-family variance will have a very large value for (1), thus β is not identifiable. In this case the large value of (1) is simply an artifact of the numerator being zero, as oppose to providing a good linear combination that have most of its weights on genetic trait components.

To accommodate large number of traits, we propose a ridge penalized principal-components approach based on heritability, where the first component is defined as

$$\mathrm{PCH}_\lambda = \arg\max_{\|\beta\|=1} \frac{\beta^T \Sigma_B \beta}{\beta^T \Sigma_W \beta + \lambda \|\beta\|^2}. \quad (2)$$

Here λ is the regularization parameter to be specified.

The function of adding the ridge penalty is explained in figure 1. Each vertical ellipse corresponds to the within-family variation, and the largest ellipse corresponds to the between-family variation. The upper arrow corresponds to the direction that maximizes the between-family variation and the lower arrow corresponds to the direction that minimizes the within-family variation. The middle arrow corresponds to the direction that provides the optimal balance between maximizing the between-family variation and minimizing the within-family variation, which is also the direction of PCH. When λ is zero, the PCH_λ is the original non-penalized leading principal component of heritability. When λ approaches infinity, the second term in the denominator of (2) dominates, and the PCH_λ approaches the linear combination that maximizes the between-family variation, Σ_B. That is, PCH_λ approaches principal-components of between-family variation, which can be defined as

$$\mathrm{PCB} = \arg\max_{\|\beta\|=1} \beta^T \Sigma_B \beta.$$

When λ ranges from zero to infinity, the PCH_λ changes between the PCH and the PCB. Therefore the effect of the parameter λ is to regularize the PCH_λ towards the direction of the maximal between-family variation, PCB. The PCB summarizes genetic information by considering the variation between the averaged traits between families. It is less optimal to the PCH analysis because the latter considers between-family variation relative to the within-family variation: given the between-family variation being the same, the traits that have more similar values for subjects within the same family (smaller within-family variation) should receive greater weights than the traits that have more distinct values for those subjects, assuming these traits are influenced by genetic factors. A hypothetical numerical example examining the effects of regularization is given in the section 3.

To find the optimal regularization parameter, a cross-validation procedure can be used. However, with no response variables the usual minimizing prediction error criterion is not directly applicable. Instead, here maximizing the 'cross-validated heritability' is used as a criterion. Let k index random partitions of families into halves, and let N denote the total number of the random partitions. Let $\hat{\beta}_\lambda^{1k}$ denote the PCH_λ computed from the first half of the data using λ in the k-th partition. Let Σ_B^{2k} and Σ_W^{2k} denote the between- and within-families variations computed from the second half of the families in the k-th partition. The regularization parameter λ can then be chosen by

$$\mathrm{CV}_\lambda = \arg\max_\lambda \frac{1}{N} \sum_k \frac{\left(\hat{\beta}_\lambda^{1k}\right)^T \Sigma_B^{2k} \hat{\beta}_\lambda^{1k}}{\left(\hat{\beta}_\lambda^{1k}\right)^T \Sigma_W^{2k} \hat{\beta}_\lambda^{1k}}. \quad (3)$$

Asymptotically, the quantity being maximized in (3) is an unbiased estimate of the true heritability, and can be regarded as the 'cross-validated heritability'.

Simulations

Simulation Methods

In all the simulations, the underlying genetic model had a single disease susceptibility locus with two alleles. The effect of the first allele on the traits was assumed to be additive: carrying one copy of the allele added a constant to the mean of the traits. Carrying the other allele had no effect on the traits. The constant effect was the same for each family, but was different across the traits. A fully informative marker perfectly linked to the underlying locus was simulated. Parents' marker genotypes were simulated based on population allele frequencies. Given their genotypes, parents' traits were simulated by adding a multivariate normal random variable to the effect of gene. This model can be expressed as

$$Y = X\mu + \varepsilon, \quad (4)$$

where $\varepsilon \sim MVN(0, \Sigma)$. Here μ is the effect size of the gene, X is the number of the disease susceptible alleles carried by a subject, and Σ is the variance-covariance matrix of the residual multivariate normal random variable. Offsprings' genotypes were simulated based on Mende-

