UC Davis

UC Davis Previously Published Works

Title

Covariance-based analyses of biological pathways.

Permalink

https://escholarship.org/uc/item/54d4d0x3

Journal

Biometrika, 102(3)

ISSN

0006-3444

Authors

Danaher, P
Paul, D
Wang, P

Publication Date

2015-09-01

DOI

10.1093/biomet/asv013

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Biometrika. Author manuscript; available in PMC 2015 September 24.

Published in final edited form as:

Biometrika. 2015 September 1; 102(3): 533–544. doi:10.1093/biomet/asv013.

Covariance-based analyses of biological pathways

P. DANAHER,

NanoString Technologies, 530 Fairview Ave. N, Seattle, Washington 98109, U.S.A

D. PAUL, and

Department of Statistics, University of California, One Shields Avenue, Davis, California 95616, U.S.A

P. WANG

Icahn Institute of Genomics and Multiscale Biology, Icahn Medical School at Mount Sinai, 1470 Madison Avenue, S8-102 New York, New York, 10029, U.S.A

P. DANAHER: pdanaher@nanostring.com; D. PAUL: debpaul@ucdavis.edu; P. WANG: pei.wang@mssm.edu

Summary

The use of high-throughput data to study the changing behavior of biological pathways has focused mainly on examining the changes in the means of pathway genes. In this paper, we propose instead to test for changes in the co-regulated and unregulated variability of pathway genes. We assume that the eigenvalues of previously defined pathways capture biologically relevant quantities, and we develop a test for biologically meaningful changes in the eigenvalues between classes. This test reflects important and often ignored aspects of pathway behavior and provides a useful complement to traditional pathway analyses.

Keywords

gene expression; pathway analysis; spiked eigenvalue

1. Introduction

A pathway refers to a set of genes or proteins jointly participating in a biological process. It is of great interest to study the behavior of pathways using high-throughput-omics data. By treating a pre-defined set of genes with shared biological function as an analytical unit, pathway-level analyses efficiently exploit prior biological knowledge, improve interpretability, and enjoy greater power by combining the signals of individual genes. Existing pathway analysis methods have focused almost exclusively on the marginal behavior of pathway genes. For example, Tomfohr et al. (2005), Lee et al. (2008) and Drier et al. (2013) suggested synthesizing the information in pathway genes into measures of pathway activity, while there is a large body of work, including Subramanian et al. (2005)

Supplementary material

Supplementary material available at *Biometrika* online provides a description of a normalization scheme, outlines of proofs of Theorems 1 and 2, a derivation of our test in the setting without spiked eigenvalues, simulations investigating the consequences of departures from our assumptions, and a table containing the full results of the breast cancer expression data analysis.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and Efron & Tibshirani (2007), aimed at identifying pathways that are enriched with differentially expressed genes. These marginal analyses, while shedding light on important questions, fail to capture the full complexity of pathway behavior.

In this paper, we propose a test to examine the joint behavior of pathway genes. This test complements the marginal analyses mentioned above and helps to provide a more comprehensive understanding of biological pathways. Specifically, we consider the problem of detecting differences in covariance among a pre-defined set of pathway genes between two classes of samples, for example between two different cancer subtypes. Tests of equality of two covariance matrices are well studied. However, the number of genes in a pathway ranges from tens to hundreds, and quite often exceeds the sample size. Under such regimes, classical tests for equality of covariance matrices no longer apply. A number of authors (Schott, 2007; Srivastava & Yanagihara, 2010; Li & Chen, 2012) have developed tests for equality of covariance matrices in the high dimension, low sample size setting. However, by testing the null hypothesis that two covariance matrices are exactly equal, without accounting for the structure induced by pathway activity, these tests provide inadequate biological insight: their rejection of the null hypothesis allows no conclusions about how pathway gene behavior differs between classes. In contrast, the proposed test is motivated from a biological model of the expression of pathway genes and focused on quantities with natural biological interpretations. The novelty of this test lies in its focus on the joint rather than the marginal behavior of pathway genes and in its consideration of disordered variability orthogonal to the effects of pathway activity.

Our biological model assumes that genes' associations with pathway activity drive the leading eigenvector of their covariance matrix. This model suggests that the first eigenvector is invariant to changes in biological conditions, while the leading and remaining eigenvalues will vary across data sets in response to within-population variability of pathway activity and variability due to other, unregulated causes, respectively. Under this model, the covariance matrix of the expression levels of pathway genes has a spiked eigen-structure (Johnstone, 2001; Baik & Silverstein, 2006; Paul, 2007), and the leading eigenvalue and the trace of the covariance matrix provide a parsimonious and biologically relevant summary of pathway genes' joint behavior. Baik & Silverstein (2006) showed that if the dimension-to-sample size ratio converges to a nonzero finite constant, and if the true spiked eigenvalues exceed a threshold, the corresponding sample eigenvalues converge with probability one to limits that depend on the true eigenvalues and the dimension-to-sample size ratio. Paul (2007) proved asymptotic normality of the leading sample eigenvalues under the same framework. We extend the latter asymptotic results to design a χ^2 test statistic based on the joint behavior of the leading eigenvalue and the trace of the sample covariance matrices of the two classes. When the proposed test rejects the null, it indicates that specific, biologically-relevant quantities differ between classes. Simulations suggest that if the spiked covariance structure holds even approximately, the proposed test has better power to detect differences in biologically important functions of the eigenvalues than existing tests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2. A model of co-expression in biological pathways

