Comput Stat (2014) 29:455–465
DOI 10.1007/s00180-013-0444-3

ORIGINAL PAPER

# Regularized principal components of heritability

**Yixin Fang · Yang Feng · Ming Yuan**

Received: 2 October 2012 / Accepted: 18 July 2013 / Published online: 23 August 2013


**Abstract** In family studies with multiple continuous phenotypes, heritability can be conveniently evaluated through the so-called principal-component of heredity (PCH, for short; Ott and Rabinowitz in Hum Hered 49:106–111, 1999). Estimation of the PCH, however, is notoriously difficult when entertaining a large collection of phenotypes which naturally arises in dealing with modern genomic data such as those from expression QTL studies. In this paper, we propose a regularized PCH method to specifically address such challenges. We show through both theoretical studies and data examples that the proposed method can accurately assess the heritability of a large collection of phenotypes.

**Keywords** Expression quantitative trait loci · Family study · High dimensional data · Linear discriminant analysis · Principal components · Sparsity

## 1 Introduction

For many common diseases, defining genetically relevant phenotypes and appropriately assessing their heritability is important yet challenging (e.g., Winawer 2006).

The research of Ming Yuan was supported in part by NSF Career Award DMS-0846234 and FRG Award DMS-1265202.

Y. Fang
New York University, New York, NY, USA

Y. Feng
Columbia University, New York, NY, USA

M. Yuan (✉)
Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA
e-mail: myuan@isye.gatech.edu

A common practice is to first obtain information on a variety of phenotypes, and then seek the most genetically relevant phenotypes by combining them through the so-called principal-components of heredity (Ott and Rabinowitz 1999).

Let $d$ be the number of phenotypes under consideration, and let $d$ dimensional vector $Y_{ij}$ be the collection of phenotypes for subject $j$ in family $i$ such that

$$Y_{ij} = \mu + A_i + E_{ij}, \qquad i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i, \tag{1}$$

where $n$ is the number of families in the sample, and $m_i$ is the number of subjects in family $i$. It is clear that the heritability of any linear combination of phenotypes, $\beta^{\mathsf{T}} Y_{ij}$, can be given as

$$h(\beta) := \frac{\beta^{\mathsf{T}} \Sigma_A \beta}{\beta^{\mathsf{T}} (\Sigma_A + \Sigma_E) \beta}, \tag{2}$$

where $\Sigma_A$ and $\Sigma_E$ are the covariance matrix of the family effect $A$ and subject effect $E$ respectively. The overall heritability of the phenotypes can then be assessed by

$$h_{\max} := \max_{\beta : \|\beta\| = 1} h(\beta),$$

where $\|\cdot\|$ is Euclidean norm, and the maximizer of $\beta$, denoted by $\beta_0$, is referred to as the principal component of heritability (PCH, for short).

In practice, covariance matrices $\Sigma_A$ and $\Sigma_E$ are often estimated by their respective sample version, leading to the sample PCH. More specifically, let $\overline{Y}_{i\cdot} = \sum_j Y_{ij}/m_i$, $\overline{Y}_{\cdot\cdot} = \sum_i \sum_j Y_{ij}/N$, and $N = \sum_{i=1}^n m_i$. The sample covariance matrices of the family and subject effects are given by

$$\widehat{\Sigma}_E = \sum \sum (Y_{ij} - \overline{Y}_{i\cdot})(Y_{ij} - \overline{Y}_{i\cdot})^{\mathsf{T}}/(N - n),$$

and

$$\widehat{\Sigma}_A = \sum \sum (Y_{ij} - \overline{Y}_{\cdot\cdot})(Y_{ij} - \overline{Y}_{\cdot\cdot})^{\mathsf{T}}/(N - 1) - \widehat{\Sigma}_E,$$

respectively. The sample PCH is then defined as

$$\widehat{\beta}^{\,\text{Sample}} = \arg \max_{\beta : \|\beta\| = 1} h_n(\beta),$$

where

$$h_n(\beta) = \frac{\beta^{\mathsf{T}} \widehat{\Sigma}_A \beta}{\beta^{\mathsf{T}} (\widehat{\Sigma}_A + \widehat{\Sigma}_E) \beta}.$$

For brevity, in what follows, we shall write $\Sigma_T = \Sigma_A + \Sigma_E$ and correspondingly

$$\widehat{\Sigma}_T = \sum \sum (Y_{ij} - \overline{Y}_{\cdot\cdot})(Y_{ij} - \overline{Y}_{\cdot\cdot})^{\mathsf{T}}/(N - 1).$$

It is clear that $h(\beta)$ is the Rayleigh quotient of positive definite matrices $\Sigma_A$ and $\Sigma_T$. Similarly, the sample heritability measurement $h_n(\beta)$ is the Rayleigh quotient of positive definite matrices $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_T$.

