Statistical Applications in Genetics and Molecular Biology

Volume 10, Issue 1

2011

Article 48

Adaptive Elastic-Net Sparse Principal Component Analysis for Pathway Association Testing

Xi Chen, *Vanderbilt University*

Recommended Citation:

Chen, Xi (2011) "Adaptive Elastic-Net Sparse Principal Component Analysis for Pathway Association Testing," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 48.

DOI: 10.2202/1544-6115.1697

Available at: <http://www.bepress.com/sagmb/vol10/iss1/art48>

©2011 De Gruyter. All rights reserved.

Adaptive Elastic-Net Sparse Principal Component Analysis for Pathway Association Testing

Xi Chen

Abstract

Pathway or gene set analysis has become an increasingly popular approach for analyzing high-throughput biological experiments such as microarray gene expression studies. The purpose of pathway analysis is to identify differentially expressed pathways associated with outcomes. Important challenges in pathway analysis are selecting a subset of genes contributing most to association with clinical phenotypes and conducting statistical tests of association for the pathways efficiently. We propose a two-stage analysis strategy: (1) extract latent variables representing activities within each pathway using a dimension reduction approach based on adaptive elastic-net sparse principal component analysis; (2) integrate the latent variables with the regression modeling framework to analyze studies with different types of outcomes such as binary, continuous or survival outcomes. Our proposed approach is computationally efficient. For each pathway, because the latent variables are estimated in an unsupervised fashion without using disease outcome information, in the sample label permutation testing procedure, the latent variables only need to be calculated once rather than for each permutation resample. Using both simulated and real datasets, we show our approach performed favorably when compared with five other currently available pathway testing methods.

KEYWORDS: gene expression, microarray, pathway analysis, sparse principal component analysis

Author Notes: This work was partially supported by the National Cancer Institute 5P30CA068485-13 and its Cancer Center Support Grant pilot project initiative for Xi Chen. The author would like to thank the reviewers for their valuable and constructive comments. Address for correspondence: Xi Chen (steven.chen@vanderbilt.edu), Department of Biostatistics, S-2323 Medical Center North, Nashville, TN 37232.

1. INTRODUCTION

The advance of high-throughput genomic technologies such as gene expression microarrays, DNA methylation arrays, SNP arrays, and next-generation sequencing enables researchers to make thousands to millions of measurements at the genome level simultaneously. Because of their ultra-high dimensionality and the complex gene-gene correlation patterns, these omics datasets have brought huge challenges for computation and statistical analysis. In particular, for gene expression data, the generation of a differentially expressed gene list is often just the beginning of data analysis and functional interpretation.

Recently, pathway or gene set analyses, which test disease association with pre-defined gene groups jointly instead of testing individual genes, have become popular approaches for analyzing gene expression data. These analyses take advantage of the fact that multiple genes in a functional group typically work cooperatively to fulfill biological functions and test disease associations with groups of genes defined by pathways or gene sets. For example, KEGG Pathway (Kanehisa and Goto, 2000) and Gene Ontology (GO) (Ashburner *et al.*, 2000) are the most commonly used databases for classifying genes into functional groups. This prior biological knowledge on pathways and gene sets can help researchers to obtain functional knowledge about genes.

Many analytical strategies have been proposed for pathway analysis. For example, Gene Set Enrichment Analysis (GSEA) (Mootha *et al.*, 2003; Subramanian *et al.*, 2005) is the one of the earliest approaches to tackle this problem and has been widely used by the research community. Other extended, improved, or related methods include the maxmean statistic (Efron and Tibshirani, 2007), the global test (Goeman *et al.*, 2004), Hotelling's T^2 test (Dinu *et al.*, 2007; Tsai and Chen, 2009), mixed models (Wang *et al.*, 2008, 2009) and others. Several recent reviews nicely summarized the progress in pathway analysis and compared the different analysis approaches (Goeman and Buhlmann, 2007; Nam and Kim, 2008; Ackermann and Strimmer, 2009).

Among the pathway analysis methods, performing dimension reduction on gene expression values and computing summary statistics (such as latent variables) for each pathway is an attractive approach. First, an often encountered problem is that the number of genes within a pathway is greater than the number of patients; this can be problematic for traditional statistical modeling which typically assumes that the number of variables is smaller than the sample size. Second, since we expect that the genes in a pathway share similar functions, using one or a few latent variables to represent each pathway also makes sense biologically. Third, beyond testing pathway association with clinical phenotypes, the estimated latent variables can also be used for further analysis such as the prediction of disease outcome or the analysis of pathway-pathway interactions.

In pathway analysis, the latent variables are often constructed based on principal components (PCs) (Tomfohr et al., 2005; Kong et al., 2006; Ma and Kosorok, 2009). Typically, all genes within a pathway are used to estimate the PCs. However, when the number of genes within a pathway is large, association signals from genes unrelated to the clinical outcome may introduce noise and obscure gene set association signal. To increase the power of gene set association testing and to effectively target association signals from functional genes, in this paper, we propose first to discover coherent sub-modules in each pathway by using a data-adaptive approach and then to construct latent variables for the sub-modules. More specifically, we will use the Adaptive Elastic-net Sparse Principal Component Analysis (AES-PCA) to summarize gene expressions within a pathway. The AES-PCA method removes noisy expression signals and also account for correlation structure between the genes. It is computationally efficient and the estimation of the PCs does not depend on clinical outcomes. Thus, for experiments with complex designs, sparse PCs obtained from AES-PCA can be conveniently integrated into regression models, where additional covariate variables can be simultaneously included in the model.