First, consider data from only one class. Denote the gene expression data for a previously defined pathway with p genes from n observations by the $n \times p$ matrix Y, and denote the data vector of the i^{th} observation by $y_i = (y_{i1},\ldots, y_{ip})$. We assume that pathway activity is the primary driver of pathway gene expression. For example, for a set of genes regulated by a common transcription factor, the primary source of variance in the expression levels of the entire gene set would be changes in the activity level of the transcription factor, and it would be reasonable to specify these relationships through a linear dependence model. We write

$$y_{ik}=\mu_k+h_k a_i+\varepsilon_{ik},\ \ i=1,\ldots,n,\ \ k=1,\ldots,p, \quad (1)$$

where μ_k is an intercept specific to gene k that can be ignored for our purposes, a_i is a random variable with mean 0 and variance σ_a^2, $h = (h_1,\ldots, h_p)$ are gene specific scaling coefficients, $\varepsilon_i = (\varepsilon_{i,1},\ldots, \varepsilon_{i,p})$ are independent random variables with mean 0 and variance σ_ε^2, and $\varepsilon_1,\ldots, \varepsilon_n, a_1,\ldots, a_n$ are mutually independent. We add the constraint $\|h\|_2^2=1$ to make h and σ_a^2 identifiable. In this model, a_i reflects the level of pathway activity, e.g. the transcription factor level, in the i^{th} sample, σ_a^2 drives the well-ordered, co-regulated component of total gene variance, and σ_ε^2 measures the unordered, noisy component of pathway gene variance. It follows that

$$\mathrm{cov}(y_i)=\mathrm{cov}(a_i h+\varepsilon_i)=\sigma_a^2 hh^T+\sigma_\varepsilon^2 I. \quad (2)$$

The first eigenvalue of cov(y_i) is $\sigma_a^2+\sigma_\varepsilon^2$, and the remaining eigenvalues equal σ_ε^2. This observation allows biological interpretations to be assigned to the eigenvalues of the pathway's covariance matrix: the first eigenvalue measures the variability in pathway genes due to changing levels of pathway activity, i.e., the well-ordered, co-regulated component of pathway gene variability, while the sum of the remaining eigenvalues captures the unordered, chaotic component of variability. This interpretation echoes Tomfohr et al. (2005), Bild et al. (2005), Bair et al. (2006) and Chen et al. (2008) in implying that an observation's projection onto the first eigenvector of the pathway's covariance matrix measures the observation's pathway activity level. Moreover, the covariance structure in (2) matches the spiked eigenvalue model of Baik & Silverstein (2006), Paul (2007), Nadler (2008), Onatski (2012) and others. This implication of the model holds nearly universally in pathway data. Thus, one can take advantage of the asymptotic theories under the spiked eigenvalue model to perform inference on σ_a^2 and σ_ε^2.

In data sets with two classes of samples, we may wish to compare these biologically meaningful quantities between classes. We therefore propose to test the null hypothesis

$$H_0:(\alpha_{1,1}, T_1)=(\alpha_{2,1}, T_2), \quad (3)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $a_{j,k}$ denotes the k^{th} population eigenvalue of class j and T_j denotes the trace of the covariance matrix of class j. The interpretation of the pair $(a_{j,1}, T_j)$ gives the alternatives to H_0 specific and useful biological meaning. For example, when $a_{1,1} > a_{2,1}$ and $T_1 - a_{1,1} \approx T_2 - a_{2,1}$, we might conclude that the pathway activity level is stronger in the first class, or, when $T_1 - a_{1,1} > T_2 - a_{2,1}$, we might conclude that the pathway is dysregulated and subject to greater nloise in the first class. By directly testing H_0 rather than the stronger null hypothesis of equality of the entire covariance matrices, we maximize power to detect changes in the modes of variability attributable to pathway activity and to noisy, non-co-regulated causes, while detecting other changes in the covariance matrix only insofar as they change our eigenvalue statistics. In particular, we ignore the eigenvectors and redistribution of weights among smaller eigenvalues. The biological model specified in (1) and (2) can also be seen as a factor analysis model with one factor and a specialized covariance structure for the idiosyncratic term ε_{ik}, even though our hypotheses and test procedure differ from the commonly used tests for factorial invariance (Meade & Bauer, 2007).

Model (1) suggests two general features of pathway data: (a) the eigenvalues of the covariance matrix resemble the spiked model; (b) the leading eigenvectors capture the effects of pathway activities on gene expression, or equivalently, the leading spiked eigenvalues capture variability in the data due to changes in pathway activity. Both features apply for a large set of pathways even when model (1) does not hold. The statistical theory behind our test relies on (a), and the biological interpretation of our test is based on (b). The Supplementary Material contains extensive empirical investigations supporting (a) and (b).

In the next sections, we introduce a test for H_0 in (3), assuming (a) and (b) hold. Moreover, in many cases, pathway genes are subject to multiple biological processes. When this occurs the covariance matrix has additional spikes, i.e., more eigenvalues become significantly larger than the noise eigenvalues. The proposed test also accommodates these scenarios.

3. A test for differences in the eigenstructure of Σ_1 and Σ_2

3·1. The single spiked eigenvalue setting

Denote the eigenvalues and trace of the sample covariance matrices by $\hat{a_{j,i}}$ and $\hat{T_j}$ (j = 1, 2; i =1,…,p). To test H_0 in (3), a natural choice is to form a test statistic using $\hat{a_{1,1}} - \hat{a_{2,1}}$ and $\hat{T_1} - \hat{T_2}$. We use a quadratic form to combine the information in these quantities.

Under H_0 in (3), without loss of generality, we assume that $\sigma_{\varepsilon,1} = \sigma_{\varepsilon,2} = 1$, or equivalently, the unspiked eigenvalues of the common covariance matrix are all equal to 1, $a_{j,2} = \cdots = a_{j,p} = 1$, ($j$ = 1, 2). To adhere to this assumption, we normalize the data as follows. We calculate a scale factor equal to the square root of the median eigenvalue of the pooled sample covariance matrix from both classes and divide all the observations by this factor; see the Supplementary Material.