Although the sample PCH approach is effective when dealing with a handful of phenotypes, it may perform rather poorly when the number of phenotypes is large (see, e.g., Jin and Fang 2011). In particular, $\hat{\beta}^{\text{Sample}}$ is only well defined if $\widehat{\Sigma}_T$ is of full rank, or in other words, the number of phenotypes to be smaller than $N$. This assumption, however, may not be appropriate for many modern genomic studies. Consider, for example, an expression QTL experiment where the expression levels of thousands of genes are treated as the so-called "gene expression phenotypes" (Cheung et al. 2003). Given that a typical experiment has only hundreds of subjects, the sample PCH approach cannot be applied.

To overcome this problem, we propose in this paper a novel regularization approach to estimating the PCH $\beta_0$. The approach is based on the notion that $\beta_0$ is sparse and therefore can be well approximated by linear combination of only a small number of phenotypes. Sparsity is a common phenomenon in high dimensional problems and a plausible assumption in most applications. We show that by appropriately exploiting the sparsity of $\beta_0$, the proposed regularization estimator can provide an accurate description of the overall heredity of a large number of phenotypes.

The rest of the paper is organized as follows. In Sect. 2, we describe the details of the proposed PCH estimate. In Sect. 3, we evaluate the empirical performance of the proposed method by both simulations and two real data analyses. We present a short discussion in Sect. 4 and relegate all technical details relegated to the "Appendix".

## 2 Regularized PCH

To exploit the sparsity of the PCH, we consider, as an alternative to the usual sample PCH, the following regularized principal component of heritability:

$$\widehat{\beta}(\lambda) = \arg\max_{\|\beta\|=1} h_{n,\lambda}(\beta), \tag{3}$$

where

$$h_{n,\lambda}(\beta) = \frac{\beta^{\mathsf{T}}\widehat{\Sigma}_A\beta}{\beta^{\mathsf{T}}\widehat{\Sigma}_T\beta + \lambda\|\beta\|_{\ell_1}^2}$$

where $\|\cdot\|_{\ell_1}$ stands for the vector $\ell_1$ norm and $\lambda \geq 0$ is a tuning parameter to be determined later. Obviously $\widehat{\beta}(\lambda)$ reduces to the usual sample PCH when $\lambda = 0$. By adding an extra term proportional to the squared $\ell_1$ norm of $\beta$, we encourage sparsity in the resulting PCH estimate. Moreover, the benefit of using the squared $\ell_1$ norm of $\beta$ is that $\widehat{\beta}(\lambda)$ is scale invariant if the units of phenotypes are changed. We first provide some theoretical justification to the proposed estimate $\widehat{\beta}(\lambda)$.

### 2.1 Theoretical properties

Since the PCH $\beta_0$ is a $d$ dimensional unit length vector, write $\beta_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0d})^{\mathsf{T}}$. We shall assume that $\beta_0$ is sparse in that most of $\beta_{0j}$s are zero. In particular, write

$$\mathrm{supp}(\beta_0) = \{1 \le j \le d : \beta_{0j} \ne 0\}$$

and denote by $s_0$ the cardinality of $\mathrm{supp}(\beta_0)$.

**Theorem 1** *Assume that* $s_0 = o((n/\log d)^{1/2})$ *and* $(n^{-1}\log d)^{1/2} \ll \lambda \ll s_0^{-1}$. *Then*

$$h(\widehat{\beta}(\lambda)) \to_p h_{\max}.$$

*In addition,*

$$h_{n,\lambda}(\widehat{\beta}(\lambda)) \to_p h_{\max}.$$

The first part of Theorem 1 suggests that, with appropriate choice of the tuning parameter, $\widehat{\beta}(\lambda)$ would indeed provide an accurate summary of the overall heritability. The second statement indicates that the overall heritability $h_{\max}$ can also be consistently estimated by $h_{n,\lambda}(\widehat{\beta}(\lambda))$.

### 2.2 Computation

We now describe how $\widehat{\beta}(\lambda)$ can be computed in practice. Following Fan et al. (2012), we consider the following approximation to (3),

$$\widetilde{\beta}(\lambda) = \arg\min \left\{ \beta^{\mathsf{T}} \widehat{\Sigma}_T \beta + \lambda \|\beta\|_{\ell_1}^2 + \gamma(\beta^{\mathsf{T}} \widehat{\Sigma}_A \beta - 1)^2 \right\}, \tag{4}$$

for some $\gamma > 0$. It is clear that $\widetilde{\beta}(\lambda)/\|\widetilde{\beta}(\lambda)\| \to \widehat{\beta}(\lambda)$ as $\gamma \to \infty$. In practice, our experience suggests that the two become fairly close for moderate or large $\gamma$. Similar observations were also made in a different context by Fan et al. (2012).