In section 2, we describe the details of the AES-PCA modeling framework and testing procedures for pathway analysis. In section 3, the AES-PCA is compared with five other widely used pathway analysis approaches using both simulation data and data from two real gene expression experiments. We provide some comments and conclusions in the last section.

2. METHODS

2.1 The adaptive elastic-net

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.1)$$

where $\mathbf{y}$ is a response vector of length n , $\mathbf{X}$ is an n by p matrix, and $\boldsymbol{\beta}$ is a vector of regression coefficients with length p . Assume both $\mathbf{y}$ and columns of $\mathbf{X}$ are standardized with mean 0. The lasso estimate is defined b

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^p |\beta_j|, \quad (2.2)$$

where the L_1 penalty $\lambda_1 \sum_{j=1}^p |\beta_j|$ shrinks the coefficients to zero as the tuning parameter λ_1 , which determines the path of the coefficient estimates, increases (Tibshirani, 1996).

Lasso can be easily applied to high-dimensional data, but several drawbacks may limit its applications. For example, the number of selected variables is bounded by the sample size. In addition, for a group of highly correlated variables, lasso tends to select one variable and ignores the others. This is undesirable for pathway analysis because when only one variable (i.e. gene) is selected, the gene group effect will be reduced dramatically. To address these limitations, Zou and Hastie, (2005) recently proposed the elastic net estimator. It can be represented as

$$\hat{\beta}^{\text{EN}} = (1 + \lambda_2) \left\{ \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2 \sum_{j=1}^p |\beta_j|^2 + \lambda_1 \sum_{j=1}^p |\beta_j| \right\}, \quad (2.3)$$

which combines both L_1 and L_2 penalties into the model.

Another direction of improvement is to correct bias in estimation by adding weight θ_j for each β_j , which corresponds to different amounts of shrinkage, to the regression coefficients. The weight vector θ can be defined as $\theta = 1 / |\hat{\beta}_{ols}|$, and $\hat{\beta}_{ols}$ is the coefficient estimator from ordinary least square. This is called adaptive lasso, which has the form

$$\hat{\beta}^{\text{Ada}} = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \sum_{j=1}^p \theta_j |\beta_j| \quad (2.4)$$

and enjoys the oracle properties (Zou, 2006).

Most recently, Zou and Zhang (2009) proposed the adaptive elastic-net estimator, which combines the two versions of lasso in Equations 2.4 and 2.5 into a single model, and it has the strengths of having both stable estimation and the oracle property. It is defined by

$$\hat{\beta}^{\text{AdaEN}} = (1 + \lambda_2) \left\{ \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2 \sum_{j=1}^p |\beta_j|^2 + \lambda_1 \sum_{j=1}^p \theta_j |\beta_j| \right\}. \quad (2.5)$$

and it has been shown that adaptive elastic-net outperformed lasso, elastic-net and adaptive lasso in terms prediction error and variable selection by simulation studies (Zou and Zhang, 2009).

2.2 Principal component analysis

PCA is one of the most commonly used dimensional reduction techniques for high-dimensional genomic data analysis. The PCs for a data matrix can be computed via Singular Value Decomposition (SVD). Let $\mathbf{X}$ be an $n \times p$ matrix with p variables (e.g. genes) measured on n samples. The SVD of $\mathbf{X}$, which has centered columns with mean zero, can be written as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad (2.6)$$

where the columns of $\mathbf{U}$ and $\mathbf{V}$ are PCs (eigen-arrays) and their corresponding loadings (eigen-genes) respectively (Alter et al., 2000). In Tomfohr et al. (2005), the first PC, which accounts for the largest amount of variations in pathway activity among all linear combinations of gene expressions, was used to construct the “metagene” to project the multi-dimensional expression data in a pathway onto a one-dimensional space. Then a t-test was used to compare metagenes for cases and controls for each pathway. Since the first PC may not be sufficient to capture enough information within a pathway, this approach was recently extended to testing a few more PCs using multivariate tests (Kong et al, 2006; Ma and Kosorok, 2009).

However, PCs computed based on all genes in a pathway may not reflect the biological context for a specific experiment or condition. Thus a new method that estimates sparse loadings for the PCs in order to naturally select variables (i.e. genes) relevant to a particular experimental condition would be more desirable for pathway analysis.

