Journal of Machine Learning Research 11 (2010) 603-624 Submitted 11/07; Revised 5/09; Published 2/10

A Rotation Test to Verify Latent Structure

Patrick O. Perry PATPERRY@STANFORD.EDU
Art B. Owen OWEN@STAT.STANFORD.EDU
Department of Statistics
Stanford Unversity
Stanford CA, 94305, USA

Editor: Aapo Hyvärinen

Abstract

In multivariate regression models we have the opportunity to look for hidden structure unrelated to the observed predictors. However, when one fits a model involving such latent variables it is important to be able to tell if the structure is real, or just an artifact of correlation in the regression errors. We develop a new statistical test based on random rotations for verifying the existence of latent variables. The rotations are carefully constructed to rotate orthogonally to the column space of the regression model. We find that only non-Gaussian latent variables are detectable, a finding that parallels a well known phenomenon in independent components analysis. We base our test on a measure of non-Gaussianity in the histogram of the principal eigenvector components instead of on the eigenvalue. The method finds and verifies some latent dichotomies in the microarray data from the AGEMAP consortium.

Keywords: independent components analysis, Kronecker covariance, latent variables, projection pursuit, transposable data

1. Introduction

The problem we consider here is one of verifying statistically that an apparent latent variable is real. The context is a microarray study, although the ideas are applicable for other high throughput biological settings, and more generally for problems where a large number of mutually correlated variables has been observed.

To fix ideas, suppose we have a matrix $Y \in \mathbb{R}^{n\times N}$ of gene expression data. Each of N genes has been measured on n microarrays. We can often assume that the n arrays come from statistically independent trials, but the N genes on any single array have a rich and unknown correlation structure. There is also an $n \times p$ matrix X of predictor variables to relate to the genes ($p < n$). We are interested in which of the predictors significantly affect the response.

The motivating context is the AGEMAP project of Zahn et al. (2007) for which $n = 40$, $N = 8932$, and $p = 3$, a project whose primary goal is to find which genes are statistically correlated with age. The covariate matrix $X \in \mathbb{R}^{n\times 3}$ has columns for intercept, age, and sex. In matrix form, we model $Y = XB + E$ where B is a p-by-N matrix of regression coefficients and E is an n-by-N matrix of Gaussian or approximately Gaussian errors.

The residuals from the linear model showed a sharp dichotomy, splitting the n AGEMAP subjects into two groups. The split could not be explained by the measured variables and it strongly suggests the presence of a binary latent variable. Binary and other latent variables can arise when

©2010 Patrick O. Perry and Art B. Owen.

the microarray data are generated at different times, by different technicians, or at different sites. Sometimes a suspected latent variable can be confirmed by looking more closely at the data or lab notes. At that point we might identify the cause and use it as an ordinary regression variable. In other cases we may not be able to pinpoint the cause, but we still want statistical confirmation that it is real. When we are confident that the variable is real then it makes sense to use a model that includes one or more latent variables.

A natural way to test for a latent variable is to compute a singular value decomposition of the residual matrix $\widehat{E} = Y - X\widehat{B}$ and decide that a latent variable is present when the largest singular value of $\widehat{E}$ is sufficiently large. Equivalently, such a test is based on the eigenvalues of the covariance matrix of the rows of $\widehat{E}$. As we show below, a test based on those eigenvalues cannot work when the rows of E are correlated Gaussian random vectors. Adding a Gaussian latent variable simply changes the covariance structure and hence is not detectable. The situation is similar to that faced in independent components analysis (Hyvarinen et al., 2001) which becomes degenerate for Gaussian components.

While the eigenvalues offer no possibility to confirm the presence of a latent variable in correlated Gaussian noise, the eigenvectors of the covariance matrix do. Our tests are based on eigenvectors, rejecting the null hypothesis when the components of an eigenvector differ significantly from what we would get with Gaussian errors. Such measures have been derived for exploratory projection pursuit by Friedman (1987), although it is important to note that Friedman's measures by themselves do not measure of statistical significance. We cannot apply standard tests of Gaussianity such as the Anderson-Darling test, because such tests require an IID sample and the components of the eigenvector are not IID. Instead we show how to use a test based on random rotations of the data. When the noise is independent across observations and comes from a multivariate Gaussian distribution, then under the null hypothesis of no latent variable, random rotations don't change the distribution of our test statistic. Importantly, our test is still valid when there are arbitrary correlations between the columns of E.

Rotation tests have been used before, by Langsrud (2005). Our main contribution is to extend rotation tests to the context of regression for the explicit purpose of detecting latent variables. We show how to apply rotations orthogonally to a given linear model and we combine rotation tests with measures of non-Gaussianity of eigenvectors.

Our principal focus is on testing for the presence versus the absence of one latent variable. The regression model is usually used when we expect no latent variables. The presence of even one latent variable would make it reasonable to switch to a factor model. We also consider sequential tests for the correct number of latent variables when there is at least one of them.

The outline of the paper is as follows: Section 2 introduces the AGEMAP data as a motivating example and introduces the regression model mixing measured and latent predictors. Section 3 develops rotation tests for the existence of latent structure in the residual matrix from a regression. The tests surveyed in Langsrud (2005) need to be modified in order to rotate orthogonally to the regression model. We also show that Gaussian latent vectors cannot be detected, and then present some test statistics for non-Gaussian latent vectors. Section 4 presents numerical simulations on examples where we know the structure. The test is able to identify large latent variables and we find that it gives reliable p-values when no latent variables are present. Section 5 discusses fitting and validating latent variables for the AGEMAP data. Section 6 presents our conclusions.

2. Background

Here we describe the AGEMAP data and introduce regression models that include both measured and latent variables.

2.1 The AGEMAP Data

The motivating application arises from the AGEMAP (Zahn et al., 2007) study of aging in mice. AGEMAP is a large microarray study conducted at the National Institute on Aging and analyzed in collaboration with the Kim lab in Stanford's department of developmental biology. The primary focus of the AGEMAP analysis was to find which genes have expression levels that change with age.

Mice of ages 1, 6, 16, and 24 months were included. There were five male mice at each age and five female mice at each age. For each of these 40 mice, 16 microarrays were prepared, one for each of 16 tissues. The tissues considered were: adrenenal glands, bone marrow, cerebellum, cerebrum, eye, gonad (ovaries/testis), heart, hippocampus, kidney, liver, lung, muscle, spleen, spinal cord, striatum, and thymus.

From each of the $40 \times 16 = 640$ microarrays, values for 8932 genes were obtained. The microarrays had more than 8932 probes, but data from multiple probes corresponding to any single gene have been averaged.

We arrange the data into tissue specific matrices $Y^{(k)}$ for $k = 1, \ldots, 16$. Each $Y^{(k)}$ has 8932 columns. The matrix $Y^{(k)}$ has n_k rows, one for each sample of tissue type k. The values of n_k are unequal due to missing data. We have $32 \le n_k \le 40$ and the average sample size is 38.625. The entry $Y_{ij}^{(k)}$ is the logarithm of the expression value for mouse i and gene j in tissue k.

2.2 Regression with Measured and Latent Variables

For a single tissue type we can drop the superscript k. Zahn et al. (2007) used multiple regression analysis to investigate the effects of aging on gene expression. The regression model for gene j is

$$Y_{ij} = \beta_{0j} + \beta_{1j} A_i + \beta_{2j} S_i + \varepsilon_{ij}, \quad 1 \le i \le n \tag{1}$$

where Y_{ij} is log expression, A_i is the age in months of mouse i, and S_i is 1 if mouse i is female and is 0 otherwise. The random error term is denoted by ε_{ij}. We write all N regression models simultaneously as

$$Y = XB + E$$

where B is a p by N matrix of regression coefficients and E is an n by N matrix of random noise. The n by p design matrix X has a column for each covariate.

Now suppose that there is a latent variable taking the value U_i for the array of the ith mouse. Then adding the latent variable to the regression (1) yields

$$Y_{ij} = \beta_{0j} + \beta_{1j} A_i + \beta_{2j} S_i + \gamma_j U_i + \varepsilon_{ij}, \quad 1 \le i \le n,$$

where both γ_j and U_i are unknown. In matrix notation we have

$$Y = XB + U\Gamma + E \tag{2}$$

where $U \in \mathbb{R}^{n\times 1}$ and $\Gamma \in \mathbb{R}^{1\times N}$. If there are ℓ latent variables then the model remains as shown in (2), except that now $U \in \mathbb{R}^{n\times \ell}$ and $\Gamma \in \mathbb{R}^{\ell\times N}$.

Each individual regression includes more parameters than observations, having n latent values U_i and a coefficient γ_j. But in aggregate only $\ell(N+n)$ parameters are added to the regression for Nn observations, so the model is not saturated for small ℓ.

2.3 Forcing Identifiability

The latent variable model in (2) is not identifiable. To see why, note that if we were to replace U by $U + X\theta$ and B by $B - \theta\Gamma$, for some $\theta \in \mathbb{R}^{p\times N}$ then we would get the same residuals. A similar indeterminacy arises from replacing U by UC and Γ by $C^{-1}\Gamma$ for an invertible $C \in \mathbb{R}^{p\times p}$. We will sometimes assume that the latent variables U satisfy $X^\mathsf{T}U = 0$, $U^\mathsf{T}U = I_\ell$, and $\Gamma^\mathsf{T}\Gamma = D = \mathrm{diag}(d_1,\ldots,d_\ell)$ where $d_1 > d_2 > \cdots > d_\ell > 0$. This makes the model identifiable apart from the signs of the columns of U. Those can be specified by making the first nonzero value in each column positive. The existence of a latent term is not affected by identifiability of U, so we won't have to force U to be identifiable to detect a latent variable. We will also assume that $X^\mathsf{T}X$ has full rank p (this can be easily arranged by removing redundant predictors).