Table 1. Four hypothetical settings to compare PCH with alternatives

Settings	μ	Σ
1	$\begin{pmatrix}1\\1\\1\\1\\1\end{pmatrix}$	$\begin{pmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}$
2	$\begin{pmatrix}1\\1\\0\\0\\0\end{pmatrix}$	$\begin{pmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&2&0.5&0.5\\0&0&0.5&2&0.5\\0&0&0.5&0.5&2\end{pmatrix}$
3	$\begin{pmatrix}1\\1\\2\\2\\2\end{pmatrix}$	$\begin{pmatrix}3.0&0.5&0&0&0\\0.5&3.0&0&0&0\\0&0&1&0.5&0.5\\0&0&0.5&1&0.5\\0&0&0.5&0.5&1\end{pmatrix}$
4	$\begin{pmatrix}1\\1\\2\\2\\2\end{pmatrix}$	$\begin{pmatrix}0.5&0&0&0&0\\0&0.5&0&0&0\\0&0&1&0.5&0.5\\0&0&0.5&1&0.5\\0&0&0.5&0.5&1\end{pmatrix}$

Table 2. Various principal-components under the four settings

	Setting 1	Setting 2	Setting 3	Setting 4
PCA	$\begin{pmatrix}0.447\\0.447\\0.447\\0.447\\0.447\end{pmatrix}$, 0.357[a]	$\begin{pmatrix}0\\0\\0.577\\0.577\\0.577\end{pmatrix}$, 0	$\begin{pmatrix}0.324\\0.324\\0.513\\0.513\\0.513\end{pmatrix}$, 0.375	$\begin{pmatrix}0.224\\0.224\\0.548\\0.548\\0.548\end{pmatrix}$, 0.395
PCB	$\begin{pmatrix}0.447\\0.447\\0.447\\0.447\\0.447\end{pmatrix}$, 0.357	$\begin{pmatrix}0.707\\0.707\\0\\0\\0\end{pmatrix}$, 0.25	$\begin{pmatrix}0.267\\0.267\\0.535\\0.535\\0.535\end{pmatrix}$, 0.380	$\begin{pmatrix}0.267\\0.267\\0.535\\0.535\\0.535\end{pmatrix}$, 0.398
PCH[b]	$\begin{pmatrix}0.447\\0.447\\0.447\\0.447\\0.447\end{pmatrix}$, 0.357	$\begin{pmatrix}0.707\\0.707\\0\\0\\0\end{pmatrix}$, 0.25	$\begin{pmatrix}0.161\\0.161\\0.562\\0.562\\0.562\end{pmatrix}$, 0.383	$\begin{pmatrix}0.603\\0.603\\0.302\\0.302\\0.302\end{pmatrix}$, 0.417
$PCH_{\lambda=1}$	$\begin{pmatrix}0.447\\0.447\\0.447\\0.447\\0.447\end{pmatrix}$, 0.357	$\begin{pmatrix}0.707\\0.707\\0\\0\\0\end{pmatrix}$, 0.25	$\begin{pmatrix}0.186\\0.186\\0.557\\0.557\\0.557\end{pmatrix}$, 0.383	$\begin{pmatrix}0.447\\0.447\\0.447\\0.447\\0.447\end{pmatrix}$, 0.410
$PCH_{\lambda=10}$	$\begin{pmatrix}0.447\\0.447\\0.447\\0.447\\0.447\end{pmatrix}$, 0.357	$\begin{pmatrix}0.707\\0.707\\0\\0\\0\end{pmatrix}$, 0.25	$\begin{pmatrix}0.241\\0.241\\0.543\\0.543\\0.543\end{pmatrix}$, 0.381	$\begin{pmatrix}0.299\\0.299\\0.523\\0.523\\0.523\end{pmatrix}$, 0.400
$PCH_{\lambda=100}$	$\begin{pmatrix}0.447\\0.447\\0.447\\0.447\\0.447\end{pmatrix}$, 0.357	$\begin{pmatrix}0.707\\0.707\\0\\0\\0\end{pmatrix}$, 0.25	$\begin{pmatrix}0.264\\0.264\\0.536\\0.536\\0.536\end{pmatrix}$, 0.380	$\begin{pmatrix}0.271\\0.271\\0.533\\0.533\\0.533\end{pmatrix}$, 0.399

[a] The numbers following the commas are heritabilities $b^T\Sigma_B b/b^T(\Sigma_B + \Sigma_W)b$.
[b] PCH is equivalent to $PCH_{\lambda=0}$.

lian transmissions, and their traits were simulated based on the model (4).