2.3 Sparse PCA

Principal components can be derived under regression-type framework. Zou et al. (2006) applied lasso regression optimization to achieve sparse loading of PCs, which is called the sparse PCA approach. For the first k PCs, we have

$$\begin{aligned} (\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg \min \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T \mathbf{x}_i\|^2 + \lambda_2 \sum_{j=1}^k |\beta_j|^2 + \lambda_1 \sum_{j=1}^k |\beta_j| \\ \text{subject to } \mathbf{A}^T \mathbf{A} = \mathbf{I} \end{aligned} \quad (2.7)$$

where $\mathbf{A}_{p \times k} = [\alpha_1, \dots, \alpha_k]$, $\mathbf{B}_{p \times k} = [\beta_1, \dots, \beta_k]$, p is the number of variables, and $\mathbf{x}_i$ is the vector for the i th sample. When the tuning parameter λ_1 equals zero then $\hat{\beta}_j$ is proportional to V_j (j th standard principal component), and if λ_1 is positive

then $\hat{\beta}_j$ can be normalized in order to obtain the j th sparse principal component. Sparse PCA based on different algorithms including adaptive sparse PCA were also developed recently (Shen and Huang, 2008; Leng and Wang, 2009). In this paper, we use the adaptive elastic-net to estimate the sparse loadings of the PCs for genes within a pathway since adaptive elastic-net has better theoretical properties and empirical performance than elastic-net and adaptive lasso (Zou and Zhang, 2009). Thus, the new objective function is

$$\begin{aligned} (\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg \min \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T \mathbf{x}_i\|^2 + \lambda_2 \sum_{j=1}^k |\beta_j|^2 + \lambda_1 \sum_{j=1}^k \theta_j |\beta_j| \\ \text{subject to } \mathbf{A}^T \mathbf{A} = \mathbf{I} \end{aligned} \quad (2.8)$$

where $\theta_j = 1/\tilde{\beta}_j$, and $\tilde{\beta}_j$ is the loading coefficients of the j th standard PC. Algorithm 1 in Zou et al. (2006) was used to optimize the equation (2.8). When $\mathbf{A} = [\alpha_1, \dots, \alpha_k]$ is fixed, each adaptive elastic-net estimate β_j can be solved by

$$\beta_j = \arg \min_{\beta} (\alpha_j - \beta)^T \mathbf{X}^T \mathbf{X} (\alpha_j - \beta) + \lambda_2 |\beta|^2 + \lambda_1 \theta_j |\beta|. \quad (2.9)$$

Then $\mathbf{B}$ and $\mathbf{A}$ are updated by computing $\mathbf{X}^T \mathbf{X} \mathbf{B} = \mathbf{U} \mathbf{D} \mathbf{V}^T$ and $\mathbf{A} = \mathbf{U} \mathbf{V}^T$ respectively until convergence. In equations 2.8 and 2.9, $\mathbf{X}$ represents an $N \times p_m$ matrix with p_m genes in the m th a priori known pathway, where $m = 1, \dots, M$, and M is the total number of a priori known pathways. The sparse PCA algorithm can be summarized as:

- (a) Initialize $\mathbf{B} = [\beta_1, \dots, \beta_k]$;
- (b) Iterative until convergence:
 - i. Hold $\mathbf{B}$ fixed and solve Equation 2.8 with respect to $\mathbf{A}$. The solution takes the form $\mathbf{A} = \mathbf{U} \mathbf{V}^T$, where $\mathbf{U} \mathbf{D} \mathbf{V}^T$ denotes the SVD of $\mathbf{X}^T \mathbf{X} \mathbf{B}$;
 - ii. Hold $\mathbf{A}$ fixed and solve with respect to $\mathbf{B}$. It can be shown that the solution also solves Equation 2.9, which can be solved using standard algorithms for the lasso such as LARS or coordinate descent.

In this paper, LARS was used compute the entire solution path with a fixed λ_2 (Efron et al., 2004). Although coordinate descent or other algorithms can be applied to improve the computation efficiency of the procedure, the proposed approach is not very computationally demanding since we process gene expression data by each pathway instead of by the whole transcriptome (Friedman et al., 2007). The tuning parameter λ_2 can be assigned to a small positive number

as suggested by Zou et al. (2006). We used BIC to choose tuning parameter λ_1 since BIC has been suggested as an optimal tuning parameter selector (Zou et al., 2007; Wang and Leng, 2007):

$$\text{BIC}_{\lambda_{1j}} = (\hat{\alpha}_{\lambda_{1j}} - \hat{\beta}_{\lambda_{1j}})^T \mathbf{X}(\hat{\alpha}_{\lambda_{1j}} - \hat{\beta}_{\lambda_{1j}}) + \frac{\log n}{n} df_{\lambda_{1j}} \quad (2.10)$$

where $df_{\lambda_{1j}}$ is the number of nonzero coefficients in the lasso estimator $\hat{\beta}_{\lambda_{1j}}$ for the j th sparse principal component. This method is referred to as adaptive elastic-net sparse PCA (AES-PCA).

Normally it is difficult to justify that only a few sparse PCs will explain all the transcription activities in the genome well. At the pathway level, however, using sparse PCA to discover latent variables that represent a few different transcription activities within a pathway is more biologically reasonable.