2.4 Noise Model and Estimation

In this section we describe the noise matrix E. One construction would be to assume that the entries of E are all independent and normally distributed with mean 0 and a different variance for each gene. We do not believe that the normality assumption causes serious difficulty for the AGEMAP data. But, assuming zero correlations among genes on the same array is not tenable. We assume instead that the rows of E are independent draws from the $\mathcal{N}(0,\Sigma_N)$ distribution where $\Sigma_N \in \mathbb{R}^{N\times N}$ is a gene-gene covariance matrix.

We will make frequent use of Kronecker notation. The random matrix $S \sim \mathcal{N}(M, A\otimes B)$ if its elements have a joint normal distribution with $\mathbb{E}(S_{ij}) = M_{ij}$ and $\mathrm{Cov}(S_{ij},S_{kl}) = A_{ik}B_{jl}$ for matrices M, A, and B of appropriate dimensionality.

Our model for the error is that $E \sim \mathcal{N}(0, I_n \otimes \Sigma_N)$. Then model (2) may be written as

$$Y \sim \mathcal{N}(XB + U\Gamma, I_n \otimes \Sigma_N).$$

The identifiability restrictions of Section 2.3 are as before. It will often be simpler to introduce an $n\times N$ matrix Z with IID $\mathcal{N}(0,1)$ entries and note that

$$Y \overset{d}{=} XB + U\Gamma + Z\Sigma_N^{\mathsf{T}/2} \tag{3}$$

where $\Sigma_N^{1/2} \in \mathbb{R}^{N\times N}$ satisfies $\Sigma_N^{1/2}(\Sigma_N^{1/2})^\mathsf{T} = \Sigma_N$ and $\Sigma_N^{\mathsf{T}/2}$ is a shorthand for $(\Sigma_N^{1/2})^\mathsf{T}$. Similarly $A^{-\mathsf{T}}$ means $(A^{-1})^\mathsf{T}$ for invertible A.

Because of our orthogonality constraint, $X^\mathsf{T}U = 0$, the least squares estimate of B is unaffected by the latent variable. That is $\widehat{B} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$. We find that $\widehat{B}$ is normally distributed with $\mathbb{E}(\widehat{B}) = B$ and the covariance between $\hat\beta_{ij}$ and $\hat\beta_{kl}$ is $((X^\mathsf{T}X)^{-1})_{ik}(\Sigma_N)_{jl}$. We summarize this via

$$\widehat{B} \sim \mathcal{N}(B, (X^\mathsf{T}X)^{-1} \otimes \Sigma_N).$$

We can estimate the latent term $\widehat{U}\widehat{\Gamma}$ from the residual matrix $\widehat{E} = Y - X\widehat{B}$. The least squares estimates correspond to Principal Components Analysis (PCA), and can be gotten from truncating

the singular value decomposition of $\widehat{E}$ to ℓ terms. This least squares procedure has been described by Gabriel (1978), who also incorporates column based covariates analogous to the row based ones in X. Another alternative is to use Independent Components Analysis (ICA). If there is some prior knowledge about the distribution of the possible latent factors, this knowledge can be used in the estimation procedure, for example by choosing a particular test statistic to use in Section 3.

3. Rotation Tests for Structure in the Residual Matrix

We have two tasks when dealing with a latent term. The primary task is to determine whether any latent structure exists in the residual matrix $Y - X\widehat{B}$. A secondary task is to estimate that latent structure when we believe it exists.

The main complication is that Σ_N, the noise covariance, is unknown. When Σ_N is known, we can multiply both sides (3) from the right by $(\Sigma_N^{\mathsf{T}/2})^{-1}$ and obtain a model with the same regression variables, the same number of factors, and errors that are IID $\mathcal{N}(0,1)$. Then if $n \gg N$ classical methods due first to Gollob (1968) and refined by Mandel (1971) based on nested hypothesis testing, may be applied. If instead $n \ll N$ the resulting IID $\mathcal{N}(0,1)$ errors can be handled by recent developments in random matrix theory due to Baik and Silverstein (2006) and Paul (2007a). For example, in this setting Rao and Edelman (2008) apply an approach based on the Akaike information criterion (AIC).

Most methods for identifying latent structure only look at the singular values of the residual matrix $\widehat{E}$. Since the correlation in the residuals is nontrivial and unknown, our setting leads us to consider other functions of $\widehat{E}$.

We will use the following elementary formula. If $W \sim \mathcal{N}(0, \Psi \otimes \Phi)$ then

$$BWC^{\mathsf{T}} \sim \mathcal{N}(0, (B\Psi B^{\mathsf{T}}) \otimes (C\Phi C^{\mathsf{T}})), \tag{4}$$

so long as the product matrix BWC^{T} is well defined.

3.1 Rotations Under the Null Hypothesis

Under the null hypothesis of no latent variable, our error term is $E \sim \mathcal{N}(0, I \otimes \Sigma_N)$. For any $n \times n$ orthogonal matrix $\mathbb{O}$, we find that $\mathbb{O}E \sim \mathcal{N}(0, [\mathbb{O}I\mathbb{O}^{\mathsf{T}}] \otimes \Sigma_N) = \mathcal{N}(0, I \otimes \Sigma_N)$ so that $\mathbb{O}E \stackrel{d}{=} E$. The original residual matrix, E, and the rotated residual matrix, $\mathbb{O}E$ have the same distribution. This fact provides our starting point.

We refer to orthogonal matrices as rotations. A stricter usage of the term requires $\det(\mathbb{O}) = 1$ but following Langsrud (2005) we allow $\det(\mathbb{O}) = -1$ as well. Such reflections, as they are sometimes called, also preserve the distribution of E so they are worth including. In a rotation test we compare some aspect of the data to its value under repeated random rotations of the data. Such rotation tests are analogous to the more familiar permutation tests. Rotation tests were first introduced by Wedderburn (1975) and Heiberger (1978). A recent survey appears in Langsrud (2005) who focuses on multiple testing issues.

Here we give a self-contained derivation of rotation tests for multiple regression. The regression context requires us to make some modifications to the method.

The data are $Y = XB + E$ with $E \stackrel{d}{=} U\Gamma + Z\Sigma_N^{\mathsf{T}/2}$ where Z_{ij} are IID $\mathcal{N}(0,1)$. The matrix X has rank $p < n$ and hat matrix $H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$. The residual matrix is $\widehat{E} = (I-H)Y = (I-H)XB + (I-H)E = (I-H)E$. We will apply many random rotations to $\widehat{E}$. This is mathematically equivalent

to rotating both X and Y each time, and then taking the residuals from the rotated variables, but of course it is faster to simply rotate the residuals. To prove equivalence:

Proposition 1 *For integers $n > p > 0$ and $N \geq 1$, let $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^{n\times N}$, $B \in \mathbb{R}^{p\times N}$ and $E \in \mathbb{R}^{n\times N}$ satisfy $Y = XB + E$. Suppose that $H = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ has rank p. Let $\mathbb{O} \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. Put $\widetilde{Y} = \mathbb{O}Y$, $\widetilde{X} = \mathbb{O}X$, $\widetilde{H} = \widetilde{X}(\widetilde{X}^\mathsf{T}\widetilde{X})^{-1}\widetilde{X}^\mathsf{T}$ and $\widetilde{E} = (I - \widetilde{H})\widetilde{Y}$. Then $\widetilde{E} = \mathbb{O}(I - H)Y$.*

Proof The rotated hat matrix satisfies $\widetilde{H} = \mathbb{O}X(X^\mathsf{T}\mathbb{O}^\mathsf{T}\mathbb{O}X)^{-1}X^\mathsf{T}\mathbb{O}^\mathsf{T} = \mathbb{O}H\mathbb{O}^\mathsf{T}$. The new residual is $\widetilde{E} = \widetilde{Y} - \widetilde{H}\widetilde{Y}$. Now $\widetilde{H}\widetilde{Y} = \mathbb{O}H\mathbb{O}^\mathsf{T}\mathbb{O}Y = \mathbb{O}HY$. Finally $\widetilde{E} = \mathbb{O}Y - \mathbb{O}HY = \mathbb{O}(I - H)Y$. ■

In the regression context, randomly rotating the residuals does not generally preserve their distribution. First because $(I - H)(I - H)^\mathsf{T} = (I - H)$, we find that under the null hypothesis ($U = 0$):

$$\widehat{E} \sim \mathcal{N}(0, (I - H) \otimes \Sigma_N).$$

But then, for the rotated residual we have

$$\mathbb{O}\widehat{E} \sim \mathcal{N}(0, [\mathbb{O}(I - H)\mathbb{O}^\mathsf{T}] \otimes \Sigma_N),$$

by Equation (4). These distributions do not usually coincide.

We will fix this problem by restricting attention to a special subset of rotation matrices. The desired rotations $\mathbb{O}$ satisfy $\mathbb{O}H\mathbb{O}^\mathsf{T} = H$ (equivalently $\mathbb{O}(I - H)\mathbb{O}^\mathsf{T} = (I - H)$) for then the rotation does not change the distribution of $\widehat{E}$. The rotations we want will fix X but rotate the space orthogonal to X. Specific construction details follow.