We note that the minimization problem in defining $\widetilde{\beta}(\lambda)$ is in general non-convex. We consider solving it by a coordinate descent type of algorithm. Without loss of generality, suppose that the first component of $\beta$, $\beta_{(1)}$, is being updated, and the remaining components, stacked as $\beta_{(-1)}$, are given. Accordingly, rewrite

$$\widehat{\Sigma}_A = \begin{pmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad \text{and} \quad \widehat{\Sigma}_T = \begin{pmatrix} t_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}.$$

Then the objective function in terms of $x = \beta_{(1)}$ becomes

$$g(x) = t_{11}x^2 + 2t_{12}^* x + t_{22}^* + \lambda(|x| + r^*)^2 + \gamma(a_{11}x^2 + 2a_{12}^* x + a_{22}^* - 1)^2, \tag{5}$$

where $r^* = \|\beta_{(-1)}\|_{\ell_1}$, $t_{12}^* = T_{12}\beta_{(-1)}$, $t_{22}^* = \beta_{(-1)}^{\mathsf{T}} T_{22}\beta_{(-1)}$, $a_{12}^* = A_{12}\beta_{(-1)}$, and $a_{22}^* = \beta_{(-1)}^{\mathsf{T}} A_{22}\beta_{(-1)}$. The objective function $g$ can then be minimized analytically. Details are given in the "Appendix". We iteratively update the coordinates until a certain convergence criterion is met. In the numerical experiments conducted in Sect. 3, we stop the iteration when $|G(\widetilde{\beta}^{\text{new}}) - G(\widetilde{\beta}^{\text{old}})|/G(\widetilde{\beta}^{\text{old}}) < 10^{-6}$, where $G(\beta)$ is the objective function in (4).

### 2.3 Tuning

Although our general theoretical results from Sect. 2.1 suggests a fairly wide range of choices of $\lambda$ would be sufficient for our purposes, in practice, fine tuning of $\lambda$ may lead to further improved finite sample performance. To this end, we introduce a bootstrap based method for choosing $\lambda$. The basic idea is that we compute $\widetilde{\beta}(\lambda)$ for a fine grid of $\lambda$s, evaluate the heritability of the resulting estimated PCH for each $\lambda$, and then choose the one that yields the highest heritability.

Let $Y_i = \{y_{i1}, \ldots, y_{im_i}\}$ be the set of observations from family $i$. Assume that $\mathbf{X} = \{Y_1, \ldots, Y_n\}$ is sampled from a population $\mathcal{P}$. To emphasize the dependence on $\mathcal{P}$, rewrite $\Sigma_A$ and $\Sigma_T$ as $\Sigma_A(\mathcal{P})$ and $\Sigma_T(\mathcal{P})$, and to emphasize the dependence on $\mathbf{X}$, rewrite $\widetilde{\beta}(\lambda)$ as $\widetilde{\beta}_\lambda(\mathbf{X})$. Let $\widehat{\mathcal{P}}$ be the empirical distribution with probability $1/n$ on each $Y_i$. The oracle guided optimal $\lambda$ is

$$\lambda_{\text{oracle}} = \arg\max_\lambda h(\widetilde{\beta}_\lambda(\mathbf{X})). \tag{6}$$

To estimate the expected heritability associated with $\widetilde{\beta}_\lambda(\mathbf{X})$,

$$\theta_\lambda(\mathcal{P}) = E_{\mathbf{X}\sim\mathcal{P}}\left\{\frac{\widetilde{\beta}_\lambda^{\mathsf{T}}(\mathbf{X})\Sigma_A(\mathcal{P})\widetilde{\beta}_\lambda(\mathbf{X})}{\widetilde{\beta}_\lambda^{\mathsf{T}}(\mathbf{X})\Sigma_T(\mathcal{P})\widetilde{\beta}_\lambda(\mathbf{X})}\right\}, \tag{7}$$

we advocate the plug-in estimate of $\theta_\lambda(\mathcal{P})$ (Efron 1979),

$$\theta_\lambda(\widehat{\mathcal{P}}) = E_{\mathbf{X}^*\sim\widehat{\mathcal{P}}}\left\{\frac{\widetilde{\beta}_\lambda^{\mathsf{T}}(\mathbf{X}^*)\Sigma_A(\widehat{\mathcal{P}})\widetilde{\beta}_\lambda(\mathbf{X}^*)}{\widetilde{\beta}_\lambda^{\mathsf{T}}(\mathbf{X}^*)\Sigma_T(\widehat{\mathcal{P}})\widetilde{\beta}_\lambda(\mathbf{X}^*)}\right\}. \tag{8}$$

Noting that $\Sigma_A(\widehat{\mathcal{P}}) = \widehat{\Sigma}_A$ and $\Sigma_T(\widehat{\mathcal{P}}) = \widehat{\Sigma}_T$, we have the following bootstrap procedure.

*Bootstrap procedure for selecting* $\lambda$

*Step* 1. Generate $B$ independent bootstrap samples, $\mathbf{X}_b^*$, $b = 1, \ldots, B$.

*Step* 2. Obtain solutions, $\widetilde{\beta}_\lambda(\mathbf{X}_b^*)$, $b = 1, \ldots, B$, over a same grid of $\lambda$s.