2.4 Current proposal for pathway testing

The following four steps describe the workflow of our proposed pathway association testing method:

- (1) Map gene expression data to collected pathway database such as KEGG, Biocarta, and others;
- (2) Summarize gene expression pathway activities for each sample using the first two sparse PCs, which are extracted using the AES-PCA method described in Section 2.3;
- (3) Test association of sparse PCs with clinical outcomes under the regression modeling framework. More specifically, the generalized linear model, the linear regression model, and the Cox proportional hazards model were applied for categorical, continuous, and survival outcome, respectively, and the deviance can be taken as summary statistic;
- (4) Estimate p-values via permutation test, by permuting outcome labels 10,000 times and computing the deviance statistic for each permutation sample.

3. SIMULATION STUDIES

The purpose of this simulation study is to examine the test size and power of the proposed approach under different settings of case-control study design. We used a similar simulation study design as in Liu et al. (2007). More specifically, each gene set included 100 genes and the expression values of these 100 genes were

generated from a multivariate normal distribution $MVN(\mu, \Sigma)$, where μ is the mean vector and Σ is the variance-covariance matrix. All elements of μ were generated from uniform distribution $U(0,10)$, and the diagonal elements of the covariance matrix were generated from uniform distribution $U(0.1,5)$. For the correlations between genes in the gene set, we simulated a block diagonal structure for the gene-gene covariance matrix Σ : two blocks consisting of 15 genes each were simulated. In each block, all genes had a pair-wise correlation ρ , which was set at 0.3, 0.5, 0.7, and 0.9.

The 15 genes in the first block and the 15 genes in the second block were selected to be the differentially expressed genes responsible for the differences between case-control groups, so the means for the first 15 genes and the second 15 genes in case group were set at $\mu + 2\gamma$ and $\mu - 2\gamma$ by adding and subtracting a constant γ , respectively. In our study, γ was set at 0, 0.1, 0.2, 0.3, 0.4, and 0.5. The total sample sizes for the study (both case and control groups) were set at $2n = 20, 50$ where n is the number of samples per group. There were a total of $2(n) \times 4(\rho) \times 6(\gamma) = 48$ scenarios, and 1000 replication datasets were generated for each scenario.

We compared the performance of the proposed approach with five other gene set testing methods. These methods included (1) global test (Goeman et al., 2004) with R package *globaltest*; (2) GlobalAncova (Hummel et al., 2008) with R package *GlobalAncova*; (3) PCA-based test (Kong et al., 2006) with R package *pcot2*; (4) GSEA (Subramanian et al., 2005) with GSEA R programs downloaded from the Broad Institute website (www.broadinstitute.org/gsea/downloads.jsp); (5) maxmean test (Efron and Tibshirani, 2007) with R package *GSA*. The methods (1) - (5) are well-known gene set or pathway analysis tools and are described in Appendix 1.

Figure 1 and Figure 2 display the comparison of the seven gene set testing methods for sample size $2n = 20$ and 50, respectively. Test sizes and power were compared at nominal p-value 0.05 for each method by estimating the proportion of p -values less than 0.05 for both null gene sets and causal gene sets. The null gene sets were represented by the scenarios where $\gamma = 0$. All methods preserved type I error rates at the 0.05 significance level. Because empirical p -values were computed based on sample label permutations in AES-PCA and other 5 methods, as expected, these methods had correct test sizes.

Figure 1 Power comparisons of six pathway testing methods when sample size $2n=20$ in simulation study. ρ represents the pair-wise correlation within causal gene blocks and γ represents the mean difference between case and control groups. The proposed AES-PCA, which is the red line, showed the best power for all scenarios. The other 5 pathway analysis methods are described in Appendix 1.

Figure 2 Power comparisons of six pathway testing methods when sample size $2n=50$ in simulation study. ρ represents the pair-wise correlation within causal gene blocks and γ represents the mean difference between case and control groups. The proposed AES-PCA, which is the red line, showed the best power for all scenarios. The other 5 pathway analysis methods are described in Appendix 1.

In terms of power, the AES-PCA model performed best across all γ levels when compared with the other 5 approaches. The *pcot2* test, which was based on the first two PCs of all the genes in each gene set, also showed reasonably good power, especially when the sample size or gene correlation increased. AES-PCA worked well because when estimating sparse PCs (which represented gene expression activities within gene sets), the gene selection step removed noisy signals from irrelevant genes in the gene sets. Both GSEA and GSA showed low power to identify differentially expressed gene sets using our simulation parameters.

4. APPLICATION

4.1 Lung cancer cell lines data

Paclitaxel is an extensively used anti-cancer drug. Gemma et al. (2006) tested for sensitivity of several anti-cancer drugs including paclitaxel on a set of non-small cell lung cancer (NSCLC) cell lines. This study also collected gene expression profiles of the cell lines using the Affymetrix HGU133A platform (GEO Accession No. GSE4127). In our analysis, we tested the association between gene expression levels and growth inhibitory activity (GI50) of paclitaxel in 28 NSCLC cell lines. After data preprocessing and filtering, 4234 genes were mapped to 186 KEGG pathways with minimum pathway size (total number of genes in the pathway) of 5. We applied AES-PCA, *globaltest*, *GlobalAncova*, GSEA and GSA for pathway analysis. *Pcot2* was not used since it was designed for binary outcome only. For each method, in addition to nominal p-values, we also computed adjusted p-values based on the method of Benjamini and Hochberg (1995) to control false discovery rate (FDR). FDR 0.1 was used as a significance cutoff to select significant pathways.