Because H has p eigenvalues equal to 1 and $n - p$ eigenvalues equal to 0 we may write it as $H = Q_1 Q_1^\mathsf{T}$ where $Q_1 \in \mathbb{R}^{n\times p}$ satisfies $Q_1^\mathsf{T} Q_1 = I_p$. Let $Q_2 \in \mathbb{R}^{n\times(n-p)}$ be a matrix such that $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ is orthogonal. For our construction, we let $\mathbb{O}_* \in \mathbb{R}^{(n-p)\times(n-p)}$ be an orthogonal matrix and then take

$$\mathbb{O} = Q_1 Q_1^\mathsf{T} + Q_2 \mathbb{O}_* Q_2^\mathsf{T}. \tag{5}$$

The matrices produced by Equation (5) are orthogonal and satisfy $\mathbb{O}H\mathbb{O}^\mathsf{T} = H$. We summarize as follows:

Proposition 2 *Let $Y \sim \mathcal{N}(XB, I_n \otimes \Sigma_N)$ where $X \in \mathbb{R}^{n\times p}$ has rank $p < n$ and $B \in \mathbb{R}^{p\times N}$. Let $\widehat{E} = (I - H)Y$ where $H = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ and let $\widetilde{E} = \mathbb{O}\widehat{E}$ where $\mathbb{O}$ satisfies* (5)*. Then both $\widehat{E}$ and $\widetilde{E}$ have the $\mathcal{N}(0, (I - H) \otimes \Sigma_N)$ distribution.*

Proof Let $E = Y - XB \sim \mathcal{N}(0, I_n \otimes \Sigma_N)$. Then $\widehat{E} = (I - H)Y = (I - H)E \sim \mathcal{N}(0, (I - H) \otimes \Sigma_N)$ as above. Now suppose that $\mathbb{O}$ satisfies (5). Then $\widetilde{E} = \mathbb{O}(I - H)E \sim \mathcal{N}(0, [\mathbb{O}(I - H)\mathbb{O}^\mathsf{T}] \otimes \Sigma_N)$. Next

$$\begin{aligned}
\mathbb{O}(I - H)\mathbb{O}^\mathsf{T} &= \left(Q_1 Q_1^\mathsf{T} + Q_2 \mathbb{O}_* Q_2^\mathsf{T}\right)(I - H)\left(Q_1 Q_1^\mathsf{T} + Q_2 \mathbb{O}_* Q_2^\mathsf{T}\right) \\
&= \left(Q_1 Q_1^\mathsf{T} + Q_2 \mathbb{O}_* Q_2^\mathsf{T}\right) Q_2 Q_2^\mathsf{T} \left(Q_1 Q_1^\mathsf{T} + Q_2 \mathbb{O}_*^\mathsf{T} Q_2^\mathsf{T}\right) \\
&= Q_2 \mathbb{O}_* \mathbb{O}_*^\mathsf{T} Q_2^\mathsf{T} \\
&= I - H,
\end{aligned}$$

and so $\widetilde{E} \overset{d}{=} \widehat{E}$. ■

To complete this section we show that Equation (5) generates all of the desired rotations.

Proposition 3 *Let $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \in \mathbb{R}^{n\times n}$ be an orthogonal matrix, where $Q_1 \in \mathbb{R}^{n\times p}$ for $n > p > 0$. Let $\mathbb{O} \in \mathbb{R}^{n\times n}$ be an orthogonal matrix and write $H = Q_1 Q_1^\mathsf{T}$. If $\mathbb{O}H\mathbb{O}^\mathsf{T} = H$ then $\mathbb{O} = Q_1\mathbb{O}_\circ Q_1^\mathsf{T} + Q_2\mathbb{O}_* Q_2^\mathsf{T}$ where $\mathbb{O}_\circ \in \mathbb{R}^{p\times p}$ and $\mathbb{O}_* \in \mathbb{R}^{(n-p)\times(n-p)}$ are orthogonal matrices.*

Proof First, any orthogonal matrix $\mathbb{O}$ can be written as $\mathbb{O} = QPQ^\mathsf{T}$ where

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$$

is orthogonal, and partitioned with $P_{11} \in \mathbb{R}^{p\times p}$, $P_{22} \in \mathbb{R}^{(n-p)\times(n-p)}$ and so on. We may take $P = Q^\mathsf{T}\mathbb{O}Q$. Now $\mathbb{O}H\mathbb{O}^\mathsf{T} = QPQ^\mathsf{T}Q_1Q_1^\mathsf{T}QP^\mathsf{T}Q^\mathsf{T} = FF^\mathsf{T}$where

$$F = QPQ^\mathsf{T}Q_1 = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}\begin{pmatrix} Q_1^\mathsf{T} \\ Q_2^\mathsf{T} \end{pmatrix} Q_1 = Q_1P_{11} + Q_2P_{21}.$$

Assume that $\mathbb{O}H\mathbb{O}^\mathsf{T} = H$. Then $Q_1^\mathsf{T}\mathbb{O}H\mathbb{O}^\mathsf{T}Q_1 = I_p$ so that $(Q_1^\mathsf{T}F)(Q_1^\mathsf{T}F)^\mathsf{T} = I_p$. But $Q_1^\mathsf{T}F = P_{11}$. Therefore $P_{11}P_{11}^\mathsf{T} = I_p$, or in other words P_{11} is an orthogonal matrix. The columns of P_{11} are unit vectors and so are those of P. Therefore $P_{21} = 0$. Similarly $P_{12} = 0$ and P_{22} is an orthogonal matrix. Taking $\mathbb{O}_\circ = P_{11}$ and $\mathbb{O}_* = P_{22}$ completes the proof. ■

From Proposition 3 we see that the suitable rotations take the form $\mathbb{O} = Q_1\mathbb{O}_\circ Q_1^\mathsf{T} + Q_2\mathbb{O}_* Q_2^\mathsf{T}$. In Equation (5), we only use $\mathbb{O}_\circ = I_p$. We don't need to vary that part of the rotation because in our application we work with

$$\mathbb{O}(I-H) = \mathbb{O}Q_2Q_2^\mathsf{T} = (Q_1\mathbb{O}_\circ Q_1^\mathsf{T} + Q_2\mathbb{O}_* Q_2^\mathsf{T})Q_2Q_2^\mathsf{T} = Q_2\mathbb{O}_* Q_2^\mathsf{T}.$$

The choice of $\mathbb{O}_\circ$ does not affect the value of $\widetilde{E} = \mathbb{O}(I-H)Y$, and so we may simply take $\mathbb{O}_\circ = I_p$.

3.2 Testing for the Existence of Structure

Here we construct a test for latent structure in the residual matrix. The null hypothesis is $H_0 : U = 0$ and the alternative is $H_1 : U \neq 0$. We will construct a rotation test by modifying the rotation tests in Langsrud (2005). As in the discussion of the randomization tests in Lehmann and Romano (2005) we find a group of rotations. It is easy to show that $O_H = \{Q_1Q_1^\mathsf{T} + Q_2\mathbb{O}_* Q_2^\mathsf{T} \mid \mathbb{O}_* \in \mathbb{R}^{(n-p)\times(n-p)}, \mathbb{O}_*^\mathsf{T}\mathbb{O}_* = I_{n-p}\}$ is a group under multiplication, because orthogonal $r-p$ by $r-p$ matrices are a group. Southworth et al. (2009) give a cautionary note on randomizations without a group structure.

Proposition 2 allows us to perform a test of the hypothesis as follows: let $T : \mathbb{R}^{n\times N} \to \mathbb{R}$ be any statistic of the residual matrix. Specific examples are given in Section 3.4. We generate independent n by n random rotations $\mathbb{O}_1, \ldots, \mathbb{O}_{R-1}$ uniformly from O_H. Then, construct a p-value as

$$\hat{p} = \frac{1}{R}\Big(1 + \sum_{i=1}^{R-1} 1\{T(\mathbb{O}_i\widehat{E}) \geq T(\widehat{E})\}\Big). \tag{6}$$

Since under the null hypothesis $\mathbb{O}_i\widehat{E}$ and $\widehat{E}$ have the same distribution and since O_H is a group, this gives us a valid p-value. The leading 1 in $\hat{p}$ counts the observed $\widehat{E}$ and prevents us from claiming $\hat{p}$ below $1/R$ if we have only seen R rotations (including the original one).

3.3 Gaussian Alternative

Here we show that when Σ_N is unknown, a Gaussian latent variable cannot be detected. This was remarked on by Snee (1982) and it is well known in the independent component analysis community; see Hyvarinen et al. (2001).

To see how the problem manifests, consider the model in (3) with just one latent variable with entries $U_i \sim \mathcal{N}(0,1)$ independently of Z. The regression model XB is assumed nonrandom. To focus on essentials we do not initially impose the normalization $U^\mathsf{T}U = I_\ell$.

Proposition 4 *For positive integers N, n, and ℓ, let $Z \sim \mathcal{N}(0, I_n \otimes I_N)$ independently of $U \sim \mathcal{N}(0, I_n \otimes I_\ell)$. Let $\Gamma \in \mathbb{R}^{\ell \times N}$. Then $U\Gamma + Z\Sigma_N^{\mathsf{T}/2} \stackrel{d}{=} Z\widetilde{\Sigma}_N^{\mathsf{T}/2}$ where $\widetilde{\Sigma}_N = \Sigma_N + \Gamma^\mathsf{T}\Gamma$.*

Proof We only need to show that $U\Gamma + Z\Sigma_N^{\mathsf{T}/2}$ has the same distribution as $Z\widetilde{\Sigma}_N^{\mathsf{T}/2}$. The matrix $U\Gamma + Z\Sigma_N^{\mathsf{T}/2}$ has independent identically distributed rows from $\mathcal{N}(0, \Sigma_N + \Gamma^\mathsf{T}\Gamma)$. Therefore $U\Gamma + Z\Sigma_N^{\mathsf{T}/2}$ has representation $\widetilde{Z}\widetilde{\Sigma}_N^{\mathsf{T}/2}$ where $\widetilde{Z} \in \mathbb{R}^{N \times n}$ has IID entries from $\mathcal{N}(0,1)$. This $\widetilde{Z}$ has the same distribution as Z. ■

If we do normalize U then nothing essential changes. We replace U by UC for a random normalizing matrix $C \in \mathbb{R}^{\ell \times \ell}$ that is independent of Z (that is, $UC \sim \mathcal{N}(0, I_n \otimes I_\ell)$ and $U^\mathsf{T}U = I$). Then we compensate by replacing Γ by $C^{-1}\Gamma$ and get $U\Gamma + Z\Sigma_N^{\mathsf{T}/2} \stackrel{d}{=} Z\widetilde{\Sigma}_N$ where $\widetilde{\Sigma}_N = \Sigma_N + \Gamma^\mathsf{T}C^{-\mathsf{T}}C^{-1}\Gamma$ is now random.