*Step* 3. Estimate $h(\widetilde{\beta}_\lambda(\mathbf{X}))$ by $\widehat{h}(\widetilde{\beta}_\lambda(\mathbf{X})) = \frac{1}{B}\sum_{b=1}^B \frac{\widetilde{\beta}_\lambda^{\mathsf{T}}(\mathbf{X}_b^*)\widehat{\Sigma}_A\widetilde{\beta}_\lambda(\mathbf{X}_b^*)}{\widetilde{\beta}_\lambda^{\mathsf{T}}(\mathbf{X}_b^*)\widehat{\Sigma}_T\widetilde{\beta}_\lambda(\mathbf{X}_b^*)}$.

*Step* 4. Select $\lambda$ as $\lambda_{\text{boot}} = \arg\max_\lambda \widehat{h}(\widetilde{\beta}_\lambda(\mathbf{X}))$.

Since the bootstrap is used for approximating the expectation only, moderate $B$, say 10 or 20, will be good enough. In practice, we should also consider "one-standard-error" rule (e.g., Hastie et al. 2009, p. 244).

**Table 1** An example of the insensitiveness of $\gamma$ in (4)

| $\gamma$ | $h(\widehat{\beta})$ | $\#\{j : \widehat{\beta}_j \neq 0\}$ | $\lvert\widehat{\beta}_1\rvert$ | $\lvert\widehat{\beta}_6\rvert$ |
|---|---|---|---|---|
| 10 | 0.548 | 8.65 | 0.444 | 0.004 |
| 20 | 0.549 | 8.35 | 0.440 | 0.000 |
| 30 | 0.543 | 8.00 | 0.438 | 0.000 |
| 50 | 0.542 | 8.30 | 0.455 | 0.000 |
| 100 | 0.537 | 8.75 | 0.417 | 0.000 |

* $\widehat{\beta} = \widetilde{\beta}(\lambda_{\text{oracle}})$

## 3 Numerical results

### 3.1 Simulations

We conduct extensive simulations to examine the performances of the regularized PCH approach, compared with the usual sample PCH approach. The data are generated from model (1), with $A_i$ and $E_{ij}$ generated from normal distributions. We consider the following three settings, where $\sigma_a$s are chosen to obtain heritability $h(\beta_0) = 0.55$.

*Setting 1 (independent).* Let $\Sigma_A = \sigma_a^2 \nu\nu^{\mathsf{T}}$ and $\Sigma_E = I_d$, where $\nu = (\mathbf{1}_5^{\mathsf{T}}, \mathbf{0}_{d-5}^{\mathsf{T}})^{\mathsf{T}}$ and $\sigma_a = 0.5$. Let $n = 100, m_i \equiv 4$, and $d = 50, 100$, or 500. In this setting, $\beta_0 = \nu/\|\nu\|$.

*Setting 2 (auto regression).* Let $\Sigma_A = \sigma_a^2 \nu\nu^{\mathsf{T}}$ and $\Sigma_E = (e_{kl})$, where $\nu = (\mathbf{1}_5^{\mathsf{T}}, \mathbf{0}_{d-5}^{\mathsf{T}})^{\mathsf{T}}$ and $e_{kl} = \rho^{|k-l|}$, for $1 \le k, l \le d$. Let $n = 100, m_i \equiv 4, d = 100$, and $(\rho, \sigma_a) = (0.2, 0.581)$, $(0.5, 0.681)$, or $(0.8, 0.623)$. In this setting, the first five components of $\beta_0$ are positive, the sixth component is negative, and the others are zero.

*Setting 3 (equal correlation).* Let $\Sigma_A = \sigma_a^2 \nu\nu^{\mathsf{T}}$ and $\Sigma_E = (e_{kl})$, where $\nu$ is to be decided, $e_{kl} = \rho$ and $e_{kk} = 1$ for $1 \le k \ne l \le p$. Let $n = 100, m_i \equiv 4, d = 100$, and $(\rho, \sigma_a) = (0.2, 0.671)$, $(0.5, 0.866)$, or $(0.8, 1.025)$. In this setting, in order to get a sparse $\beta_0$ like the one in Setting 1, let $\widetilde{\nu} = \Sigma_E \beta_0$ and $\nu$ is the scaled $\widetilde{\nu}$ with first component being 1.

First, we conduct a simulation to show that the approximation (4) is not sensitive to $\gamma$ by generating 100 repetitions under Setting 1 with $d = 100$. Under this setting, $\beta_0 = 0.447(\mathbf{1}_5^{\mathsf{T}}, \mathbf{0}_{95}^{\mathsf{T}})^{\mathsf{T}}$. Given $\gamma = 10, 20, 30, 50$, or 100, $\lambda$ is chosen as $\lambda_{\text{oracle}}$ in (6), and for convenience denote the solution $\widetilde{\beta}(\lambda_{\text{oracle}})$ as $\widehat{\beta}$. Table 1 summarizes the average heritability and Rayleigh quotient associated with $\widehat{\beta}$, average angle between $\widehat{\beta}$ and $\beta_0$, average number of non-zero components of $\widehat{\beta}$, and average absolute values of its 1st and 6th components. This simulation backs up our claim and hereafter we use $\gamma = 20$.