Under this model, the family-specific component Σ_B and the subject-specific component Σ_W can be computed as (see Appendix)

$$\Sigma_B = Cov(X_{11}, X_{12})\mu\mu^T, \quad (5)$$

and

$$\Sigma_W = \Sigma + \{Var(X_{11}) - Cov(X_{11}, X_{12})\}\mu\mu^T, \quad (6)$$

where X_{11} and X_{12}, respectively, are the numbers of the disease susceptible alleles carried by any two siblings from a same family. Let p denote the population frequency of the disease susceptible allele. In the first two sets of simulations, p was assumed to be 1/2. Therefore $\Sigma_B = \mu\mu^T/4$ and $\Sigma_W = \Sigma + \mu\mu^T/4$: The allele frequency was decreased to 0.2 in the third set of simulations involving linkage analysis. It's easy to see from these calculations that when there are no environmental factors, the family-specific component is induced by parental genotype status at the disease susceptibility, while the subject-specific component is induced by the differences of offsprings' inheritance of the parental disease susceptible alleles.

To examine properties of the PCH and the PCH_λ, and compare them with some other alternatives, consider four hypothetical settings summarized in table 1. In each of the four settings, there were five trait items. The first setting corresponds to the scenario in which all five items were of the same importance. In the second setting, only the first two items had genetic effect. In the third setting, the genetic components in the last three items had a larger effect than the first two, but the variance of the noise components for these items which contributed to the within-family variation were smaller. In the fourth setting, the genetic components in the last three items were larger than the first two, and the variance of the noise components in these items were also larger.

Using Σ_B and Σ_W computed from (5) and (6), the values of PCA, PCB, PCH and PCH_λ (with λ = 1, 10 or 100) in the above settings were summarized in table 2. In the first setting, because all items were independent and identically distributed, it was not surprising that all methods chose the same linear combination. In the second

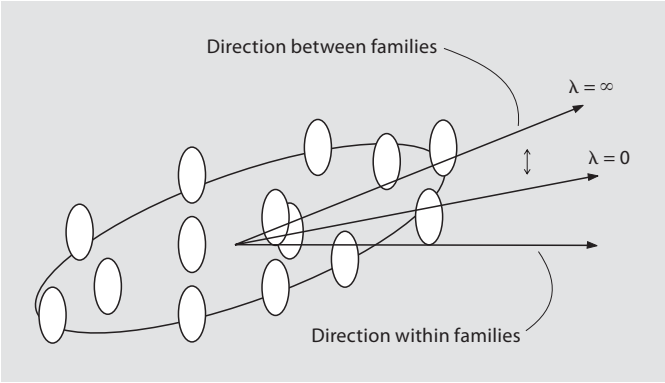

Fig. 1. Interpretation of the regularization parameter λ in PCH_λ.

setting, the usual PCA chose the last three items which had no genetic effect, while the other three methods chose the first two items that had genetic effect. In this setting, the PCA was not an appropriate choice of analysis method. Also note that the PCB and the PCH provided the same results because the normal residual variances contributing to the within-family variation for the two genetic components were identical (equals to 1). In the last two settings, by comparing the heritability, PCH was better than PCA and PCB. This is because the PCH considered the between-family variation relatively to the within-families variation. In the third setting, both PCB and PCH put larger weights on the last three trait items because these items had larger effect than the first two, and also happened to have smaller within-family variation. In the fourth setting, PCH put more weights on the first two items, because they had smaller within-family variations. In contrast, PCB put more weights on the last three items, because the PCB did not capture the within-family variation information, resulting in lower heritability.