The implication of Proposition 4 is that if we don't know anything about Σ_N, or Γ, then a Gaussian latent variable is impossible to detect. There is no mathematical difference between a Gaussian latent vector and a changed correlation structure. Put another way, such latent variables are already well accounted for in the correlation structure.

Latent variables of practical interest typically exhibit non-Gaussian traits like clumping or outliers. Also, if a latent variable corresponds to a roughly-linear time trend, then it will be nearly uniformly distributed if the points are sampled at regular time intervals. Therefore this restriction still leaves many interesting testing problems.

3.4 Choice of the Test Statistic, T

Since a Gaussian latent variable is covered by the correlation model and is not detectable, any effective test statistic T must be tuned for non-Gaussian latent variables. A non-Gaussian latent variable makes for an error term $U\Gamma + Z\Sigma_N^{\mathsf{T}/2}$ that does not have a rotationally invariant distribution.

In principle any function $T(E)$ can be used. However, the choice of T will often be dictated by what we deem to be interesting structure. Here we describe four different possibilities for T. We first apply a simplification procedure to reduce E to a vector $u(E)$. Then, we apply a function to reduce u to a scalar. The end result is a scalar-valued function $T(E)$.

We would like $u(E)$ to be representative of latent structure in E. An obvious choice is the first left singular vector of E, which corresponds to the first principal component of E. A second choice is to apply ICA to E, treating the columns as mixtures of n-dimensional sources, and have $u(E)$ be the first estimated source. In both cases $u(E)$ is a unit vector.

We cannot simply use a test for normality of the components of $u(E)$, such as the Anderson-Darling test, because the components we get are not independent $\mathcal{N}(0,1)$ even under H_0.

Instead, we propose two functions for reducing $u = u(E)$ to a scalar. The first is the L^1 norm of the vector:

$$T_{L^1}(u) = \sum_{i=1}^{n} |u_i|.$$

As a point of reference, for independent $u_i \sim \mathcal{N}(0, \frac{1}{n})$ we would get $T_{L^1} \doteq \sqrt{2n/\pi}$. So, the expected L^1 norm of a uniformly distributed n-dimensional unit vector is approximately $\sqrt{2n/\pi} \doteq 0.798\sqrt{n}$. Larger values of T_{L^1} correspond to distributions whose expected absolute value is large compared to their root mean square. Uniform distributions on $[-1,1]$ or $\{-1,1\}$ behave this way. Conversely, small values of T_{L^1} arise from very heavy tailed distributions like the Cauchy which have outliers.

The rotation based p-value (6) is sensitive to large values of T_{L^1} and should therefore catch dichotomies and light tailed latent variables. To detect heavy tailed alternatives we could use (6) with $1/T_{L^1}$. Because we are potentially interested in both kinds of non-Gaussian latent structure we take

$$\hat{p} = \frac{2}{R} \min\left(1 + \sum_{i=1}^{R-1} 1\{\widetilde{T}_i \geq \widehat{T}\},\ 1 + \sum_{i=1}^{R-1} 1\{\widetilde{T}_i \leq \widehat{T}\}\right), \tag{7}$$

where $\widehat{T} = T(\widehat{E})$ and $\widetilde{T}_i = T(\mathbb{O}_i\widehat{E})$ and $T(\cdot)$ subsumes all the computation in T_{L^1}. The leading 2 in (7) compensates for using the more extreme of two tails.

The second test statistic comes from Exploratory Projection Pursuit (Friedman, 1987). T_{EPP} is a distance measure on densities, represented as a Legendre-series and then truncated to 4 terms:

$$T_{\text{EPP}}(u) = \sum_{j=1}^{4} \left(j + \frac{1}{2}\right) \left(\mathbb{E}P_j(R)\right)^2$$

where, P_j is the j-th Legendre polynomial, R is a random variable uniformly distributed over the discrete set $\{2\Phi(u_i) - 1\}_{i=1}^{n}$, Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution, and $\mathbb{E}$ denotes expectation over the randomness in R. The Legendre polynomials can be computed using the recurrence relation

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \quad \text{and,} \\ (j+1)P_{j+1}(x) &= (2j+1)xP_j(x) - jP_{j-1}(x). \end{aligned}$$

In computing T_{EPP} we use

$$\mathbb{E}P_j(R) = \frac{1}{n} \sum_{i=1}^{n} P_j(2\Phi(u_i) - 1).$$

T_{EPP} is designed to be close to 0 when the histogram of u looks Gaussian, and it gets bigger the more "non-Gaussian" u is. The full derivation of T_{EPP} is given by Friedman (1987).

Because only large values of T_{EPP} are interesting we use (6) directly without making a two-tailed modification.

We can combine PCA or ICA with $T_{\text{EPP}}(u)$ or $T_{L^1}(u)$ and get four different test statistics. We write $T_{\text{EPP}}(E)$ or $T_{L^1}(E)$ when the context dictating PCA or ICA for u is clear. The choice of T is independent of the procedure for estimating the latent variable. In particular, it is possible to detect the existence of latent structure in $\widehat{E}$ using a PCA-based test statistic and then fit the structure using ICA. Indeed, in simulations it turns out to be better to use PCA for u. This will be further explored in Section 4.

3.5 Identifying the Rank of the Latent Term

When we are able to reject the null hypothesis we conclude that some latent structure exists, but we do not know the rank of U. To estimate the number of latent variables we consider a sequential approach, based on subtracting estimated latent variables and looking for latent structure in the residuals.

First, we fit the model $\widehat{Y}_0 = X_0\widehat{B}_0$ and get the residual matrix $\widehat{E}_0$. The subscripts denote a model with 0 latent terms. Next, we test for latent structure in $\widehat{E}_0$. If we determine that any structure exists, we fit a single latent variable $\hat{u}_1$.

At this stage, we treat $\hat{u}_1$ as a known covariate. We create a new covariate matrix $X_1 = [X_0 \;\; \hat{u}_1]$ by appending $\hat{u}_1$ as a column onto the old covariate matrix. Now, we fit the a new model $\widehat{Y}_1 = X_1\widehat{B}_1$ and get a new residual matrix $\widehat{E}_1$. We proceed in a sequential matter: test for structure in $\widehat{E}_i$; upon identifying structure, fit a single latent variable, treat it as a covariate, and get a new residual matrix $\widehat{E}_{i+1}$. We stop when there is no latent structure in $\widehat{E}_i$.

In the context of PCA, fitting $\hat{u}_2$ sequentially after adjusting $\widehat{E}$ for $\hat{u}_1$ is equivalent to fitting $\hat{u}_1$ and $\hat{u}_2$ simultaneously. For other estimation methods, this equivalence may not hold. The procedure we describe is still valid, but there may be some loss in power. If this is a concern, the practitioner can adjust the testing procedure accordingly.

One has to be careful when testing for more than one latent term. In particular, for some settings when $n \ll N$, it is impossible to consistently estimate the latent variables $\hat{u}_i$. When we do not have a good estimate of $\hat{u}_i$, treating it as a known covariate will introduce a potentially serious error. When this happens, only the p-value for the first term is reliable. This point is illustrated in the example of Section 4.3, where the error in the p-value distribution was small.

3.6 Caveats

When we reject the null hypothesis, then either there is strong enough latent structure in the data, or the noise is far from Gaussian. Therefore, rejecting the null hypothesis is *necessary* to deem latent structure to be real, but not sufficient. Often there is ambiguity between what constitutes non-Gaussianity and what can be explained by a latent variable. An outlier can be modeled using a latent variable that has support on a single observation. Bi-modal noise can be re-cast as a clumping latent effect.

3.7 Related Work

Rank determination methods have been the subject of much interest in crop science. For a recent survey see Crossa and Cornelius (2002). Those methods tend to focus on the amount of variance explained by the first principal component. In an eigen-analysis of $Y - X\widehat{B}$, they focus on the size of the eigenvalues. There has been considerable difficulty with getting tests to have the right level, as described for example by dos S Dias and Krzanowski (2003). The core problem is that there is no good way to count the degrees of freedom for such data sets, despite recent progress in random matrix theory including El Karoui (2007), Paul (2007b), and Nadler (2007). Owen and Perry (2009) apply a cross-validation-based approach to rank determination for the truncated SVD and non-negative matrix factorization. That work requires independent noise, not the correlated noise we consider here. Efron (2009) uses permutations to test whether some microarrays are independent of each other.

4. Empirical Testing

In this section we examine the performance of rotation tests on constructed examples where we know the answer. Some readers may prefer to read the real data example of Section 5 first. In our constructed examples, the response satisfies

$$Y \sim \mathcal{N}(XB + U\Gamma, I_n \otimes \Sigma_N),$$

with parameters described below.

4.1 Microarray Model

This example is designed to resemble microarray studies. We take $p = 2$, $n = 20$ and $N = 256$. This value of N is small, to allow a larger number of simulated cases. The matrix X has a first column of 1s. The second column has values $1, \ldots, n$. We take B to be a $p \times N$ matrix of 0s. Having $B = 0$ is no loss of generality, because the analysis works on residuals after regression on X and the residuals are unaffected by B.

We construct latent variables $U \in \mathbb{R}^{n\times 3}$. The first latent variable, u_1, is the first column of U and has independent elements distributed as Cauchy random variables. The second latent variable, u_2, has elements which are either -1 or 1 with equal probability. The third latent variable, u_3, has elements that are independent and exponentially distributed with mean 1. Thus, u_1 is an "outlier" effect, u_2 is a "clumping" effect, and u_3 is some other latent effect.