For notational convenience, in the rest of this subsection we use a_j and $\hat{a_j}$ to mean $a_{j,1}$ and $\hat{a_{j,1}}$, respectively. According to Baik & Silverstein (2006), the first sample eigenvalue is a biased estimate of its population counterpart: $\hat{a_j} \to a_j + \gamma_j a_j/(a_j - 1)$, where $p, n \to \infty$, $p/n_j \to \gamma_j \in (0, \infty)$ and $\alpha_j > 1 + \gamma_j^{1/2}$, ($j$ = 1, 2). Define $b_a = (\gamma_1 - \gamma_2)a_0/(a_0 - 1)$, where a_0 is the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

first eigenvalue shared by both classes under H_0 and satisfies $\alpha_0 > 1 + \max(\gamma_1^{1/2}, \gamma_2^{1/2})$. Then under H_0, $(\hat{a_1} - \hat{a_2}) \to b_a$ almost surely. This limiting value $b_a \neq 0$ when $\gamma_1 \neq \gamma_2$. To test H_0, we focus on the bias-corrected quantity $(\hat{a_{1,1}} - \hat{a_{2,1}} - \hat{b_a})$ and propose the test statistic $Q^T \hat{\sum}_Q^{-1} Q$, where

$$Q = (Q_\alpha, Q_T)^T = \left(\hat{\alpha}_1 - \hat{\alpha}_2 - \hat{b}_\alpha, \hat{T}_1 - \hat{T}_2\right)^T, \quad (4)$$

and $\hat{b_a}$ and $\hat{\Sigma_Q}$ are appropriate consistent estimates for b_a and $\Sigma_Q = \mathrm{cov}(Q)$, respectively.

We now describe the construction of $\hat{b_a}$. We first propose the following estimator for a_0,

$$\bar{\alpha}_0 = (w_1 \bar{\alpha}_1 + w_2 \bar{\alpha}_2),\ w_j = n_j/(n_1 + n_2), \quad (5)$$

where $\bar{a}_j$ is the asymptotic method of moments estimator for a_j, namely, $\bar{a}_j = [1 + \hat{a_j} - \hat{\gamma_j} + \{(1 + \hat{a_j} - \hat{\gamma_j})^2 - 4\hat{a_j}\}^{1/2}]/2$, which is obtained by solving the equation $\hat{a_j} = a_j\{1 + \hat{\gamma_j}/(a_j - 1)\}$. Here and henceforth, $\hat{\gamma_j} = p/n_j$ for $j = 1, 2$. Substituting a_0 with $\bar{a}_0$ in the expression for b_a yields the estimate

$$\hat{b}_\alpha = (\hat{\gamma}_1 - \hat{\gamma}_2)\bar{\alpha}_0/(\bar{\alpha}_0 - 1). \quad (6)$$

If $\hat{\alpha}_j < (1 + \hat{\gamma}_j^{1/2})^2$, then $\bar{a}_j$ is complex-valued, which indicates that the population covariance is either unspiked or has small, undetectable, spikes.

Define the 2 × 2 symmetric matrix Σ_Q with diagonal elements $\tau_{Q\alpha\alpha}$ and τ_{QTT}, respectively, and off-diagonal element $\tau_{Q\alpha T}$. Theorem 2 yields the consistent estimates

$$\hat{\tau}_{Q_{\alpha\alpha}} = \sum_{j=1}^{2} \left(\frac{2}{n_j}\right) \frac{\bar{\alpha}_0 \theta_j^2 \rho_j}{1 + \bar{\alpha}_0 \hat{\gamma}_j / \{(\bar{\alpha}_0 - 1)^2 - \hat{\gamma}_j\}} \quad (7)$$

$$\text{and} \quad \hat{\tau}_{Q_{\alpha T}} = \sum_{j=1}^{2} \left(\frac{2}{n_j}\right) \frac{\bar{\alpha}_0 \theta_j \rho_j}{1 + \bar{\alpha}_0 \hat{\gamma}_j / \{(\bar{\alpha}_0 - 1)^2 - \hat{\gamma}_j\}}. \quad (8)$$

We estimate $\hat{\tau_{QTT}}$ using equation (10).

After we obtain $\hat{\Sigma_Q}$, we propose to reject H_0 for large values of $Q^T \hat{\sum}_Q^{-1} Q$. According to Theorem 2, under H_0, the asymptotic joint normality of $\hat{a_1} - \hat{a_2} - \hat{b_a}$ and $\hat{T_1} - \hat{T_2}$ suggests that $Q^T \sum_Q^{-1} Q \to \chi_2^2$ in distribution. Then to the extent that $\sum_Q^{-1}$ is estimated accurately, our test statistic $Q^T \hat{\sum}_Q^{-1} Q$ may be compared to the quantiles of a χ_2^2 distribution to obtain a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

p-value. A permutation test may also be employed. Simulations in Section 5 show the proposed test to have accurate Type-1 error at all sample sizes when our assumptions hold, suggesting that accurate estimation of $\sum_Q^{-1}$ is not a hurdle for the test's performance.

3·2. Test robust to the number of spiked eigenvalues

We generally expect that genes in a pathway are jointly associated with not just one but a number of biological processes, which implies the existence of multiple spiked eigenvalues. To accommodate an unspecified number of spiked eigenvalues in the proposed test, we first estimate the number of spiked eigenvalues and then apply a modified expression for var($\hat{T}_j$).

To estimate M_j, the number of spiked eigenvalues in class j, we choose a threshold of $(1+\hat{\gamma}_j^{1/2})^2+\{2\log(n_j)/n_j\}^{1/2}$, and with I denoting the indicator function, define

$$\hat{M}_j=\sum_{m=1}^{p} \mathrm{I}\left[\hat{\alpha}_{j,m}>(1+\hat{\gamma}_j^{1/2})^2+\{2\log(n_j)/n_j\}^{1/2}\right]. \quad (9)$$

This estimator may have difficulty classifying the eigenvalues near $(1+\hat{\gamma}_j^{1/2})^2$. However, the treatment of such small spiked eigenvalues will not appreciably affect our estimates of var($\hat{T}_j$). We then use independence of $\hat{T}_1$ and $\hat{T}_2$ to estimate τ_{QTT} with

$$\hat{\tau}_{Q_{TT}}=\sum_{j=1}^{2}\frac{2}{n_j}\left(\sum_{m=1}^{\hat{M}_j}\hat{\alpha}_{j,m}^2+p-\hat{M}_j\right). \quad (10)$$

Some alternative methods for estimating the number of spikes, e.g. the proposal by Kritchman & Nadler (2008), have good power of detection and could be used instead of the estimator (9), but the approach detailed above does not depend on the Gaussian assumption.