It can also be seen from table 2 that when λ was zero, the PCH_λ was equivalent to the PCH; when λ was very large, the PCH_λ approached the PCB; and when λ ranged between zero and infinity, the PCH_λ ranged between the PCH and the PCB. This property was also illustrated in figure 1. The cross-validation procedure (3) can help to determine a best value that balance the need to maximize heritability and the need to stabilize estimates for high-dimensional data.

To examine the performance of the proposed ridge penalized PCH approach, two sets of simulations were carried out under settings 3 and 4. In real life situations, we have no knowledge of which transcripts are controlled by a master QTL and which are not. The simulation experiments were designed to investigate whether the proposed approach can be applied to a range of expressions and distinguishing the genetically influenced components from the noise components. We consider the cases when the noise components outnumber the genetic components, because in real life applications, the genetic components would only be a proportion of all the traits under investigation. In all simulations, only a small proportion of traits were simulated to have genetic effect. Through these experiments, we show the potential of using PCH_λ as a multivariate screening tool on hundreds of transcripts without knowing which ones may be controlled by genetic factors and which ones are independent of the genetic factors. Ideally, the combined traits would have large weights for the genetic traits and small weights for the non-genetic traits, so that the combined trait would have large heritability.

In both simulations, 25 families with four siblings in each family were used. In each simulation, there were 200 replications. In each replicate, λ was chosen by maximizing 'crossed validated heritability' defined in (3). The number of random partitions was 40. Fifty values of λ were equally spaced on the intervals (0, 100) and (0, 120), respectively, in each set of simulation. In each replication, the λ yielding highest cross-validated heritability, CVλ, was chosen to compute PCH_λ.

In the first set of simulations, we augmented the setting 3 in table 1 to simulate trait constellations with 50 and 100 components. In each constellation, the first five components were influenced by a single genetic locus and the other components were random noise with no genetic effect. The augmented vector of the genetic effect size was therefore $\mu = (1, 1, 2, 2, 2, 0, \ldots, 0)^T$. The augmented variance-covariance matrix of the multivariate Gaussian random variable was

$$\begin{pmatrix} \Sigma_s & 0 \\ 0 & \Sigma_e \end{pmatrix}$$

Here Σ_s was the original 5×5 matrix in setting 3 (see table 1), and Σ_e was

$$\begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

In this set of simulations ρ was a 4-dimensional matrix with each element equals to 0.1. In these experiments among the 45 or 95 non-genetic traits, the Gaussian random variables for the first five non-genetic components were correlated, and the remaining 40 or 90 non-genetic traits were independent.

Table 3. Results for the 5 genetic components and the subsequent 5 non-genetic components of PCH_λ: settings 3 and 4[a]

β	50 traits		100 traits	
	mean $(\hat{\beta})$	MSE $(\hat{\beta})$	mean $(\hat{\beta})$	MSE $(\hat{\beta})$
Setting 3				
0.161	0.150	0.032	0.195	0.027
0.161	0.212	0.025	0.248	0.019
0.562	0.578	0.139	0.526	0.180
0.562	0.529	0.150	0.488	0.180
0.562	0.540	0.147	0.524	0.183
0	0.036	0.018	0.053	0.008
0	0.015	0.013	0.022	0.019
0	0.016	0.014	0.088	0.009
0	0.032	0.015	0.008	0.009
0	0.005	0.014	0.027	0.008
Setting 4				
0.603	0.463	0.186	0.392	0.237
0.603	0.529	0.189	0.452	0.240
0.302	0.370	0.029	0.361	0.046
0.302	0.355	0.033	0.311	0.048
0.302	0.418	0.032	0.324	0.047
0	0.006	0.014	0.028	0.009
0	0.021	0.014	0.071	0.010
0	0.018	0.016	0.051	0.010
0	0.065	0.014	0.054	0.008
0	0.019	0.019	0.011	0.010

[a] Only the weights of the first 10 traits were reported.