Then we report the main simulation results in Tables 2, 3 and 4. For each setting, 100 repetitions are generated. For the regularized PCH, tuning parameter $\lambda$ is selected via the proposed bootstrap procedure with $B = 20$ (the one-standard-error rule is applied). Note that for heritability, the closer to 0.55 the better, and for angle, the closer to 0 the better. The numbers of selected phenotypes and the false negative rates are also reported.

**Table 2** Simulation Setting 1

| $d$ | Method | $h(\widehat{\beta})$[a] | $\text{Ang}(\widehat{\beta}, \beta_0)$ | $\#\{j : \widehat{\beta}_j \neq 0\}$ | FN[c] |
|---|---|---|---|---|---|
| 50 | PCH | 0.456 (0.025)[b] | 34.776 (4.088) | 50 (0) | 0 |
| | Regularized PCH | 0.539 (0.011) | 14.082 (4.539) | 10.80 (4.81) | 0 |
| 100 | PCH | 0.324 (0.050) | 51.456 (5.098) | 100 (0) | 0 |
| | Regularized PCH | 0.534 (0.011) | 16.348 (4.246) | 7.94 (4.30) | 0 |
| 500 | PCH | 0.002 (0.003) | 88.130 (1.441) | 500 (0) | 0 |
| | Regularized PCH | 0.531 (0.014) | 17.274 (4.899) | 22.75 (14.86) | 0 |

[a] $\widehat{\beta} = \widetilde{\beta}(\lambda_{\text{boot}})$; [b] Average (standard deviation); [c] Average false negatives

**Table 3** Simulation Setting 2

| $\rho$ | Method | $h(\widehat{\beta})$ | $\text{Ang}(\widehat{\beta}, \beta_0)$ | $\#\{j : \widehat{\beta}_j \neq 0\}$ | FN |
|---|---|---|---|---|---|
| 0.2 | PCH | 0.332 (0.044) | 55.485 (4.349) | 100 (0) | 0 |
| | Regularized PCH | 0.532 (0.011) | 20.170 (4.667) | 21.07 (9.25) | 0 |
| 0.5 | PCH | 0.319 (0.057) | 60.466 (6.148) | 100 (0) | 0 |
| | Regularized PCH | 0.523 (0.017) | 25.347 (7.594) | 22.04 (9.36) | 0.12 |
| 0.8 | PCH | 0.332 (0.051) | 54.394 (5.493) | 100 (0) | 0 |
| | Regularized PCH | 0.501 (0.096) | 17.907 (11.359) | 21.07 (11.83) | 1.16 |

**Table 4** Simulation Setting 3

| $\rho$ | Method | $h(\widehat{\beta})$ | $\text{Ang}(\widehat{\beta}, \beta_0)$ | $\#\{j : \widehat{\beta}_j \neq 0\}$ | FN |
|---|---|---|---|---|---|
| 0.2 | PCH | 0.332 (0.054) | 60.705 (4.762) | 100 (0) | 0 |
| | Regularized PCH | 0.536 (0.012) | 22.943 (7.795) | 12.43 (7.57) | 0.03 |
| 0.5 | PCH | 0.328 (0.059) | 71.636 (6.191) | 100 (0) | 0 |
| | Regularized PCH | 0.527 (0.011) | 49.711 (11.422) | 19.57 (7.20) | 0.59 |
| 0.8 | PCH | 0.320 (0.061) | 80.292 (4.820) | 100 (0) | 0 |
| | Regularized PCH | 0.534 (0.007) | 75.801 (10.406) | 21.76 (7.48) | 2.67 |

From Table 2, we see that the regularized PCH performs much better than the usual PCH, especially when the dimension $d$ is large, in that the regularized PCH, $\widetilde{\beta}(\lambda_{\text{boot}})$, has heritability close to $h_{\max} = 0.55$ and is close to $\beta_0$. In addition, the sparsity property of the regularized PCH is promising, because it is easier to interpret the results when only a few phenotypes are involved. Moreover, when achieving the sparsity, the regularized PCH have never missed the genetically related phenotypes, as shown by the zero false negative rate.

From Tables 3 and 4, we also see that the regularized PCH performs much better than the usual PCH. However, the regularized PCH misses some important phenotypes in a few repetitions, especially under Setting 3 when $\rho = 0.8$, where the $\widetilde{\beta}(\lambda_{\text{boot}})$ is not very close to $\beta_0$. One explanation is that when the important phenotypes and the unimportant phenotypes are highly correlated, it is hard to distinguish them.

### 3.2 Asthma data

Asthma is a complex disease that is likely genetically heterogeneous. The asthma data used here were originally collected as part of the Collaborative Study on the Genetics of Asthma (CSGA 1997). As in Reilly et al. (2007), we use only the 27 multi-generational Caucasian families that were collected in Minnesota. These families had 169 asthmatic members, and the average family size was 6.3. Four phenotypes considered in Reilly et al. (2007) were the logarithm of the percent predicted of the following variables: volume exhaled during the first second of a forced expiratory maneuver (FEV1), forced expiratory vital capacity (FVC), maximum expiratory flow when half of the FCV has been exhaled (FEFM) and forced expiratory flow rate over the middle half FCV (FF25).