For GSEA and GSA, all pathways had adjusted p-values greater than 0.5. Similarly, all adjusted p-values from *GlobalAncova* were greater than 0.15, and the results from *globaltest* were very similar except one pathway: “taurine and hypotaurine metabolism” with a p-value of 0.04. There were eight genes in this pathway and the gene set significance was mostly influenced by two genes — *CSAD* and *GADI* — based on the diagnostic plot implemented in the R package *globaltest*.

In contrast, the AES-PCA method models for association of the outcome with a group of correlated genes instead of a few single genes. Three pathways were selected at FDR 0.1 level by our proposed approach. The most significant one was “DNA replication pathway”, which was directly involved in the anti-cancer mechanism since paclitaxel is a mitotic inhibitor. The other two were “selenoamino acid metabolism pathway” and “renin-angiotensin system (RAS)

pathway” with 22 and 16 genes, respectively. Antagonism of RAS can suppress tumor growth and metastasis and has been shown to be effective in several studies recently (George et al., 2010). Thus, our analysis results displayed drug sensitivity associations with several pathways, and these significant pathways can be used as possible molecular signatures to guide therapeutic options.

4.2 Colon cancer survival data

This colon cancer dataset included 55 patients who were diagnosed with colorectal adenocarcinoma from stage I to IV (Smith et al., 2010). The median follow-up time was 50.2 months and disease-free survival time and censoring status were used as survival analysis outcomes. Gene expressions of 55 patients were measured by Affymetrix HGU133 plus 2.0 GeneChip Expression Arrays (GEO Accession No. GSE17537). We had 4796 genes mapped to 186 KEGG pathways with a minimum pathway size of 5.

The purpose of this analysis is to identify pathways associated with survival times. Pathway analysis was conducted for all methods except *GlobalAncova* and *pcot2*, which are not applicable to survival outcomes. We followed the same procedures as those for our lung cancer cell lines data analysis described in Section 4.1.

All adjusted p-values from GSEA and GSA were greater than 0.7. The lowest adjusted p-value in globaltest was 0.3184 (for the “type II diabetes mellitus” pathway). AES-PCA identified 46 pathways at FDR 0.05 level. The top eight pathways and genes selected by sparse PC1 and PC2 are listed in Appendix 2. Several top pathways are directly associated with cancer development and metastasis. Much previous experimental evidence suggests that the PPAR pathway, which is the most significant pathway in our list, plays a critical role in colorectal tumor growth. The crucial gene in this pathway is peroxisome proliferator-activated receptor gamma (*PPARG*). The protein encoded by this gene belongs to the peroxisome proliferators-activated receptor subfamily of nuclear receptors and directly activates a wide range of genes leading to acceleration of colorectal tumor growth (Wu, 2000; Gupta et al., 2001; Gupta et al., 2004). *PPARG* is regulated by Wnt/ β -catenin signaling pathway, which is also important for development of colorectal cancer (Jansson et al., 2005). Another characteristic from the top pathways in Appendix 2 is that epidermal growth factor receptor (*EGFR*) related genes including *ERBB2*, *KRAS*, *PIK3* family and *MAP2K* family were frequently selected by AES-PCA. Anti-EGFR monoclonal antibodies such as cetuximab and panitumumab can block *EGFR* tyrosine kinase activation and then inhibit the activation of RAS-MAP2K and PI3K-AKT signaling pathways to prevent cancer cell proliferation and invasion (Ciardiello and Tortora, 2008). Mutation status of *KRAS* is also a predictor of resistance to

EGFR monoclonal antibodies in colorectal cancer (Lièvre et al., 2006). These *EGFR* related genes were likely implicated in targeted therapies in metastasis of colon cancer of 55 patients. In Figure 3, we display the first and second AES-PCs loading for the selected genes in PPAR pathway. For the 68 genes in this pathway, there were 13 genes with nonzero loadings for the first PC and 9 for the second, and none of those genes overlapped between these two PCs, which may represent different biological processes within this pathway.

Figure 3 Gene plots of first and second adaptive elastic-net sparse PCs for the PPAR pathway.

5. DISCUSSION

In the analysis of gene pathways, it is often advantageous to discover the subgroups or sub-modules within each pathway before testing pathway association with clinical outcomes because (i) pathway membership information is often based on previous literature and may not be very accurate; (ii) it is biologically meaningful to find multiple groups of genes in a pathway with distinct functions, and gene selection can help researchers focus on a few key players in a large pathway; and (iii) detecting pathway association signals with clinical outcomes by removing noisy signals from irrelevant genes can often improve statistical power.

Principal component analysis is often used as a dimension reduction approach to summarize variations from multiple genes within a pathway while

modeling correlations among genes. Constructing PCA scores and testing their association with clinical outcomes after supervised variable (gene) selection was shown to be an effective approach for pathway testing (Chen et al., 2008; Chen et al., 2010). However, correcting for the bias introduced by using outcome information in the gene selection step is not trivial. A permutation test can be used for testing pathway association with clinical outcomes, but because supervised variable selection needs to be performed for each permutation re-sample, the computation can be challenging.