Table 4. Results for the 5 genetic components and the subsequent 5 non-genetic components of $PCH_\lambda = 0$ (no regularization): setting 3[a]

β	50 Components		100 Components	
	mean $(\hat{\beta})$	MSE $(\hat{\beta})$	mean $(\hat{\beta})$	MSE $(\hat{\beta})$
0.161	0.018	0.034	0.029	0.032
0.161	0.065	0.034	0.002	0.030
0.562	0.066	0.346	0.170	0.340
0.562	0.097	0.357	0.027	0.329
0.562	0.185	0.335	0.059	0.332
0	0.269	0.022	0.036	0.009
0	0.091	0.022	0.003	0.010
0	0.010	0.018	0.047	0.012
0	0.087	0.018	0.261	0.011
0	0.092	0.018	0.030	0.010

[a] Results of PCH for setting 4 were similar to setting 3.

In the second set of simulations, we augmented setting 4 in table 1 to simulate trait constellations with 50 and 100 components. In this set, the variance-covariance matrix of the non-genetic trait components Σ_e had a diagonal structure with elements equal to 0.35. Therefore, in these experiments the Gaussian random variables for all non-genetic components were independent.

The previous two sets of simulations investigate the proposed methods by comparing the estimated loadings of PCH_λ to that of PCA and PCH. We conducted a third set of simulations to investigate the type I error rate and the power of proposed methods through testing of linkage using the combined traits as phenotypes. We generated 100 families, where there were 20 families with 3 siblings, 30 families with 3 siblings, 25 families with 4 siblings and 25 families with 5 siblings. The allele frequency was decreased to 0.2. The number of replications in these analyses was 1,000. In each replication, we computed PCA using standard principal-components analysis, PCH using methods in [7], and PCH_λ using the proposed methods. The λ was selected by cross validation. We then subject the combined traits from these three approaches to a Haseman-Elston regression to test for linkage [10].

The first setting of the linkage analysis was an analogy of scenario 2 and was summarized in table 5, setting 5. We simulated 200 traits with 10 components having positive genetic effect. The mean of the genetic traits was 1.5 and the variance was 0.25. The next 15 non-genetic traits had mean 0, correlation 0.5 and variance 3.0. The remaining non-genetic traits had mean 0 and variance 0.5.

The second setting of the linkage analysis was an analogy of scenario 3 and was summarized in table 5, setting 6. The number of traits was still 200, but there were 25 traits with positive genetic effects. The first 10 traits had mean 1 and variance 3, and the next 15 traits had mean 2 and variance 1. The correlation among these traits was 0.1. The remaining 175 traits were un-correlated and had mean 0 and variance 2.

Simulation Results

The empirical mean and mean-squared-error of the PCH_λ in the first set of simulations were recorded in the top panel of table 3. The first five rows recorded the estimated coefficients for the genetic trait components, while the next 5 rows recorded coefficients for the subsequent non-genetic trait components. The other coefficients for the remaining non-genetic trait components were omitted from the table. The minimal value of these coeffi-

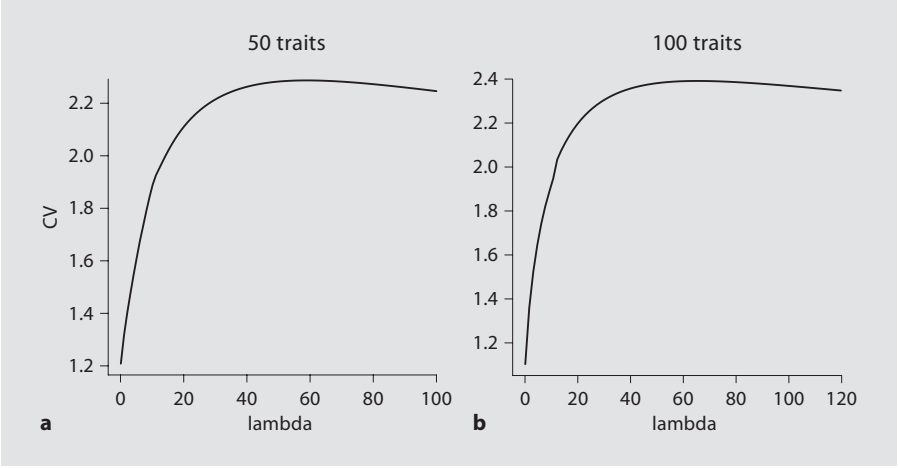

Fig. 2. Cross-validation to choose λ: setting 3, 50 and 100 trait components.