The F-test statistics for these four phenotypes are 1.759, 2.173, 1.281, and 2.102, respectively. When we apply the regularized PCH starting from $\lambda_{\max} = 5.002$ (formula for $\lambda$ is in the "Appendix"), first FVC enters, then FF25, and then FEFM. The bootstrap procedure (without one-standard-error rule because of low-dimension) selects $\lambda_{\text{boot}} = 0.048$, and the final selected subset is (FVC, FF25, FEFM).

### 3.3 Gene-expression data

The dataset was provided by Genetic Analysis Workshop 15 in 2006, and it was originally analyzed in Cheung et al. (2003, 2005) and Morley et al. (2004). In the dataset, gene expression levels of 3,554 genes in 14 large families were measured. There were 194 subjects in the sample and the average family size was 13.9.

As pointed out by Morley et al. (2004), "the correlation in expression level of these genes supports the observation that they share common transcriptional regulators. However, the regulatory regions defined by mapping are still large, and these might be subgroups of co-regulated phenotypes that are influenced by distinct, but every closely linked, regulators." Therefore, linear combinations of these phenotypes showing high familial aggregation can have larger power in linkage analysis (Wang et al. 2007).

First we conduct F-test for 3,554 gene expressions individually. The five-number summary of the F scores is: 0.26 (Min), 1.69 (1st Quartile), 2.48 (Med), 3.44 (3rd Quartile) and 18.43 (Max). To demonstrate the application of the regularized PCH, for simplicity, here we consider only the top 300 gene expressions of the largest F-scores (range 4.83–18.43). In practice, it is also reasonable to consider such reduced set of phenotypes, in the spirit of Sure Independence Screening (SIS; Fan and Lv 2008).

For this reduced set of phenotypes, we apply the regularized PCH starting with $\lambda_{\max} = 20.6011$. When $\lambda$ decreases, phenotypes are entering the model one by one, with the phenotype of the largest F score entering first. The bootstrap procedure selects $\lambda$ as $\lambda_{\text{boot}} = 0.0095$ and only 59 components of $\widetilde{\beta}(\lambda_{\text{boot}})$ are non-zero. The bootstrap selection of $\lambda$ and the normalized $\widetilde{\beta}(\lambda_{\text{boot}})$ are displayed in Fig. 1.

## 4 Discussion

Here the regularized PCH approach is proposed to remedy the usual sample PCH approach for the situations where the number of phenotypes is large. The idea is to

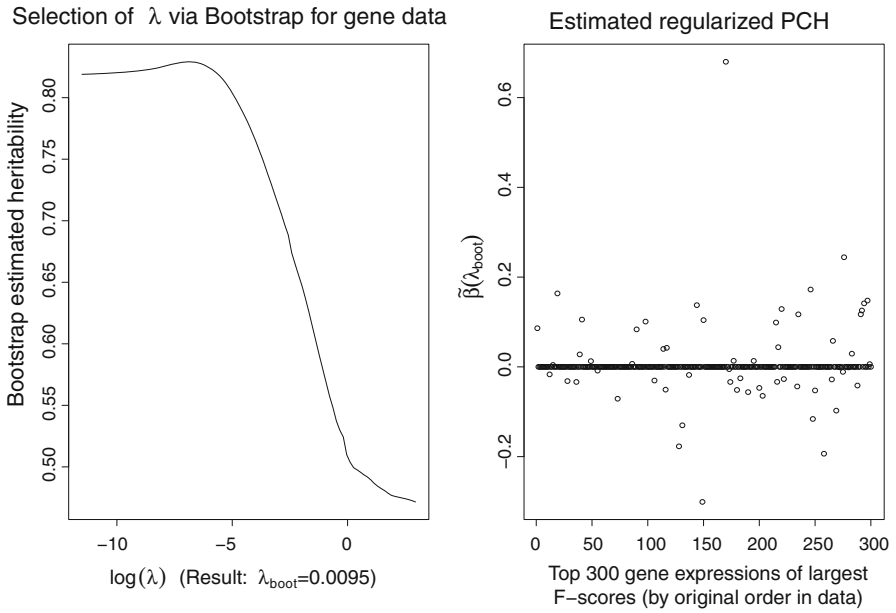


**Fig. 1** Gene data

add a novel penalty term to the usual sample PCH approach, based on the notion that the PCH $\beta_0$ is sparse. Such regularization is a common procedure in high dimensional problems.

We use the bootstrap procedure to select the tuning parameter. Another option is the cross-validation procedure proposed in Wang et al. (2007), where data were randomly divided into two halves, one for training and the other for examining the heritability.

It would be interesting to further develop the regularized PCH approach to obtain the $k$th principal component of heritability, $k = 2, 3, \ldots$, which is orthogonal to the previous obtained $k - 1$ principal components of heritability.