In this paper, we developed a different strategy to do variable (gene) selection and PCA-based dimension reduction for pathway testing. Our approach based on adaptive elastic-net sparse PCA (AES-PCA) is attractive because it can efficiently perform gene selection by capturing the variations and incorporating the correlations of genes in a pathway. In addition, because AES-PCA is an unsupervised procedure that does not use any outcome information in the variable selection step, it is also straightforward to implement a permutation test based on AES-PCA scores.

In the simulation studies, we compared the proposed AES-PCA with five other popular and related pathway testing approaches. Our results showed that AES-PCA has the best power across all scenarios. In particular, in comparison with the PCA-based method from Kong et al. (2006), the better performance of AES-PCA suggests that instead of computing PCA scores based on all genes in a pathway, the variable selection step in AES-PCA helped it gain power by removing noise gene signals from irrelevant genes.

The default setting in the AES-PCA method is to test pathway association based on the first and second PCs, which is the same as the *pcot2* approach described in Kong et al. (2006). Although adding more PCs may increase the power of the test, in our simulation study and real datasets analysis, the two AES-PCs implementation showed a satisfactory performance. In addition, the few extracted PCs are relatively easy to interpret biological activities within a pathway.

Another advantage of the proposed AES-PCA pathway testing approach is that it is applicable to different experimental designs. In this paper, we illustrated the analysis for binary, continuous and survival outcomes. For studies with larger sample sizes, future studies can be conducted to fit extracted sparse PCs or latent variables in more complicated models to account for correlations between pathways or to detect pathway-pathway interactions. It would also be interesting to compare the sub-modules in pathways identified by our approach with gene modules identified by gene network methods.

APPENDIX

Appendix 1. Description of pathway testing methods used in simulation studies

We describe the five available pathway analysis methods used in the simulation study. For clarity, we followed the notations in each method's original paper.

GSEA

Gene set enrichment analysis (GSEA) is one of the earliest developed and a widely used pathway analysis tool (Subramanian et al. 2005). Let N be the total number of genes in a gene expression experiment; k be the number of samples; and S be the pathway or gene set of interest. To calculate enrichment score $ES(S)$: (1) for each gene, compute correlation r_j between gene expression and phenotype, then rank N genes by r_j to have a ranking list $L = \{g_1, \dots, g_N\}$; (2) for each position i in the list L , two running sums for genes within gene set S and outside S can be calculated separately according to the following formula

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_R}, \text{ where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - N_H)}$$

The enrichment score $ES(S)$ is the maximum of $P_{\text{hit}} - P_{\text{miss}}$ over all position i . When $p=0$, $ES(S)$ is the standard Kolmogorov-Smirnov statistic. The null hypothesis for GSEA of gene set S is: genes in S are randomly distributed in gene list L . The empirical p-value for $ES(S)$ can be computed by permutation of sample labels.

GSA

Efron and Tibshirani (2007) proposed the use of the maxmean statistic as well as combining sample and gene permutations to improve GSEA, especially for the Kolmogorov-Smirnov statistic. The steps are as follows:

1. Compute a summary statistic z_j for each gene, for example two-sample t -statistic;
2. For each gene set S with m genes, define the summary statistic maxmean $S_{\text{max}} = \max\{\bar{s}^{(+)}, \bar{s}^{(-)}\}$

where $s(z) = (s^{(+)}, s^{(-)}), s^{(+)} = \max(z, 0), s^{(-)} = -\min(z, 0)$ and $\bar{s}^{(+)} = \frac{1}{m} \sum s_j^{(+)}, \bar{s}^{(-)} = \frac{1}{m} \sum s_j^{(-)}$;

3. Standardize $S_{\max}$ by its randomization mean and standard deviation: $S'_{\max} = (S_{\max} - \text{mean}_s) / \text{stdev}_s$ where mean_s and stdev_s are the mean and standard deviation of $S_{\max}$ by gene label permutation, respectively;
4. Permute sample labels B times and recompute $S'_{\max}$ each time, then estimate p -values and false discovery rates for all pathways.

Globaltest

Goeman et al. (2004) proposed a generalized linear model for testing the association of phenotype with all genes in a pathway or gene set. Let Y be the $n \times 1$ vector for clinical outcome, for example 0/1 for two-group comparison, and let $X_{n \times m}$ be the gene expression matrix for m genes in a pathway S . The regression model can be written as

$$E(Y_i | \beta) = h^{-1} \left(\alpha + \sum_{j=1}^m x_{ij} \beta_j \right)$$

where β_j is the regression coefficient for each gene. The null hypothesis for testing pathway association with clinical outcome is

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0$$

which states that none of the genes in the pathway is associated with the outcome. To simplify the testing procedure, the authors assume all β are from a common distribution with expectation 0 and variance τ^2 , thus the null hypothesis becomes

$$H_0 : \tau^2 = 0.$$

A score test is implemented to test the null hypothesis by assuming a diagonal covariance matrix $\tau^2 I_m$ for vector β (independent genes assumption). Either asymptotic approximation or permutation test are available in R package *globaltest* to estimate p -values for pathways.