Table 5. Simulation settings for linkage analysis

Settings	μ	Σ
5	$\begin{pmatrix} 1.5_{10} \\ 0_{15} \\ 0_{175} \end{pmatrix}$	$\begin{pmatrix} \mathrm{diag}\left(0.25_{10\times 1}\right) & 0_{10\times 15} & 0_{10\times 175} \\ 0_{15\times 10} & \mathrm{diag}\left(2.5_{15\times 1}\right)+0.5_{15\times 15} & 0_{15\times 175} \\ 0_{175\times 10} & 0_{175\times 15} & \mathrm{diag}\left(0.5_{175\times 1}\right) \end{pmatrix}$
6	$\begin{pmatrix} 1_{10} \\ 2_{15} \\ 0_{175} \end{pmatrix}$	$\begin{pmatrix} \mathrm{diag}\left(2.9_{10\times 1}\right)+0.1_{10\times 10} & 0.1_{10\times 15} & 0_{10\times 175} \\ 0.1_{15\times 10} & \mathrm{diag}\left(0.9_{15\times 1}\right)+0.1_{15\times 15} & 0_{175\times 15} \\ 0_{175\times 10} & 0_{175\times 15} & \mathrm{diag}\left(2_{175\times 1}\right) \end{pmatrix}$

cients was 0.0001 and the maximal was 0.058. As desired, the coefficients for the genetic components were large, and the coefficients for the non-genetic components were small. The behavior of the cross-validation criterion (3) in a typical simulation was depicted in figure 2a. The maxima in this particular simulation occurred at λ = 58.63, and the value of CV_λ in (3) was 2.29.

When increasing the number of traits to 100, the estimated PCH_λ had similar behaviors with having 50 traits. The minimal value of the non-genetic coefficients omitted from the top panel of table 3 was 0.0003, and the maximal was 0.088. The behavior of the cross-validation in a typical simulation depicted in figure 2b also was similar. The maxima was at λ equaled to 65.2, and the value of (3) was 2.39.

The performance of the PCH_λ in the second set of simulations was recorded in the bottom panel of table 3. This set was an augmentation of setting 4. The mean-squared-error for the coefficients were slightly larger than that in the setting 3. The general behavior of the PCH_λ was similar to that in the first set. It can be seen from the both sets of simulations that the proposed PCH_λ method had substantially larger coefficients for the traits with genetic effect than for the non-genetic traits.

It is interesting to compare the penalized PCH with the non-penalized PCH, or $PCH_{\lambda=0}$. The $PCH_{\lambda=0}$ was computed by an eigen-analysis of the generalized inverse of Σ_W multiplied by Σ_B. Here the same set of data that generated the top panel of table 3 was used. The mean and the mean-squared-error of the estimated leading $PCH_{\lambda=0}$ were summarized in table 4. It can be seen that these estimates without regularization were random fluctuations and did not capture the genetic trait components.

Results from the first set of linkage analysis were summarized in the top panel of table 6. The type I error rates from all three approaches were approximated the same as the specified α level. The power for testing for linkage using proposed PCH_λ was above 95% for all values of α levels, while the power using standard principal component analysis (PCA) was only 24% with α level equals to 0.05. The power for PCH was very low and approximately the

Table 6. Results of linkage analysis: settings 5 and 6

Setting	α level	PCH_λ		PCH		PCA	
		Type I error	power	Type I error	power	Type I error	power
5	0.1	0.105	0.986	0.091	0.101	0.094	0.237
	0.05	0.055	0.98	0.046	0.054	0.045	0.163
	0.025	0.029	0.975	0.025	0.028	0.023	0.121
	0.01	0.006	0.971	0.013	0.014	0.011	0.087
	0.005	0.005	0.966	0.009	0.005	0.005	0.071
6	0.1	0.102	0.988	0.109	0.096	0.102	0.533
	0.05	0.052	0.984	0.047	0.055	0.057	0.428
	0.025	0.032	0.977	0.021	0.028	0.028	0.355
	0.01	0.014	0.967	0.006	0.016	0.010	0.300
	0.005	0.005	0.956	0.003	0.007	0.003	0.266

same as the type I error rate. In this setting, the PCA did not capture most of the genetic variability since the variance for the non-genetic traits were larger than the variance for the genetic traits resulting PCA giving more weights to the non-genetic traits. The PCH was computed using generalized inverse because of the singularity of Σ_W. Without regularization, the PCH approach will randomly combine all the traits. Therefore PCH did not capture any genetic variation and the power for testing for linkage using PCH as a phenotype is similar to the type I error rate.