We outline below the proposed procedure for testing H_0 in (3), which is robust to the number of spiked eigenvalues.

1. Calculate the eigenvalues $\{\hat{a}_{j,k}\}$ and trace $\hat{T}_j$ of the sample covariance matrix $\hat{\Sigma}_j$ (j = 1, 2).
2. Calculate $\hat{b}_a = (\hat{\gamma}_1 - \hat{\gamma}_2)\bar{a}_0/(\bar{a}_0 - 1)$, where $\bar{a}_0$ is defined in equation (5).
3. Calculate Q according to (4).
4. Estimate Σ_Q:
   a. Estimate the number of spiked eigenvalues in each class according to (9); and then calculate $\hat{\tau}_{Q_{TT}}$ according to (10).
   b. Calculate θ_j and ρ_j, j = 1, 2, as defined by Theorem 2. Compute $\hat{\tau}_{Q_{aa}}$; and $\hat{\tau}_{Q_{aT}}$ according to equation (7) and (8) respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Compute the test statistic $Q^T\hat{\sum}_Q^{-1}Q$. To attain a p-value, compare its value to the quantiles of a χ^2_2 distribution. Alternatively, permute the class labels and recompute the test statistic many times, and compare the quantiles of the resulting statistics to the true $Q^T\hat{\sum}_Q^{-1}Q$.

Sometimes the first eigenvalue might be inadequate to capture variability due to pathway coregulation. For such occasions we could use the top M eigenvalues and test an extended null hypothesis $H_{0M}: (a_{1,1},\ldots,a_{1,M},T_1) = (a_{2,1},\ldots,a_{2,M},T_2)$; see the Supplementary Material.

4. Theoretical results

In this section, we outline theoretical results for the asymptotic behavior of $(\hat{a_1} - \hat{a_2} - \hat{b_a})$ and $(\hat{T_1} - \hat{T_2})$ under the spiked eigenvalue setting implied by our biological model under the null hypothesis and assuming Gaussian data.

We first consider a single class. Denote the population eigenvalues by $\{\alpha_i\}_{i=1}^p$ and their sample equivalents by $\{\hat{\alpha}_i\}_{i=1}^p$. We assume $a_1 > a_2 = \cdots = a_p = 1$. Write $a_1 \equiv a$, $\hat{a_1} \equiv \hat{a}$, and let $T = \sum_{k=1}^p \alpha_k$, $\hat{T} = \sum_{k=1}^p \hat{\alpha}_k$. In Theorem 1, we lay the groundwork for our method by specifying the joint asymptotic distribution of $(\hat{a}, \hat{T})$. This result is of interest beyond its application to the proposed test, and to our knowledge it gives the first published expression for the joint asymptotic distribution of $\hat{a}$ and $\hat{T}$.

Theorem 1

Suppose that $p, n \to \infty$ such that $n^{1/2}|p/n - \gamma| \to 0$ where $\gamma \in (0, 1)$. Assume $a > 1 + \gamma^{1/2}$. Let $\rho = a\{1 + \gamma/(a-1)\}$. Then

$$\sum_{\alpha T,n}^{-1/2} \begin{pmatrix} n^{1/2}(\hat{\alpha}-\rho) \\ \hat{T}-T \end{pmatrix} \to N(0, I_2),$$
$$\text{in distribution, where } \sum_{\alpha T,n} = \begin{pmatrix} \sigma_{\alpha\alpha,n} & \sigma_{\alpha T,n} \\ \sigma_{\alpha T,n} & \sigma_{TT,n} \end{pmatrix}, \quad (11)$$
$$\sigma_{\alpha\alpha,n} = \frac{2\alpha\rho}{1+\frac{\alpha\gamma}{(\alpha-1)^2-\gamma}}, \quad \sigma_{TT,n} = 2\left(\frac{\alpha^2}{n} + \frac{p-1}{n}\right), \quad \sigma_{\alpha T,n} = n^{-1/2}\alpha\rho\left\{\frac{2+K(\rho,\gamma)}{1+\frac{\alpha\gamma}{(\alpha-1)^2-\gamma}}\right\},$$

$$K(\rho,\gamma) = \frac{1}{2\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{k_\gamma(x,y)}{(\rho-x)^2}dxdy. \quad (12)$$

Here $k_\gamma(x, y)$ is a bounded, nonnegative function with support $\{(1 - \gamma^{1/2})^2, (1 + \gamma^{1/2})^2\} \times \{(1 - \gamma^{1/2})^2, (1 + \gamma^{1/2})^2\}$.

Corollary 1

Under the assumptions of Theorem 1, in distribution,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\begin{pmatrix} n^{1/2}(\hat{\alpha}-\rho) \\ \hat{T}-T \end{pmatrix}^T \sum_{\alpha T,n}^{-1} \begin{pmatrix} n^{1/2}(\hat{\alpha}-\rho) \\ \hat{T}-T \end{pmatrix} \to \chi_2^2.$$

Remark 1—The conclusions in Theorem 1 and Corollary 1 remain unchanged even if we replace γ by $\hat{\gamma} = p/n$, and a by $\bar{a} = (1/2)\ 1 + \hat{a} - \hat{\gamma} + \{(1+ \hat{a} - \hat{\gamma})^2 - 4\hat{a}\}^{1/2}$, which is obtained by solving the equation $\hat{a} = \bar{a}\{1+ \hat{\gamma}/(\bar{a} - 1)\}$.