## 5 Appendix

### 5.1 Proof of theorem

Assume the average family size $N/n \to m_0$ and $\min_{\|\beta\|=1} \beta^{\mathsf{T}} \Sigma_T \beta \geq \delta > 0$. Rewrite

$$h_{n,\lambda}(\beta) = \frac{\beta^{\mathsf{T}} \Sigma_A \beta + \beta^{\mathsf{T}} \Delta_A \beta}{\beta^{\mathsf{T}} \Sigma_T \beta + \beta^{\mathsf{T}} \Delta_T \beta + \lambda \|\beta\|_1^2},$$

where $\Delta_A = \widehat{\Sigma}_A - \Sigma_A$ and $\Delta_T = \widehat{\Sigma}_T - \Sigma_T$. Under some mild conditions, we have (e.g., Bickel and Levina 2004) $\|\Delta_A\|_\infty = O_p(\sqrt{(\log d)/n})$ and $\|\Delta_T\|_\infty = O_p(\sqrt{(\log d)/n})$, where $\|\cdot\|_\infty$ is the element-wise super-norm. From $\beta^{\mathsf{T}} \Delta_A \beta \leq$

$\|\beta\|_{\ell_1}\|\Delta_A\beta\|_\infty \le \|\Delta_A\|_\infty\|\beta\|_{\ell_1}^2$, we have $\beta^{\mathsf{T}}\Delta_A\beta \le O_p(\sqrt{(\log d)/(n-1)})\|\beta\|_{\ell_1}^2$. Similarly, $\beta^{\mathsf{T}}\Delta_T\beta \le O_p(\sqrt{(\log d)/(N-n)})\|\beta\|_{\ell_1}^2$.

Thus, if $\lambda \gg O_p(\sqrt{(\log d)/n})$ and $\min_{\|\beta\|=1}\beta^{\mathsf{T}}\Sigma_T\beta \ge \delta > 0$, we have $\max_\beta |h_{n,\lambda}(\beta) - h_{0,\lambda}(\beta)| = o_p(1)$, where $h_{0,\lambda}(\beta) = \beta^{\mathsf{T}}\Sigma_A\beta/(\beta^{\mathsf{T}}\Sigma_T\beta + \lambda\|\beta\|_{\ell_1}^2)$. Further, because $\lambda\|\beta\|_{\ell_1}^2 \le \lambda s_0\|\beta\|^2$, if $\lambda \ll 1/s_0$ and $\min_{\|\beta\|=1}\beta^{\mathsf{T}}\Sigma_T\beta \ge \delta > 0$, we have $\max_\beta |h_{0,\lambda}(\beta) - h(\beta)| = o(1)$. Together, we have

$$\max_\beta |h_{n,\lambda}(\beta) - h(\beta)| = o_p(1).$$

Therefore, noting that $0 \le h(\beta_0) - h(\widehat{\beta}(\lambda)) \le h(\beta_0) - h_{n,\lambda}(\beta_0) + h_{n,\lambda}(\widehat{\beta}(\lambda)) - h(\widehat{\beta}(\lambda))$, we have $h(\widehat{\beta}(\lambda)) \to_p h_{\max}$, and noting that $h_{n,\lambda}(\beta_0) - h(\beta_0) \le h_{n,\lambda}(\widehat{\beta}(\lambda)) - h(\beta_0) \le h_{n,\lambda}(\widehat{\beta}(\lambda)) - h(\widehat{\beta}(\lambda))$, we have $h_{n,\lambda}(\widehat{\beta}(\lambda)) \to_p h_{\max}$.

### 5.2 The coordinate descent algorithm

We discuss the path solution to the optimization problem (4). Write two input matrices $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_T$ as $(a_{ij})$ and $(t_{ij})$ respectively. Note that when $\lambda$ is larger than some value say $\lambda_{\max}$, $\widetilde{\beta}(\lambda) = \mathbf{0}_d$. Here we derive the formula for $\lambda_{\max}$. If $\beta_{(-1)} = \mathbf{0}_{d-1}$, then $g(x) = t_{11}x^2 + \lambda x^2 + \gamma(a_{11}x^2 - 1)^2$, which achieves minimum at $x = \sqrt{[-\frac{t_{11}+\lambda-2a_{11}\gamma}{2\gamma a_{11}^2}]_+}$, where $[a]_+$ equals $a$ if $a \ge 0$ and zero otherwise. Therefore,

$$\lambda_{\max} = \max_{1\le j\le d}\{2a_{jj}\gamma - t_{jj}\}.$$

On a fine grid of $L$ values of $\lambda$, say $\exp\{\log(\lambda_{\max})(1:L)/L\}$, we computer the solution path starting with $\lambda = \lambda_{\max}$ backwardly and $\widetilde{\beta}(\lambda_{\max}) = \mathbf{0}_d$. Then we calculate the solution associated with the consecutive $\lambda$ on the grid, using the solution we obtain most recently as the initial for minimization.