Results from the second set of linkage analysis were summarized in the bottom panel of table 6. The power for PCH_λ was similar to that in the first set. Note that the usual PCA captured more genetic variation in this setting than that in the previous one. However, it still had lower heritability leading to loss of information. The power for testing linkage using PCA was 53% with α level 0.1, which is considerably lower than the power of PCH_λ. The magnitude of power gain of PCH_λ for other α levels is large. PCH with no regularization again had no power for detecting linkage due to the random combination of traits as we also see from setting 5. Both sets of linkage analysis suggest that the ridge regularized PCH has good power for detecting linkage compared to the un-regularized PCH and the usual PCA.

Discussion

Here is proposed a penalized principal-components approach based on heritability that can be applied to high-dimensional traits. The method finds a linear combination of original traits that maximizes the familial-component variation of the trait constellation relative to the regularized subject-component variation. The optimal regularization parameter can be chosen by a cross-validation procedure that maximizes an unbiased estimate of the heritability. The computation of the PCH_λ is fastened by a singular value decomposition of $\Sigma_W + \lambda I = UDV^T$. Without regularization, the matrix D is not invertible. It was shown that the proposed penalized PCH method had scores with substantially larger weight on the genetic trait components, while the PCH analysis without regularization failed to distinguish the genetic trait components from the non-genetic components. Simulations show that using PCH_λ as phenotype in testing for linkage has good power, while using the standard PCA results in loss of power and using PCH has no power at all (when there is a large number of traits thus Σ_W is singular).

Here the variation of the family-specific component was estimated by the sample between-family variance-covariance matrix, and the variation of the subject-specific component was estimated by the sample within-family variance-covariance matrix. For more complicated pedigree data other than nuclear families, the family-specific and subjects-specific component variation can be estimated using the variance components model [11]

$$Y = G + E,$$

where Y is a $k \times p$ matrix of p traits measured from k subjects from a pedigree, G is the unobserved polygenic effect, and E is the residual environmental effect. The covariance of Y can be decomposed as $\Sigma_Y = \Sigma_G \otimes 2\Phi + \Sigma_E \otimes I$, where Φ is the kinship matrix with ϕ_{ij} denoting the kinship coefficient [Chapter 5 of 12] between the i-th

and the j-th subject. The matrix Σ_G is the family-specific or the polygenic effect variation, and Σ_E is the subject-specific variation or the residual environmental effect variation. They can be estimated by maximum likelihood estimation using normal distribution working assumption. After estimating Σ_G and Σ_E, the proposed penalized PCH approach for general pedigrees can then be defined as

$$\arg \max_{\|\beta\|=1} \left(\frac{\beta^T \Sigma_G \beta}{\beta^T \Sigma_E \beta + \lambda \|\beta\|^2} \right).$$

When dealing with extremely large number of traits, it is desirable to obtain PCH_λ with sparse loadings. In these cases, a LASSO (L_1) type penalty can be added. The resulting sparse PCH solves

$$\arg \max_{\|\beta\|=1} \left(\frac{\beta^T \Sigma_B \beta}{\beta^T \Sigma_W \beta + \lambda_1 \|\beta\|^2} - \lambda_2 \sum_k |\beta_k| \right). \quad (7)$$

This is similar to the elastic-net approach introduced by Zou et al. [8]. The cross-validation procedure in (3) can be applied to choose λ_1 and λ_2. However, the computation is intensive because it is required to search λ_1 and λ_2 on a 2-dimensional grid.