Remark 2—If $a \gg 1+ \gamma^{1/2}$, the contribution of the term $K(\rho, \gamma)$ in the expression for $\sigma_{aT,n}$ is asymptotically negligible. In this case, $\sigma_{aT,n}$ can be replaced by

$$\tilde{\sigma}_{\alpha T,n} = n^{-1/2} 2\alpha\rho[1+\alpha\gamma/\{(\alpha-1)^2-\gamma\}]^{-1}.$$

We then apply the results of Theorem 1 to the two-class case to calculate the null distribution of our test statistic. Specifically, under H_0 in (3), without loss of generality, we assume that the common covariance matrix has eigenvalues a_0, 1,…, 1, where $\alpha_0 > 1+ \max\{\gamma_1^{1/2}, \gamma_2^{1/2}\}$, and $\gamma_j = \lim_{nj\to\infty} p/n_j \in (0, 1)$. Under the alternative, the non-spiked eigenvalues could take values other than 1.

Theorem 2

Suppose that $p, n_1, n_2 \to \infty$ such that $n_j^{1/2}|p/n_j - \gamma_j| \to 0$ where $\gamma_j \in (0, 1)$, for $j = 1, 2$. Let $\hat{a}_{jk}$ denote the k-th largest eigenvalue of the sample covariance matrix of class j, and $\hat{T}_j = \sum_{k=1}^{p} \hat{\alpha}_{jk}$. Let $\hat{b_a}$ be defined by (6). Introduce

$$\sum_{Q,n} = \begin{pmatrix} \sigma_{Q_{\alpha\alpha},n} & \sigma_{Q_{\alpha T},n} \\ \sigma_{Q_{\alpha T},n} & \sigma_{Q_{TT},n} \end{pmatrix},$$

with $\sigma_{Q_{TT},n} = 2(1/n_1 + 1/n_2)\,(\alpha_0^2 + p - 1)$;

$$\sigma_{Q_{\alpha\alpha},n} = w_2\theta_1^2 \left\{ \frac{2\alpha_0\rho_1}{1+\frac{\alpha_0\gamma_1}{(\alpha_0-1)^2-\gamma_1}} \right\} + w_1\theta_2^2 \left\{ \frac{2\alpha_0\rho_2}{1+\frac{\alpha_0\gamma_2}{(\alpha_0-1)^2-\gamma_2}} \right\},$$

where $w_j = n_j/(n_1 + n_2)$, $\rho_j = a_0\{1+ \gamma_j/(a_0 - 1)\}$,

$$\theta_j = 1+\left(\frac{1}{\gamma_1}+\frac{1}{\gamma_2}\right)^{-1} \frac{(\gamma_1-\gamma_2)}{(\alpha_0-1)^2} \frac{\kappa_j}{\gamma_j}, \quad \kappa_j = \frac{1}{2}\left[1+\frac{\rho_j-1-\gamma_j}{\{(1+\rho_j-\gamma_j)^2-4\rho_j\}^{1/2}}\right];$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and $\sigma_{Q_{\alpha T,n}} = n_1^{-1/2} w_2^{1/2} \theta_1 \left[\frac{\alpha_0 \rho_1 \{2 + K(\rho_1, \gamma_1)\}}{1 + \frac{\alpha_0 \gamma_1}{(\alpha_0 - 1)^2 - \gamma_1}} \right] + n_2^{-1/2} w_1^{1/2} \theta_2 \left[\frac{\alpha_0 \rho_2 \{2 + K(\rho_2, \gamma_2)\}}{1 + \frac{\alpha_0 \gamma_2}{(\alpha_0 - 1)^2 - \gamma_2}} \right]$,

where $K(\rho, \gamma)$ is as in (12). Then, in distribution,

$$\sum_{Q,n}^{-1/2} \begin{Bmatrix} \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} (\hat{\alpha}_{11} - \hat{\alpha}_{21} - \hat{b}_\alpha) \\ \hat{T}_1 - \hat{T}_2 \end{Bmatrix} \to N(0, I_2). \quad (13)$$

Remark 3—In Theorem 2, we can replace γ_j by $\hat{\gamma_j} = p/n_j$, and a by $\bar{a}_0$ defined through (5) without altering the conclusions.

Remark 4—The statements of both theorems remain valid even if $\gamma_j \in [1, \infty)$ for $j = 1, 2$, though the proofs change slightly. Moreover, the conclusions of Theorem 2 continue to hold even when $\gamma_j = 0$, $j = 1, 2$, with $\rho_j = 0$ and the terms $K(\rho_j, \gamma_j)$ are absent from the expressions.

Remark 5—If $a_{j1} \to \infty$, but $a_{j1} = o(p)$, for $j = 1, 2$, both theorems hold.

Remark 6—If $\alpha_0 \gg 1 + \max(\gamma_1^{1/2}, \gamma_2^{1/2})$, the contribution of the terms $K(\rho_j, \gamma_j)(j = 1, 2)$, in the expression for $\sigma_{Q_{aT,n}}$ is asymptotically negligible. In this case, we replace $\sigma_{Q_{aT,n}}$ by

$$\tilde{\sigma}_{Q_{\alpha T,n}} = n_1^{-1/2} w_2^{1/2} \theta_1 \frac{2\alpha_0 \rho_1}{1 + \frac{\alpha_0 \gamma_1}{(\alpha_0 - 1)^2 - \gamma_1}} + n_2^{-1/2} w_1^{1/2} \theta_2 \frac{2\alpha_0 \rho_2}{1 + \frac{\alpha_0 \gamma_2}{(\alpha_0 - 1)^2 - \gamma_2}}.$$

This is the expression used in defining the test statistic $Q^T \hat{\sum}_Q^{-1} Q$.