The latent coefficient matrix, Γ, has independent elements distributed as $\mathcal{N}(0,1)$. We do not think that non-normal Γ would make the signal artificially easy to detect, but taking Gaussian Γ removes any such worry. As described, U, Γ, and X do not satisfy the identifiability conditions of Section 2.3. The existence of an unnormalized latent variable implies that a normalized one exists, and so the testing problem is unaffected.

For Σ_N we need a 256×256 correlation matrix. The true correlation patterns for microarray data are not known. The sample sizes to date are far too small to allow confident description of the patterns. Owen (2005) looks at what gene-gene correlations are like in real data. We mimic two features of microarray data. First, genes are often thought to belong to relatively small clusters. Second, the mean of the squared estimated off-diagonal sample correlations is often seen to be a small multiple of $1/n$. The value $1/n$ is very close to what we would expect in the event that all true correlations were zero. To encode the first property, we take

$$\Sigma_{ij} = \begin{cases} 1 & i = j, \\ \rho & \lfloor (i-1)/32 \rfloor = \lfloor (j-1)/32 \rfloor, \quad \& \quad i \neq j, \\ \rho & i - j \equiv 0 \bmod 32, \quad \& \quad i \neq j, \\ 0 & \text{else.} \end{cases}$$

In words, gene i belongs to two clusters: one cluster of 8 genes corresponding to the least-significant digit of $i-1$ in base 32, and one cluster of 32 genes corresponding to the most-significant digit of $i-1$ in base 32. Gene i has 38 non-zero correlations with other genes. The value of $\rho > 0$ is chosen so that signal is about 30% of the noise:

$$\frac{\sum_{i=1}^N \sum_{j=1}^N \Sigma_{ij}^2 - N}{N(N-1)/n} = 0.30.$$

Thus $38\rho^2 = (N-1)/(0.30\,n)$ so $\rho = \sqrt{0.30\,(N-1)/(38n)} \doteq 0.317$.

4.2 Rotation Tests

The true model has three latent variables. We are interested in what happens when testing for the first, second, third, and fourth latent terms. We look at two different choices for the test statistic: T_{EPP} in conjunction with PCA, and T_{EPP} in conjunction with ICA. The results are summarized in ROC curves in Figure 1.

The upper-right panel shows the results from testing the residual matrix after one latent term has been removed. Here, we see that testing with u_{PCA} results in about 75% of the replicates having estimated p-values less than 0.2, while testing with u_{ICA}, results in about 55%. Generally, the PCA test gave us higher power. We also found that FastICA can get stuck in a local minimum. This is what lead to its surprisingly poor performance in the upper left panel. The latent variables of the randomly-rotated data are more non-Gaussian than the latent variable estimated from the original data.

The lower-right panel shows the estimated p-values after three latent terms have been fit and removed from the residual matrix. As expected, the estimated p-values are close to the specified false-positive rates. Comparing the lower left panel to the others, we see, unsurprisingly that the smallest latent vector is hardest to detect while the largest is easiest to detect. Finally, testing for a fourth latent variable gives us a uniform p-value, which is exactly what we want since there are only three latent terms.

A word is in order about how we removed the first latent variable when testing for the presence of the second. We tried removing vectors as estimated by PCA and also by ICA. There was not much difference in performance, and PCA has the computational advantage that the estimated second vector does not change when we remove the first. Therefore when testing for the k'th vector, whether by ICA or PCA, we always used PCA to remove the first $k-1$ of them.

4.3 Testing Under and Near the Null Hypothesis

In the previous simulation, the signal-to-noise ratio between the latent effect terms and the random error is relatively high, and so the p-values for non-existent latent terms are faithful. In this simulation, we demonstrate that the p-values for testing for multiple latent variables are slightly liberal if the signal strength is too weak, but these p-values are still within tolerable accuracy.

We generate an $n \times N$ data matrix Y according to the model

$$Y = (N\lambda)^{1/2} u\gamma^{\mathsf{T}} + Z\Sigma_N^{\mathsf{T}/2},$$

with $n = 20$ and $N = 200$. There is a single latent variable u which has elements equal to -1 or $+1$ with equal probability. The coefficient vector γ is a uniformly distributed random unit vector in $\mathbb{R}^N$. The noise covariance Σ_N is a diagonal matrix with $(\Sigma_N)_{ii}$ independent from all other entries and exponentially-distributed with mean 1; $\Sigma_N^{1/2}$ is its square root. The noise variable matrix Z has IID $\mathcal{N}(0,1)$ elements. We choose λ to be a fixed scalar, specified below.

The theory in Section 3.2 tells us that the p-value from a rotation test of a single latent term is uniformly-distributed when $\lambda = 0$. However, it tells us nothing about p-values for a second term. Regardless of the value of λ, we would like them to be uniformly distributed, so that the test is faithful to the specified false positive rate. The issue is whether errors in the estimated first latent vector spoil the test for the second. Results in Onatski (2007) suggest that as the sample size goes to infinity, p-values from the second and higher terms will be faithful when we fit with PCA. Our sample size is only 20, so we do an empirical test.

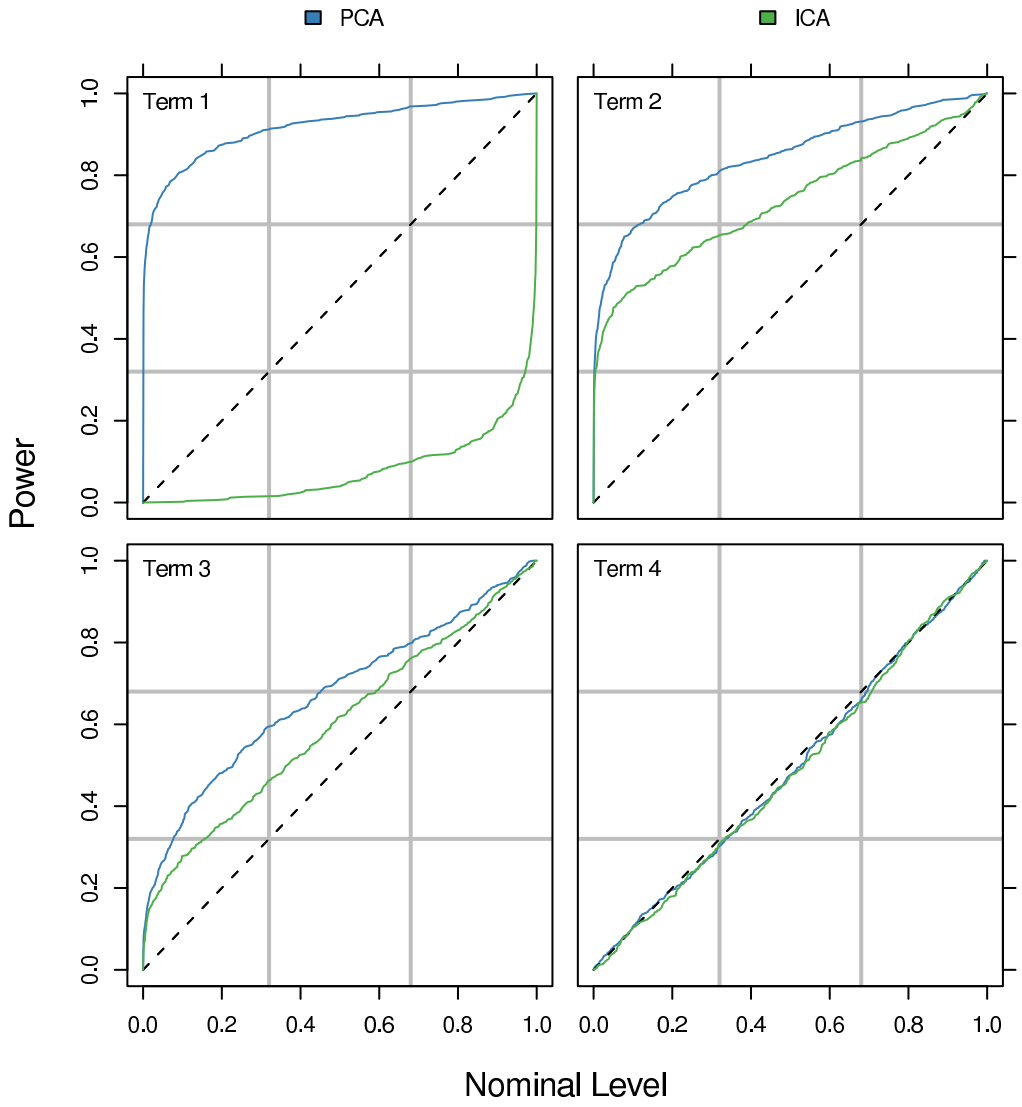

Figure 1: MULTI-FACTOR ROTATION TESTS WITH PCA AND ICA. We simulate data from a model with three latent terms and then apply rotation tests for latent structure. The test statistic is $T_{\text{EPP}}(u)$ applied to either the first principal component u_{PCA}, or the first independent component u_{ICA} (via FastICA) of the residual matrix. The plots show estimated ROC curves after 0, 1, 2, or 3 principal components have been fit and removed. The x axis of each plot is the specified false-positive rate. The y axis is the proportion of replicates with an estimated p-value below that level, using 500 total replicates of the data set. The plots are discussed further in the text.

For all λ in the set $\{0, 0.5, 1, 5, 10, 50, 100, 500, 1000\}$, we perform the following simulation, which we repeat 1000 times:

1) Generate data Y as described above.