GlobalAncova

Mansmann and Meister (2005) proposed a pathway testing approach based on the ANCOVA model. In a two-group comparison scenario, let x_{ijk} denote the gene expression value for the i th group ($i = 1, 2$), j th gene ($j = 1, 2, \dots, N_{\text{genes}}$) and k th sample ($k = 1, 2, \dots, M$). The linear model can be written as

$$x_{ijk} = \mu_{ij} + \theta_j z_{ik} + e_{ijk}$$

where z_{ik} is the clinical covariate and $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ with group effect α , gene effect β and interaction effect γ . The null hypothesis is

$$\mu_{1j} = \mu_{2j} \Leftrightarrow \alpha_i = \gamma_{ij} = 0.$$

Either asymptotic or permutation tests of the F-statistic are available in R package *GlobalAncovat* to estimate p -values for pathways.

PCOT2

Kong et al. (2006) applied dimension reduction by PCA and Hotelling's T^2 test for pathway testing of binary clinical outcomes. Specifically, singular value decomposition was used in each pathway to extract the two-dimensional vector u_{ij} representing the first and second principal components for the j th sample ($j = 1, \dots, n_i$) in the i th group ($i = 1, 2$). The test statistic is based on Hotelling's T^2 :

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{u}_1 - \bar{u}_2)' S^{-1} (\bar{u}_1 - \bar{u}_2)$$

where $\bar{u}_i$ is the mean vector and S is the pooled empirical covariance matrix. The null hypothesis is

$$H_0 : \bar{u}_1 = \bar{u}_2$$

and the test statistic would follow an F -statistic. A permutation test is the default setting for R package *pcot2* to calculate p -values.

Appendix 2. Top eight pathways from the AES-PCA method for colon cancer data

KEGG pathways	Size	FDR	Selected Genes
PPAR signaling pathway	68	0.0001	PC1: <i>ACOX1, ACOX2, CPT2, EHHADH, PPARG, ACADM, SLC27A2, ACAA1, UPC1, PCK2, DBI, HMGCS2, SCP2</i> ; PC2: <i>ILK, CD36, UBC, ACSL1, ANGPTL4, LPL, OLR1, FABP4, FABP7</i>
ERBB signaling pathway	86	0.0093	PC1: <i>PAK3, MAP2K7, NRG2, CRKL, MAPK8, CAMK2A, MAP2k1</i> ; PC2: <i>STAT5B, PAK2, SHC1, PIK3CB, ERBB2, KRAS, MAP2K2, NRAS, GSK3B, ARAF, ABL1</i>
Aldosterone-regulated sodium reabsorption	42	0.0093	PC1: <i>PIK3CA, PIK3R5, PIK3CG, SGK1, ATP1A3, IGF1, PRKCB, HSD11B1</i> ; PC2: <i>KCNJ1, PDPK1, KRAS, INS, INSR, PIK3R2, IRS2</i>
Thyroid cancer	29	0.0093	PC1: <i>CDH1, TCF7L2, MAP2K1, RXRB, TP53, MAPK1, NRAS</i> ; PC2: <i>TFG, KRAS, CCDC6</i>
Regulation of actin cytoskeleton	208	0.0112	PC1: <i>PIK3CG, PDGFRA, CD14, SSH1, MRAS, ITGB2, ARPC5, ITGB1, ITGA4, PIP4K2A, MYLK, PAK4, MSN, CFL2</i> ; PC2: <i>ITGA2B, KRAS, RAC1, FGF1, FGF14, MYL10, FGF21, PPP1CC, ARPC3, PAK3, CRKL, CHRM2, ARPC5L</i>
Neurotrophin signaling pathway	123	0.0124	PC1: <i>RAPGEF1, RELA, MAPK1, MAPKAPK2, PTPN11, RHOA, SHC1, TP53, YWHAQ, GSK3B, MAPK14, CRK, CALM1</i> ; PC2: <i>MAPK8, KRAS, NTF4, NTRK3, KIDINS220, CRKL, MAP3K1, MAPK12, YWHAQ</i>
Endometrial cancer	52	0.0165	PC1: <i>TP53, CTNNA1, MAPK1, CDH1, SOS2, MAPK2k2, GSK3B</i> ; PC2: <i>PEN, PIK3Cb, PIK3CD, ERBB2, ELK1, TCF7L2, KRAS, SOS1, PIK3R5, APC, MAP2K1, ARAF</i>
Endocytosis	179	0.0165	PC1: <i>CHMP4C, CHMP4A, VPS4A, STAM, EGFR, CHMP2A, CHMP2B, SRC, HSPA1L, ADRB3, EPS15, CLTA, LDLR, TSG101, CXCR2, ZFYVE20, RAB4A, RAB11A, STAMBP, VPS25, RNF41</i> ; PC2: <i>HLAC, HLAG, ARRB2, ARF6, AP2M1, AGAP2, PSD, EHD1, AP2A1</i>