After combining traits by the penalized principal-components of heritability approach, genetic analysis can then be applied to the combined phenotypes. When it is desirable to incorporate information from available genotypes to guide the formation of the principal-components, a supervised principal-components analysis as proposed by Bair et al. [13] may be considered. A canonical correlation analysis that searches for a linear combination of traits having maximal correlations with genotypes can also be considered. With high-dimensional data, sparse canonical correlation analysis was proposed in Hastie et al. [14]. However, these methods are computationally intensive.

To interpret results from the combined traits, one would check the loadings for each trait component. Traits with large loadings are expected to have larger genetic effects based on heritability. If the combined traits are mapped to a genomic location (hotspot), these traits with large loadings are expected to be the traits mapped to this hotspot.

In reality, it may be rare that thousands of traits are controlled by a single locus. Here we have only considered the first penalized principal components of heritability. The subsequent second and third PCH_λ capture variation in the traits that is not explained by the first PCH_λ. Using the second or the third PCH_λ as phenotypes in a linkage analysis may reveal a second locus contributing to the variation in traits that is orthogonal to the first PCH_λ.

Directly applying proposed methods to thousands of traits may not be desirable because the number of the noise components may overwhelm limited number of genetic components. Having too many noise components may dilute the effect of genetic traits and cause all trait components to have small weights. In this case, heritability based clustering [15] can be applied first to divide traits into clusters with hundreds of components, and the penalized principal components of heritability approach can be applied to traits within the clusters. Another approach to directly handle thousands of traits is to apply a LASSO type penalty as in (7).

Finally, it is worth to note that the proposed methods can not be used to determine which traits to include for collineared traits. For example, for two perfectly correlated traits, without prior subject information the cross-validation procedure can not distinguish which trait is more important than the other.

An R source code computing PCH_λ and cross-validation is available from the authors upon request.

Appendix

Assume that the frequencies of the two alleles at the disease susceptibility locus are p and q, respectively. Let X_{ij} be the number of the disease susceptible alleles carried by the j-th subject in the i-th family. Considering six different combinations of parental genotypes, one can verify that

$$E(X_{ij}) = 2p^2 + 2pq,\ E(X_{ij}^2) = 4p^2 + 2pq,$$

and

$$E(X_{ij}X_{ik}) = 2p^4 + 9p^3q + 10p^2q^2 + pq^3.$$

Noting that $Var(Y_{ij}) = Var(X_{ij})\mu\mu^T + \Sigma$, it is easy to obtain formulae (4) and (5).

References

1 Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS: Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 2003;33:422–425.

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

3 Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G: Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet 2005;37:225–232.

4 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.
5 Jiang C, Zeng ZB: Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 1995;140:1111–1127.
6 Dick DM, Nurnberger J Jr, Edenberg HJ, Goate A, Crowe R, Rice J, Bucholz KK, Kramer J, Schuckit MA, Smith TL, Porjesz B, Begleiter H, Hesselbrock V, Foroud T: Suggestive linkage on chromosome 1 for a quantitative alcohol-related phenotype. Alcohol Clin Exp Res 2002;26:1453–1460.
7 Ott J, Rabinowitz D: A principal-components approach based on heritability for combining phenotype information. Hum Hered 1999;49:106–111.
8 Zou H, Hastie T, Tibshirani R: Sparse Principal Component Analysis. Technical Report, Department of Statistics, Stanford University, 2004.
9 Mardia KV, Kent JT, Bibby JM: Multivariate Analysis. London, Academic Press, 1979.
10 Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1972;2:3–19.
11 Almasy L, Dyer TH, Blangero J: Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genet Epidemiol 1997;14:953–958.
12 Lange K: Mathematical and statistical methods for genetic analysis.New York, Springer Verlag, 2002.
13 Bair E, Hastie T, Paul D, Tibshirani R: Prediction by supervised principal components. J Am Stat Assoc 2005;473:119–137.
14 Hastie T, Buja A, Tibshirani R: Penalized discriminant analysis. Ann Stat 1995;23:73–102.
15 Wang Y, Fang Y, Wang S: Clustering and principal component analysis for mapping co-regulated genome-wide variation using family data. BMC Genet, in press.