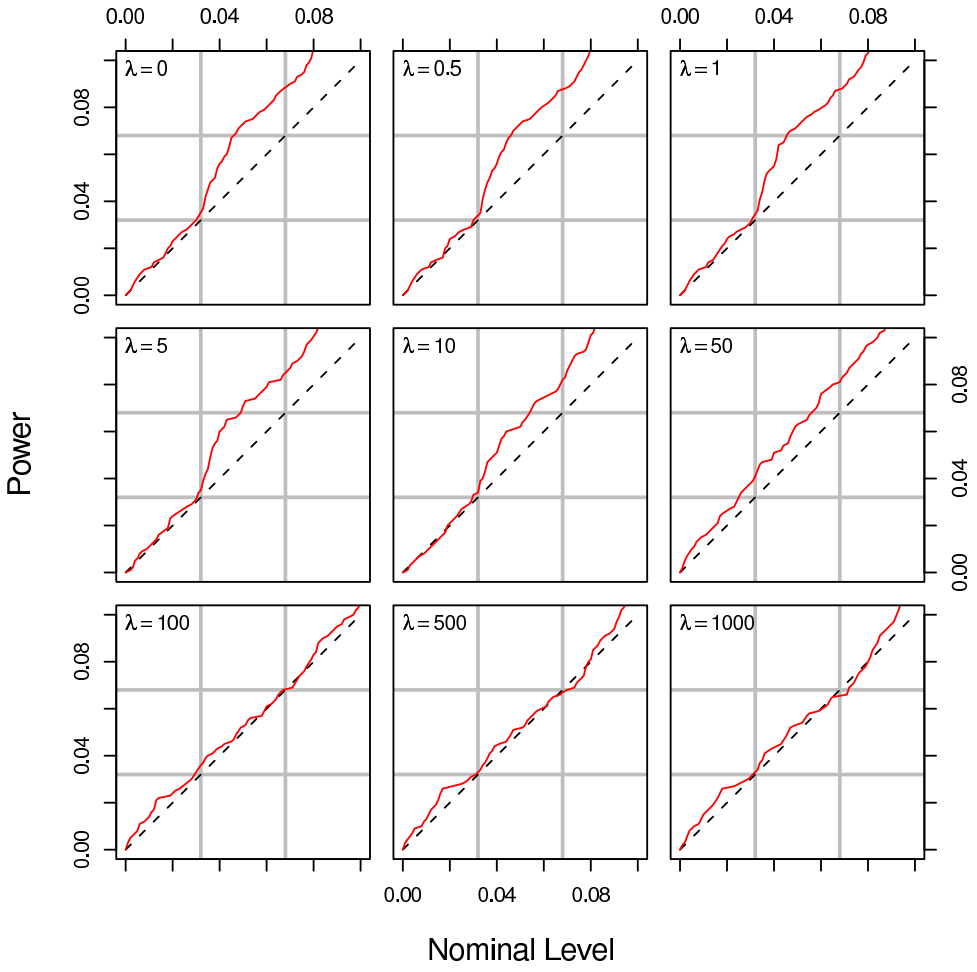

Figure 2: TESTING FOR A NON-EXISTENT SECOND TERM. We estimate a factor generated according to $(N\lambda)^{1/2}u\gamma^{\mathsf{T}}$ and then test for a second term. Depending on the factor strength (which is related to our ability to estimate), the p-values for validating the second term may be slightly liberal. This seems to be the case for $\lambda \leq 50$. For $\lambda \geq 100$, the p-values appear to be faithful.

2) Fit a single latent variable $\hat{u}$, using the first term in the SVD of Y.
3) Construct the matrix of residuals as $\widehat{E} = Y - \hat{u}\hat{u}^{\mathsf{T}}Y$.
4) Test for the existence of more latent terms using a PCA-based rotation test using T_{EPP} as our test statistic and treating $\hat{u}$ as a known covariate. Record the p-value estimated from 999 random rotations.

We would like to assess the implications of treating $\hat{u}$ as a known covariate. When λ is big, this is a reasonable assumption since the term is easy to estimate, but when λ is small this is not the case.

We summarize the results in Figure 2. We can see that for $\lambda \leq 50$, the small p-values are slightly liberal. When $\lambda \geq 100$, the p-values appear to be faithful. When the first latent variable is strong, then we have a reliable test for the second.

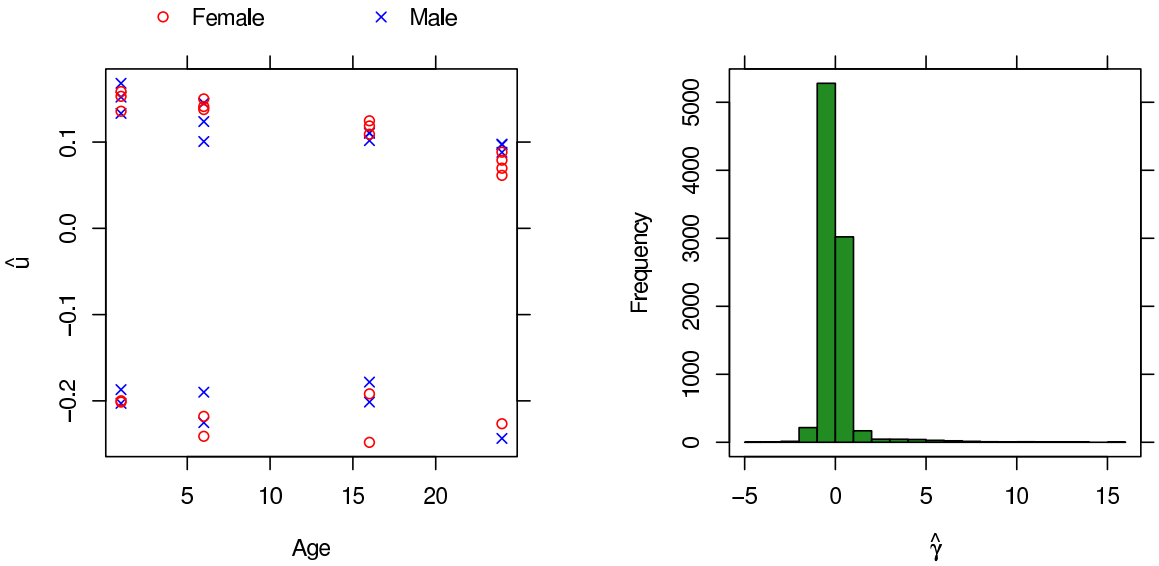

Figure 3: CEREBELLUM LATENT VARIABLE. The left plot shows the latent variable estimated for the cerebellum in each of 39 mice plotted versus the ages of those mice. We can see an apparent dichotomy that is unrelated to gender. The right plot shows a histogram of the regression coefficients for the latent variable. The long tail of the histogram indicates that a large number of genes (about 100) are related to the dichotomy.

5. Latent Variables for the AGEMAP Mice

Now that we have seen how rotation tests work in simulations, we apply them to the data described in the beginning of the paper. Recall that in the AGEMAP data set there are 16 tissue types and 32–40 mice per tissue with known age and sex. Here we will see patterns that certainly appear unlikely to be artifacts. Then we verify them by the rotation test.

We fit 16 regression models of gene activation on age and sex with one latent variable, one for each tissue. The result is that for each tissue type from $k = 1, \ldots, 16$, we have an estimated latent variables vector $U^{(k)} \in \mathbb{R}^{n_k}$.

The latent variables for tissue 2 (the cerebellum) have a striking pattern. There is one value $\widehat{U}_i^{(2)}$ for each of $n_2 = 39$ mice for which a cerebellum array was available. Figure 3 shows that latent variable plotted versus age and with plot symbols encoding the sex of the mouse. It is clear that the mice are split into two different groups, one with a high value of the latent variable and one low.

Often when one sees two distinct groups in microarray data, they correspond to male versus female samples, and certain genes that are sex related, such as on those on the Y chromosome in males or Xist genes that silence a second X chromosome for females. That cannot be the case here because the estimated latent variable is orthogonal to both the sex and age variables by construction, meaning the sum of its coefficients over male samples must equal the negative of the sum over females.

There are high and low values for the latent variable for the cerebellum. The second panel of Figure 4 shows the histogram of these latent values. It is clearly bimodal. The other 15 panels in Figure 4 show the corresponding histograms for the other 15 tissues.