Continuing the discussion in Sect. 2.2, the problem narrows down to find the minimum point of function (5). Now we examine the local minimum point(s) of $g(x)$ when $\beta_{(-1)} \ne \mathbf{0}_{d-1}$. Letting its derivative be zero, we have $c_3x^3 + c_2x^2 + c_1x + c_0 = 0$, where $c_3 = 2\gamma a_{11}$, $c_2 = 6\gamma a_{11}a_{12}^*$, $c_1 = t_{11} + \lambda + 4\gamma {a_{12}^*}^2 + 2\gamma a_{11}(a_{22}^* - 1)$, and $c_0 = t_{12}^* + \lambda\text{sign}(x)r^* + 2\gamma(a_{22}^* - 1)a_{12}^*$. Here $\text{sign}(x)$ is the sign of $x$. Now the problem becomes finding solutions to this cubic equation. Let $\widehat{\beta}_{(1)}$ be the updated $\beta_{(1)}$ after one coordinate descent step.

The unique non-differentiable point of $g(x)$, $x = 0$, needs special concerns. The set of all subgradients of $g(x)$ at $x = 0$ is $\{t_{12}^* + \lambda\xi r^* + 2\gamma(a_{22}^* - 1)a_{12}^* : |\xi| \le 1\}$ (for the definition of subgrandient, see e.g., Bertsekas 1995). At any $x \ne 0$, $g(x)$ is differentiable. Letting $P = c_1/c_3 - c_2^2/(3c_3^2)$ and $Q = 2c_2^3/(27c_3^3) - c_1c_2/(3c_3) + c_0/c_3$, the equation becomes $y^3 + Py + Q = 0$ with transformation $y = x + c_2/(3c_3)$.

The problem of finding the solutions to $y^3 + Py + Q = 0$ has been solved by many mathematicians. Here we follow Kavinoky and Thoo (2008). Define $\Delta = Q^2/4 + P^3/27$, $S = \sqrt[3]{-Q/2 + \sqrt{D}}$, and $T = \sqrt[3]{-Q/2 - \sqrt{D}}$. Always, there are three

roots (real or complex): $y_1 = S + T$, $y_2 = -(S+T)/2 + \sqrt{-3/4}(S-T)$, $y_3 = -(S+T)/2 - \sqrt{-3/4}(S-T)$. If $\Delta > 0$, there is one real root, if $D = 0$, there are two real roots (one of them is minimum point), and if $D < 0$, there are three real roots (two of them are local minimum points). Let $x_i = y_i - c_2/(3c_3)$, $i = 1, 2, 3$. Also let $x_{10} = \min\{x_1, x_2, x_3\}$ and $x_{20} = \max\{x_1, x_2, x_3\}$ (the middle one is a local maximum point).

*Case (i).* If $|t^*_{12} + 2\gamma(a^*_{22} - 1)a^*_{12}| \leq \lambda r^*$, zero is a local minimum point of $g(x)$ and $\widehat{\beta}_{(1)} = 0$.

*Case (ii).* If $t^*_{12} + 2\gamma(a^*_{22} - 1)a^*_{12} > \lambda r^*, \widehat{\beta}_{(1)} < 0$. If $D \geq 0$ ($D$ depends on $\text{sign}(\widehat{\beta}_{(1)}) = -1$), $x_1$ is minimum point and $\widehat{\beta}_{(1)} = x_1$. Otherwise, $\widehat{\beta}_{(1)}$ equals the negative one if $x_{10}$ and $x_{20}$ are of different signs, and equals the one of smaller absolute value if both are negative.

*Case (iii).* If $t^*_{12} + 2\gamma(a^*_{22} - 1)a^*_{12} < -\lambda r^*, \widehat{\beta}_{(1)} > 0$. If $D \geq 0$ ($D$ depends on $\text{sign}(\widehat{\beta}_{(1)}) = 1$), $x_1$ is minimum point and $\widehat{\beta}_{(1)} = x_1$. Otherwise, $\widehat{\beta}_{(1)}$ equals the positive one if $x_{10}$ and $x_{20}$ are of different signs, and equals the one of smaller absolute value if both are positive.

## References


Bertsekas D (1995) Nonlinear programming. Athena Scientific, Belmont, MA

Bickel P, Levina E (2004) Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10:989–1010

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

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Fan J, Feng Y, Tong X (2012) A road to classification in high dimensional space: the regularized optimal affine discriminant. J R Stat Soc Ser B 74:745–771

Fan J, Lv J (2008) Sure independence screening for ultrahigh-dimensional feature space. J R Stat Soc Ser B 70:849–911

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York

Jin M, Fang Y (2011) Variable selection in canonical discriminant analysis for family studies. Biometrics 67:124–132

Kavinoky R, Thoo JB (2008) The number of real roots of a cubic equation. AMATYC Rev 29:3–8

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

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

Reilly C, Miller MB, Liu Y, Oetting WS, King R, Blumenthal M (2007) Linkage analysis of a cluster-based quantitative phenotypes constructed from pulmonary function test data in 27 multigenerational families with multiple asthmatic members. Hum Hered 64:136–145

The Collaborative Study on the Genetics of Asthma (CSGA) (1997) A genome-wide search for asthma susceptibility loci in ethnically diverse populations. Nat Genet 15:389–392

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

Winawer MR (2006) Phenotype definition in epilepsy. Epilepsy Behav 8:462–476