Remark 7—Both the theorems can be easily extended to cases with multiple spiked eigenvalues. See the Supplementary Material for details.

The proof of Theorem 1 uses the asymptotic expansions of the leading sample eigenvalues in Paul (2007) and the behavior of linear spectral statistics of sample covariance matrices described in Bai & Silverstein (2010). Theorem 2 follows from this and an application of the delta method.

5. Simulations

In this section, we describe simulations investigating the Type-1 error and power of the prposed test and the tests of Schott (2007) and Srivastava & Yanagihara (2010).

We consider three different sets of covariance structures. For each set, we use the same baseline covariance matrix Σ_1 and introduce different perturbations to generate Σ_2. We define Σ_1 according to the biological model in Section 2. To simulate data with p genes, we set $\sum_1 = \sigma_a^2 h h^T + I$, with $\sigma_a^2 = 35 p^{-1/2}$ and $h_{p \times 1} = \{-0.5, 1/(p-1) - 0.5, \ldots, (p-2)(p-1) -$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

0.5, 0.5}. In Σ_1, $\sigma_a^2 hh^T$ represents the variability due to pathway activities, and I represents the unordered, noisy component of pathway gene variance. In the first perturbation, which we call the added noise setting, we let $\Sigma_2 = \Sigma_1 + 0.2I$, so gene expression is subject to broader disorder in the second class. In the second perturbation, the lost co-regulation setting, we simulate pathway dysregulation by letting $\Sigma_2 = 0.7\Sigma_1 + 0.3\text{diag}(\Sigma_1)$ so that overall variability is unchanged but less well-ordered. This perturbation substantially decreases the first eigenvalue while leaving the trace unchanged. In real data, a change like this could arise from deactivation of pathway regulatory elements like transcription factors. In the third perturbation, the additional biological process setting, we let $\Sigma_2 = \Sigma_1 + gg^T$, where $g = \{g_i\}$ is defined as $g_i = 0.75$ for $i \in 1,\ldots 0.4p$ and $g_i = 0$ otherwise. In this setting, 40% of the genes in the pathway participate in a secondary biological process represented by the gg^T component.

We consider $p = 20$, 50 and 100. The corresponding first eigenvalues of Σ_1 under three different dimensions are 15.4, 22.5 and 30.8 respectively. For each p, we consider sample sizes $n_1 \in \{20, 30, 50, 75, 100, 130\}$ and $n_2 = 0.66n_1$. For each (p, n) and (Σ_1, Σ_2), we simulate 10,000 pairs of multivariate normal datasets and apply the proposed test as well as the methods of Schott (2007) and Srivastava & Yanagihara (2010) to test the differences between the two covariance matrices. We apply the robust version of the test described in in Section 3·2 for the added noise and the lost co-regulation settings, and we apply the multiple-spike version described in the Supplementary Material with $M = 2$ for the additional biological process setting. Under all three settings, we preprocess the data using the normalization scheme described in the supplementary material and derive the p-values according to the theoretical χ^2 distributions. Additionally, we examine the tests' Type-1 error rates in these settings by defining $\Sigma_0 = n_1/(n_1 + n_2)\Sigma_1 + n_2/(n_1 + n_2)\Sigma_2$, generating datasets of size n_1 and n_2 from Σ_0, and running the tests on these null datasets.

Fig. 1 displays the results of these simulations. The first row of plots displays type-I error rates of the three methods. The method of Schott (2007) is conservative, the method of Srivastava & Yanagihara (2010) is liberal, and the proposed test has the most accurate levels under all settings. The second row of plots displays powers of the three tests based on theoretical null distributions. The proposed test outperforms the others in the added noise and lost co-regulation settings and is competitive in the additional biological process setting. In the third row of plots, instead of using theoretical approximations to determine each test statistic's threshold for significance, we compute adjusted power as the percentage of test statistics under the alternative hypothesis exceeding the 0.05 quantile of the empirical null distribution of the test statistics. In this way, the type-I errors of all tests are perfectly controlled at 0.05, so the power comparison is more fair and direct. The proposed test easily outperforms the other two in term of adjusted power under all settings and all n, p combinations.

The proposed test nearly dominates the methods of Schott (2007) and Srivastava & Yanagihara (2010) in these simulations. In other simulations, we found that the methods of Schott (2007) and Srivastava & Yanagihara (2010) perform well in cases where single elements of the covariance matrix differ substantially between classes. However, changes in the biological quantities we are interested in will most often manifest as widespread, small

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

differences in the covariance matrix, a setting which these earlier methods are not optimized to detect.

We also evaluate the effects of various departures from model (1) on the performance of the proposed test through simulations. Specifically, we consider the effects of variability in the unspiked eigenvalues, unequal error variances, non-normality of the data and multiple spiked eigenvalues. We find that the proposed test is robust to all these departures except for non-normality of the data. Thus we recommend the permutation test in highly kurtotic data.

6. Application to a breast cancer dataset

We apply the proposed test to a breast cancer gene expression dataset (Loi et al., 2007), which has microarray measurements on breast tumor samples from 277 patients treated with tamoxifen and 137 untreated patients. The interest is to identify different regulation patterns between patients with or without tamoxifen treatment. We normalized all observations to have equal median and median absolute deviance. Outliers can drive the first eigenvalue of a dataset, destroying its interpretation under our biological model. We therefore truncated each gene's data in each class at four standard deviations from its mean. This rule truncated 6.4% of the data.