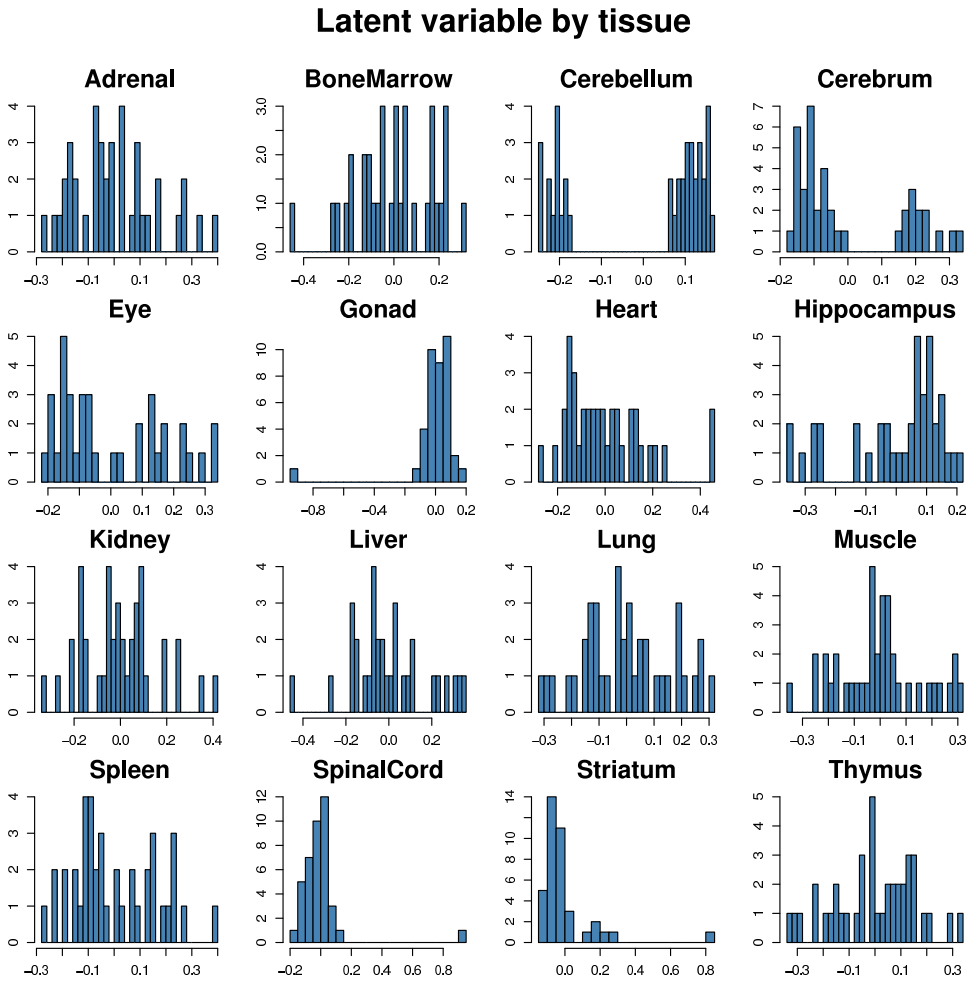

Figure 4: LATENT VARIABLES BY TISSUE. This figure shows histograms of the latent variables found in microarray data from 16 mouse tissues. In each histogram the latent variable values from up to 40 mice are given.

Some of the other histograms have interesting and interpretable structure too. The histograms for spinal tissue, gonad and striatum all show outliers. The biggest latent effect in these tissues is that the expression of one mouse was quite different from the other mice and that difference is reflected in a large number of genes. It is not simply one unusual animal. Three different mice were the outliers in the three different tissues.

The histogram for the cerebrum shows an apparent dichotomy, similar to but less pronounced than the one for the cerebellum. For both of these tissues, the latent variable is splitting the mice into two groups. Both dichotomies are somewhat imbalanced with one group roughly twice as large as the other. Such an effect would be explainable if the same latent factor were affecting both of these brain tissues. Figure 5 plots the estimated latent variable from the cerebrum versus that for the cerebellum. There is one point for each of the 39 mice in which both tissues were measured.

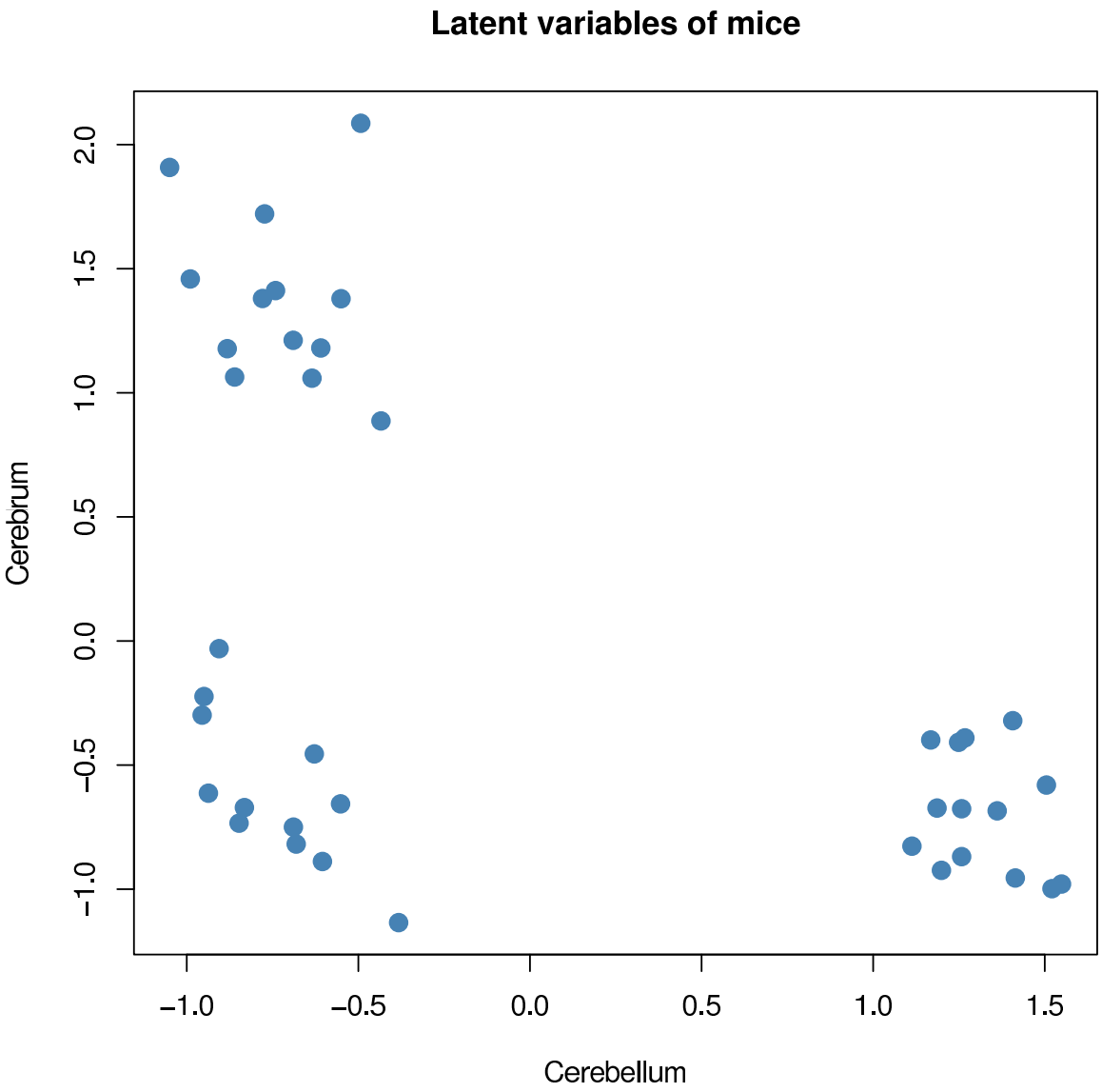

Figure 5: LATENT VARIABLES OF MICE. This figure plots the latent variable from the cerebrum versus that from the cerebellum for the 39 mice for which both arrays were available. There appear to be three kinds of mice in this population.

Figure 5 shows that the apparent dichotomy in the cerebrum is not the same as the one in the cerebellum. The pattern is not a simple double dichotomy either. Rather there appears to be a trichotomy. It is visually striking that there are no mice in the upper right hand corner of Figure 5. The counts of the four corners of Figure 5 are set out in Table 1.

About one third of the mice have the rare cerebellum type, one third have the rare cerebrum type and the remaining mice the common form for both tissues. Were the types independent we would expect about one ninth of the mice to be rare for both. A p value based on Fisher's exact test is

Count	Common cerebellum	Rare cerebellum
Rare cerebrum	13	0
Common cerebrum	12	14

Table 1: Counts of mice in the corners of Figure 5.

0.00094. While we can't be sure that the next mouse won't be the rare type for both brain parts, the failure to observe one here is statistically significant.

Although we don't have laboratory notes to identify the meaning of these groupings, the fact that the joint behavior of two dichotomies from different tissues forms such an interesting pattern lends additional support to the rotation results.

We applied rotation tests for all 16 tissues in the AGEMAP data set using two different measures of non-normality.

We demonstrate the calculation of p-values for the spleen and cerebellum data in Figure 6. The rotation distributions for the other 15 tissues have approximately the same shape, so we do not display them here. Instead, we summarize the results in Table 2. The estimated latent variables for the cerebellum, cerebrum, and eye tissues exhibited dichotomies. This shows up in significantly high values of T_{L^1} and T_{EPP}. The gonad, spinal cord, and striatum latent variables have clear outliers, which manifest as a significantly low values of T_{L^1}, and a significantly high values of T_{EPP}. The spleen latent variable potentially has an outlier at age 5 months, and is found to be marginally significant according to T_{EPP}, and not significant according to T_{L^1}. The discrepancy is because of the two- versus one-tailed p-value.

The only case where T_{L^1} and T_{EPP} give drastically different results is with the latent variable estimated from the hippocampus data. Using T_{L^1}, the variable is nowhere near significant ($\hat{p} = 0.586$), but using T_{EPP}, the variable is unquestionably significant ($\hat{p} < 0.001$). T_{EPP} finds the skewed histogram interesting, while T_{L^1} does not. A possible explanation for why the latent variable is insignificant according to T_{L^1} is this: T_{L^1} is simultaneously measuring presence of outliers and presence of clumping. Outliers correspond to low values of T_{L^1}, and clumping corresponds to high values of T_{L^1}. In the hippocampus data, we see both outliers *and* clumping. The two features "cancel out", giving a moderate value of T_{L^1}. T_{EPP}, on the other hand, does not distinguish between the different kinds of non-Gaussianity. The two features act in tandem to give a high value of the test statistic.

6. Conclusions

We find that it is possible to test for latent variables in correlated Gaussian noise by a rotation test using a projection pursuit index applied to the components of the first singular vector, instead of the usual test based on the size of the largest singular value. This test detects the lack of rotational invariance of the matrix of errors. The rotations must be done orthogonally to the regression variables. Testing for one latent variable is theoretically justified and reliable. Testing for additional terms is possible, but can give somewhat liberal p-values if the signal strength is too weak.