REFERENCE

- Ackermann M, Strimmer K. 2009. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10:47.
- Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *P Natl Acad Sci USA* 97:10101-10106.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25-29.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B* 57:289-300.
- Chen X, Wang L, Smith JD, Zhang B. 2008. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics* 24:2474-2481.
- Chen X, Wang L, Hu B, Guo M, Barnard J, Zhu X. 2010. Pathway-based analysis for genome-wide association studies using supervised principal components. *Genet Epidemiol* 34:716-724.
- Ciardiello F, Tortora G. 2008. Drug therapy: EGFR antagonists in cancer treatment. *New Engl J Med* 358:1160-1174.
- Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y. 2007. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* 8:242.
- Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Ann Stat* 32:407-451.
- Efron B, Tibshirani R. 2007. On testing the significance of sets of genes. *Ann Appl Stat* 1:107-129.
- Friedman J, Hastie T, Hoefling H, Tibshirani R. 2007. Pathwise coordinate optimization. *Ann Appl Stat* 1:302-332.
- Gemma A, Li C, Sugiyama Y, Matsuda K, Seike Y, Kosaihiira S, Minegishi Y, Noro R, Nara M, Seike M, Yoshimura A, Shionoya A, Kawakami A, Ogawa N, Uesaka H, Kudoh S. 2006. Anticancer drug clustering in lung cancer based on gene expression profiles and sensitivity database. *BMC Cancer* 6:174.

- Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20:93-99.
- Goeman JJ, Buhlmann P. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23:980-987.
- George AJ, Thomas WG, Hannan RD. 2010. The renin-angiotensin system and cancer: old dog, new tricks. *Nat Rev Cancer* 10:745-759.
- Gupta RA, Brockman JA, Sarraf P, Willson TM, DuBois RN. 2001. Target genes of peroxisome proliferator-activated receptor gamma in colorectal cancer cells. *J Biol Chem* 276:29681-29687.
- Gupta RA, Wang DZ, Katkuri S, Wang HB, Dey SK, DuBois RN. 2004. Activation of nuclear hormone receptor peroxisome proliferator-activated receptor-delta accelerates intestinal adenoma growth. *Nat Med* 10:245-247.
- Hummel M, Meister R, Mansmann U. 2008. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* 24:78-85.
- Jansson EA, Are A, Greicius G, Kuo IC, Kelly D, Arulampalam V, Pettersson S. 2005. The Wnt/beta-catenin signaling pathway targets PPAR gamma activity in colon cancer cells. *P Natl Acad Sci USA* 102:1460-1465.
- Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27-30.
- Kong SW, Pu WT, Park PJ. 2006. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 22:2373-2380.
- Leng C, Wang H. 2009. On General Adaptive Sparse Principal Component Analysis. *J Comput Graph Stat* 18:201-215.
- Lièvre A, Bachet JB, Le Corre D, Boige V, Landi B, Emile JF, Cote JF, Tomasic G, Penna C, Ducreux M, Rougier P, Penault-Llorca F, Laurent-Puig P. 2006. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res* 66:3992-3995.
- Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. 2007. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics* 8:431
- Ma S, Kosorok MR. 2009. Identification of differential gene pathways with principal component analysis. *Bioinformatics* 25:882-889.

- Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. 2003. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34:267-273.
- Nam D, Kim SY. 2008. Gene-set approach for expression pattern analysis. *Brief Bioinform* 9:189-197.
- Shen HP, Huang JHZ. 2008. Sparse principal component analysis via regularized low rank matrix approximation. *J Multivariate Anal* 99:1015-1034.
- Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang AX, Lu PC, Johnson JC, Schmidt C, Bailey CE, Eschrich S, Kis C, Levy S, Washington MK, Heslin MJ, Coffey RJ, Yeatman TJ, Shyr Y, Beauchamp RD. 2010. Experimentally Derived Metastasis Gene Expression Profile Predicts Recurrence and Death in Patients With Colon Cancer. *Gastroenterology* 138:958-968.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P Natl Acad Sci USA* 102:15545-15550.
- Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B* 58:267-288.
- Tomfohr J, Lu J, Kepler TB. 2005. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6:225.
- Tsai CA, Chen JJ. 2009. Multivariate analysis of variance test for gene set analysis. *Bioinformatics* 25:897-903.
- Wang H, Leng C. 2007. Unified LASSO estimation by least squares approximation. *J Am Stat Assoc* 102:1039-1048.
- Wang L, Zhang B, Wolfinger RD, Chen X. 2008. An integrated approach for the analysis of biological pathways using mixed models. *PLoS Genet* 4:e1000115.
- Wang L, Chen X, Wolfinger RD, Franklin JL, Coffey RJ, Zhang B. 2009. A unified mixed effects model for gene set analysis of time course microarray experiments. *Stat Appl Genet Mol Biol* 8:47.
- Wu GD. 2000. A nuclear receptor to prevent colon cancer. *New Engl J Med* 342:651-653.

- Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 67:301-320.
- Zou H. 2006a. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 101:1418:1429.
- Zou H, Hastie T, Tibshirani R. 2006b. Sparse principal component analysis. *J Comput Graph Stat* 15:265-286.
- Zou H, Hastie T, Tibshirani R. 2007. On the "degrees of freedom" of the lasso. *Ann Stat* 35:2173-2192.
- Zou H, Zhang HH. 2009. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat* 37:1733-1751.