Curated databases of gene relationships like KEGG (Kanehisa & Goto, 2000), Reactome (Matthews et al., 2009), and Biocarta (Nishimura, 2001) often build pathways from genes involved in distantly related biological processes. Consequently, these curated pathways tend to be subject to complex co-regulation better described with network estimation tools (Peng et al., 2009; Danaher et al., 2014) than with this paper's biological model. In lieu of KEGG pathways, we sought sets of genes that could be expected to exhibit the tight co-regulation implied by our model. Cheng et al. (2013a) identified attractor metagenes, sets of genes that tended to cluster together across multiple breast cancer gene expression datasets. We expected that genes clustered together across datasets would often share a biological function, and examination of Cheng et al. (2013a)'s metagenes confirmed this hypothesis. For example, the ID55 metagene contains exclusively histone genes; and the ID88 metagene contains several genes from the cytochrome P450 family, and, intriguingly, ESR1, one of the most-studied genes in breast cancer. The biological relevance of these attractor metagenes was further demonstrated by Cheng et al. (2013b), who used attractor metagenes to inform a successful breast cancer prognostic algorithm. Given their biological meaning and apparent consistency with our biological model, we took these metagenes as the basic units of our analysis, and we ran our method and a traditional gene set analysis (Efron & Tibshirani, 2007) on every metagene with more than 5 genes.

Table 1 displays selected results; the Supplementary Material has complete results. A 2.67GHz laptop took 11 minutes to compute p-values for the 24 metagenes analyzed using 10000 permutations. The proposed biological model and test revealed a rich picture of changes in co-expression far beyond what traditional Gene Set Analysis provided. The ID88 metagene has higher total variance but a lower first eigenvalue under tamoxifen. This pattern of increased noise and decreased variability due to pathway activity strongly suggests pathway dysregulation. The histone metagene, ID55, saw increases in both overall

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

variability and its first eigenvalue under tamoxifen, suggesting more dynamic histone activity levels in the tamoxifen group. Histones are central to cancer proliferation; this result could be explained by patients heterogeneously responding to the drug. The mesenchymal transition attractor metagene followed a similar pattern, with increased variability under tamoxifen almost entirely due to an increased first eigenvalue.

The p-values returned by the χ^2 approximation and the permutation test generally tracked each other, with a Spearman correlation between them of 0.88. However, the permutation test returned uniformly higher p-values than the purely theoretical test, and one metagene, ID79, showed a markedly increased p-value under the permutation test. The liberal χ^2 p-values appear to be driven by excessively kurtotic data, and they suggest the use of the permutation test over the χ^2 approximation in highly kurtotic data.

7. Discussion

The proposed test is a powerful complement to traditional, marginal effects-based analyses like gene set analysis or tests comparing overall pathway activity levels. Given the high dimensionality and complex behavior of biological pathways, it seems appropriate to apply analyses focused on varied aspects of pathway behavior. A complete analysis of a pathway would include a summary of single-gene behavior, a comparison of overall pathway activity levels between disease states (Lee et al., 2008), a test for changes in covariance structure like the method proposed here, and ideally several other analyses yet to be discovered.

While the proposed test is motivated from the biological model (1), it can be applied to the broad class of pathways for which the first eigenvalue is spiked and reflects variability due to heterogeneous pathway activity levels. Nevertheless, not every gene set adheres to these assumptions. For example, many of the larger KEGG pathways contain genes too distally related to show discernible co-regulation. Our test is better applied to gene sets very likely to experience co-regulation, for example more narrowly-defined KEGG pathways and data-derived gene sets like the attractor metagenes of Cheng et al. (2013a) and the cancer signatures of Wolf et al. (2014). These data-derived gene sets are often highly biologically interpretable, and they have been shown to predict patient outcomes (Cheng et al., 2013b; Clarke et al., 2013; Wolf et al., 2014). It is possible to check a gene set's suitability for analysis with the proposed test by comparing the prominence of its first eigenvalue $(\hat{a}/\hat{T})$ to the $\hat{a}/\hat{T}$ of random gene sets. Various biological and technical variables will induce eigenstructure in sets of unrelated genes. If a gene set's first eigenvalue is more prominent than seen in random gene sets, the gene set is likely experiencing co-regulation. When the values of these technical, e.g. regent lot, or biological, e.g. cancer subtype, variables are known, it is possible to scrub their influence from the data by regressing each gene on these variables and performing the proposed test on the residuals.

A useful extension of this work would be the development of tests for differences in more targeted quantities than the somewhat broad $(\hat{a}, \hat{T})$. For example, a test for changes in $(\hat{T} - \hat{a})$ could be considered to directly look for increased dysregulation, or non-co-regulated variability, between classes. The asymptotic normality of $\hat{T}$ and $\hat{a}$ would make these tests simple to derive.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

An approach to this problem rooted in factor analysis could also be productive, although the factor analysis literature lacks the results for high-dimensional data that enabled our approach.

SETPath, an R package implementing the test, is on CRAN (R Core Team, 2013).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grants from the National Science Foundation and the National Institutes of Health supported this research. PD primarily worked on this method while part of the Department of Biostatistics at the University of Washington. Reviewers provided valuable input.

References

Bai, ZD.; Silverstein, JW. Spectral Analysis of Large Dimensional Random Matrices. Springer; 2010.

Baik J, Silverstein JW. Eigenvalues of large sample covariance matrices of spiked population models. Journal of Multivariate Analysis. 2006; 97:1382–1408.

Bair E, Hastie TJ, Paul D, Tibshirani RJ. Prediction by supervised principal components. Journal of the American Statistical Association. 2006; 101:119–137.

Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2005; 439:353–357. [PubMed: 16273092]

Chen X, Wang L, Smith JD, Zhang B. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics. 2008; 24:2474–2481. [PubMed: 18753155]

Cheng WY, Yang THO, Anastassiou D. Biomolecular events in cancer revealed by attractor metagenes. PLoS Computational Biology. 2013a; 9:e1002920. [PubMed: 23468608]

Cheng WY, Yang THO, Anastassiou D. Development of a prognostic model for breast cancer survival in an open challenge environment. Science Translational Medicine. 2013b; 5:181ra50.

Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, ODriscoll L, Gallagher WM, Hennessy BT, Moriarty M, Crown J, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis. 2013; 34:2300–2308. [PubMed: 23740839]

Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014; 76:373–397. [PubMed: 24817823]

Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proceedings of the National Academy of Sciences. 2013; 110:6388–6393.

Efron B, Tibshirani RJ. On testing the significance of sets of genes. The Annals of Applied Statistics. 2007; 1:107–129.

Johansson K. Shape fluctuations and random matrices. Communications in Mathematical Physics. 2000; 209:437–476.

Johnson D, Graybill F. An analysis of a two-way model with interaction and no replication. Journal of the American Statistical Association. 1972; 67:862–868.

Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. Annals of Statistics. 2001; 29:295–327.

Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000; 28:27–30. [PubMed: 10592173]

Kritchman S, Nadler B. Determining the number of components in a factor model from limited noisy data. Chemometrics and Intelligent Laboratory Systems. 2008; 94:19–32.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Computational Biology. 2008; 4:e1000217. [PubMed: 18989396]

Li J, Chen S. Two sample tests for high-dimensional covariance matrices. The Annals of Statistics. 2012; 40:908–940.

Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, et al. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. Journal of Clinical Oncology. 2007; 25:1239–1246. [PubMed: 17401012]

Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research. 2009; 37:D617–D622.

Meade AW, Bauer DJ. Power and precision in confirmatory factor analytic tests of measurement invariance. Structural Equation Modeling. 2007; 14:611–635.

Nadler B. Finite sample approximation results for principal component analysis: a matrix perturbation approach. Annals of Statistics. 2008; 36:2791–2817.

Nadler B. On the distribution of the ratio of the largest eigenvalue to the trace of a Wishart matrix. Journal of Multivariate Analysis. 2011; 102:363–371.

Nishimura D. Biocarta. Biotech Software and Internet Report. 2001; 2:117–120.

Onatski A. Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics. 2012; 168:244–258.

Paul D. Aymptotics of sample eigenstructure for a large dimension spiked covariance model. Statistica Sinica. 2007; 17:1617–1642.

Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression model. Journal of the American Statistical Association. 2009; 104:735–746. [PubMed: 19881892]

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2013.

Roy S. On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics. 1953; 24:220–238.

Schott J. A test for the equality of covariance matrices when the dimension is large relative to the sample size. Computational Statistics and Data Analysis. 2007; 51:6535–6542.

Srivastava M, Yanagihara H. Testing the equality of several covariance matrices with fewer observations than the dimension. Journal of Multivariate Analysis. 2010; 101:1319–1329.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102:15545–15550.

Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics. 2005; 6:225. [PubMed: 16156896]

Wolf DM, Lenburg ME, Yau C, Boudreau A, vant Veer LJ. Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. PloS One. 2014; 9:e88309. [PubMed: 24516633]

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

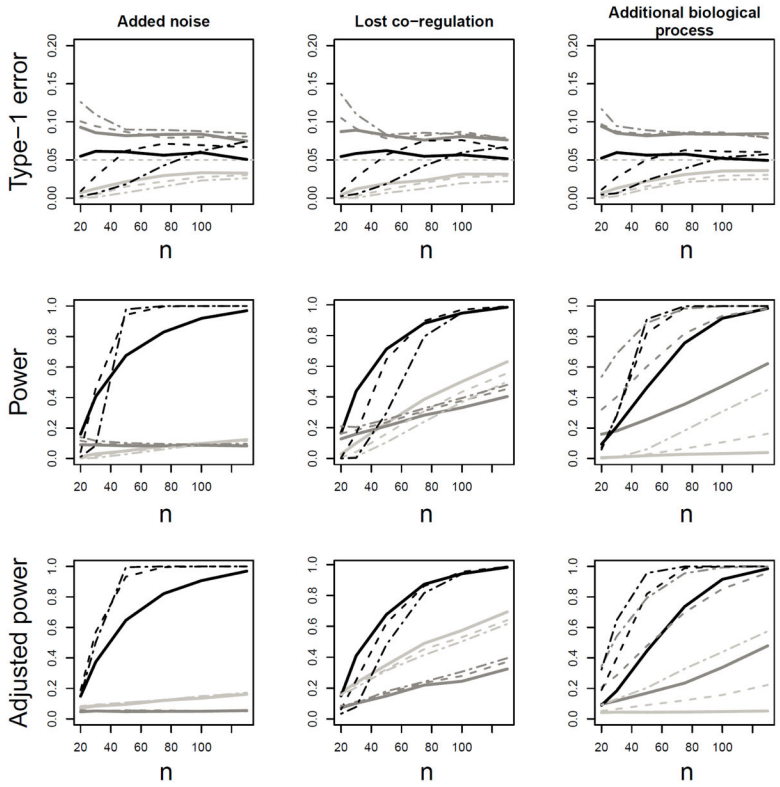

Fig. 1.

Performance of the proposed test (black lines), the method of Schott (2007) (dark grey lines) and the method of Srivastava & Yanagihara (2010) (light grey lines). Solid, dashed and dashed/dotted lines display results under $p = 20$, 50 and 100, respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Eigenvalue statistics and p-values calculated using the proposed test and Gene Set Analysis. Theoretical and permutation-based p-values from the proposed test are under p_{χ^2} and p_{perm}, respectively, and p-values from Gene Set Analysis are under p_{GSA}.

Metagene	Size	α_0^*	α_1^*	T_0^*	T_1^*	p_{χ^2}	p_{perm}	p_{GSA}
ID88	15	38.43	23.38	58.26	62.82	0.000	0.000	0.010
ID55	18	119.67	139.97	156.72	192.28	0.000	0.000	0.240
MTA**	19	127.17	167.85	157.72	205.85	0.007	0.000	0.440

* α_0 and T_0 are for the untreated group; while α_1 and T_1 are for the TAM treated group.

** MTA stands for Mesenchymal Transition Attractor.

Biometrika. Author manuscript; available in PMC 2015 September 24.