For microarray data, a normal distribution is often a very reasonable model. Some researchers apply transformations for the explicit purpose of making the data more normally distributed. For data that is not close to normally distributed a strategy of looking for latent variables by measuring how non-Gaussian they are is not recommended. It might uncover eigenvectors with especially

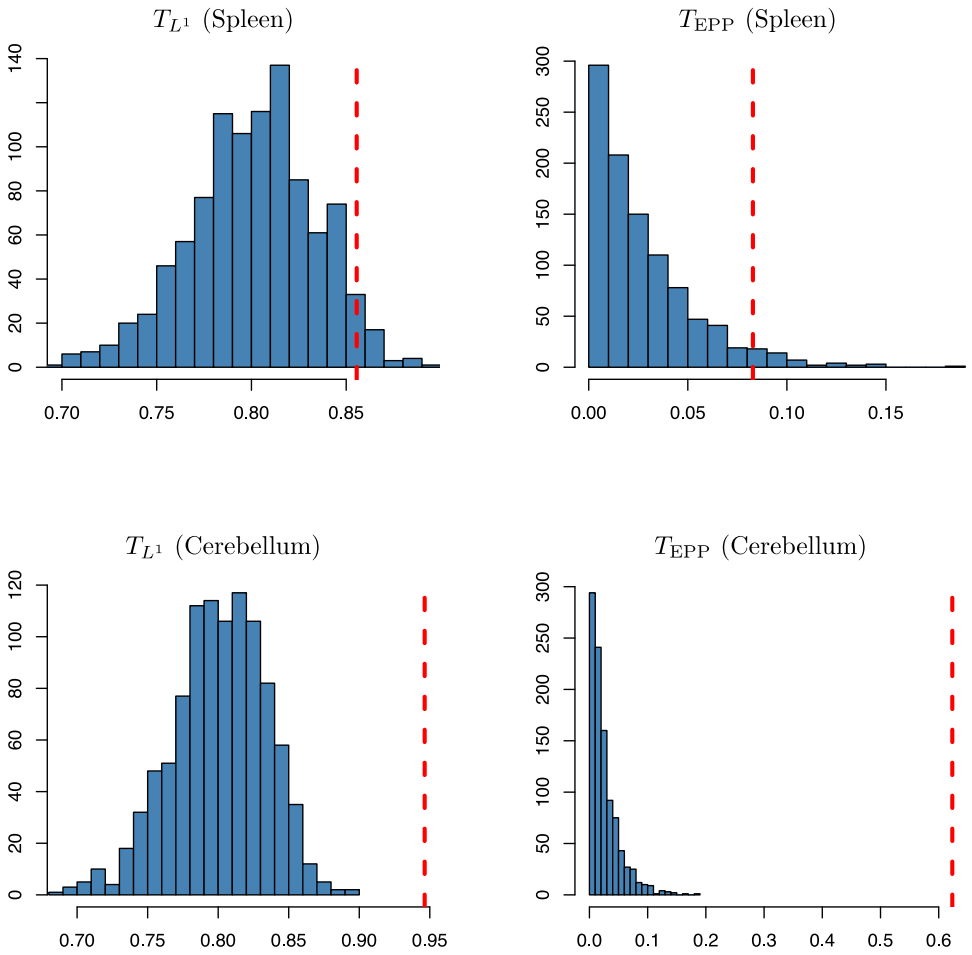

Figure 6: This figure shows histograms of 1000 realizations of the test statistics T_{L^1} and T_{EPP} after applying random rotations to the estimated latent variable. The top row comes from the spleen data, and the bottom row from the cerebellum. The dashed line shows the value of the test statistic at the observed data. In the top-left plot, the area in the right tail is 0.049, giving a two-sided p-value of 0.098. In the top-right plot, the tail area and the one-sided p-value are equal to 0.045. Formal descriptions of the p-values can be found in Equations (6) and (7). The latent variable is found to be barely significant at the 0.05 level according to T_{EPP}, but not according to T_{L^1}. In the bottom plots, the observed data falls at the extreme of the rotation histograms, and is found to be strongly significant in both cases.

non-Gaussian components but their interpretation is more difficult without a Gaussian background to compare them to.

Our original interest was to see if thousands of genes could be used to define a genomic "true age" of a sample of mouse tissue as a latent variable in the residuals from a regression that did not include age. It turned out that the dominant latent variable bore no resemblance to chronological

		T_{L^1}		T_{EPP}	
Tissue	R^2	T	$\hat{p}$	T	$\hat{p}$
Adrenal	0.109	0.7986	0.816	0.0204	0.460
Bone Marrow	0.227	0.8098	0.988	0.0035	0.931
Cerebellum	0.593	0.9464	0.002^{**}	0.6233	0.001^{**}
Cerebrum	0.520	0.8961	0.002^{**}	0.4344	0.001^{**}
Eye	0.165	0.8927	0.004^{*}	0.2343	0.001^{**}
Gonad	0.197	0.4788	0.002^{**}	0.3899	0.001^{**}
Heart	0.287	0.7866	0.562	0.0611	0.117
Hippocampus	0.208	0.8222	0.586	0.2281	0.001^{**}
Kidney	0.247	0.7702	0.320	0.0145	0.598
Liver	0.311	0.7758	0.384	0.0608	0.170
Lung	0.169	0.8085	0.944	0.0130	0.625
Muscle	0.318	0.7601	0.172	0.0583	0.121
Spleen	0.328	0.8555	0.098	0.0829	0.045^{*}
Spinal Cord	0.309	0.4641	0.002^{**}	0.3864	0.001^{**}
Striatum	0.319	0.5851	0.002^{**}	0.4020	0.001^{**}
Thymus	0.266	0.8125	0.838	0.0264	0.386

Table 2: Test statistics and p-values from the rotation tests applied to the AGEMAP data. In all cases, 1000 random rotation were used to construct the a reference histogram, and approximate p-values were estimated. Significant results at the 0.05 level are marked with a single asterisk. In instances where the observed test statistic was at the extreme end of the histogram, we have marked the p-values with two asterisks. In the second column of the table, we indicate how much of the residual is explained by the latent term. We can see that high R^2 does not necessarily indicate significance according to the rotation test. A Bonferroni correction for multiple testing would multiply the p-values by 32 and would find most of the same latent variables significant.

age. We never uncovered a biological explanation for the dichotomies and other latent variables that we saw. But, the rotation tests confirm that these striking anomalies would not arise from correlated Gaussian noise. Several of the tissues did not have apparent latent variables. Accordingly results like those in Table 2 help one focus on where to search for physical causes underlying apparent latent variables.

It may happen that a latent variable is statistically significant when judged by a rotation test but only explains a negligible amount of the response variation. This seems unlikely to happen in practice and did not happen for the AGEMAP data, according to the R^2 column in Table 2. But, when it does happen one can always declare the variable statistically but not practically significant.

It is natural to ask if rotation tests extend to nonlinear models. Our method is strongly geared to linear models because of the way we construct our rotations, so we see no straightforward extension.

Acknowledgments

We thank Stuart Kim, Jacob Zahn, and four anonymous referees for their comments. This work was supported by the NSF under grants DMS-0906056 and DMS-0604939.

References

J. Baik and J.W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

J. Crossa and P. L. Cornelius. Linear-bilinear models for the analysis of genotype-environment interaction data. In M. S. Kang, editor, *Quantitative Genetics, Genomics and Plant Breeding in the 21st Century, an International Symposium*, pages 305–322, Wallingford UK, 2002. CAB International.

C. T. dos S Dias and W. J. Krzanowski. Model Selection and Cross Validation in Additive Main Effect and Multiplicative Interaction Models. *Crop Science*, 43(3):865–873, 2003.

B. Efron. Are a set of microarrays independent of each other? *Annals of Applied Statistics*, 3(3): 922–942, 2009.

N. El Karoui. Tracy-Widom limit for the largest eigenvalue of a large class of complex Wishart matrices. *Annals of Probability*, 35(2):663–714, 2007.

J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82: 249–266, 1987.

K. R. Gabriel. Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society, B*, 40(2):186–196, 1978.

H. F. Gollob. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33:73–115, 1968.

R. M. Heiberger. Algorithm AS 127: Generation of random orthogonal matrices. *Applied Statistics*, 27(2):199–206, 1978.

A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. J. Wiley New York, 2001.

Ø. Langsrud. Rotation tests. *Statistics and Computing*, 15:53–60, 2005.

E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

J. Mandel. A new analysis of variance model for non-additive data. *Technometrics*, 13:1–18, 1971.

B. Nadler. personal communication. Discussions at AIM workshop on Random Matrices and Higher Dimensional Inference, April 2007.

A. Onatski. Asymptotics of the principal components estimator of large factor models with weak factors and i.i.d. Gaussian noise. Available at http://www.columbia.edu/˜ao2027/inference45.pdf, August 2007.

A. B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society, Series B*, 67:411–426, 2005.

A. B. Owen and P. O. Perry. Bi-cross-validation of the SVD and the non-negative matrix factorization. *Annals of applied statistics*, 3(2):564–594, 2009.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007a.

D. Paul. Spiked covariance: some extensions. Unpublished technical note, June 2007b.

N.R. Rao and A. Edelman. Sample eigenvalue based detection of high dimensional signals in white noise using relatively few samples. *IEEE Transactions on Signal Processing*, 56(7 Part 1):2625–2638, 2008.

R. D. Snee. Nonadditivity in a Two-Way Classification: Is It Interaction or Nonhomogeneous Variance? *Journal of the American Statistical Association*, 77(379):515–519, 1982.

L. K. Southworth, S. K. Kim, and A. B. Owen. Properties of balanced permutations. *Journal of Computational Biology*, 16(4):625–638, 2009.

R. W. M. Wedderburn. Random rotations and multivariate normal simulation. Technical report, Rothamsted Experimental Station, 1975.

J.M. Zahn, S. Poosala, A.B. Owen, D.K. Ingram, A. Lustig, A. Carter, A.T. Weeraratna, D.D. Taub, M. Gorospe, K. Mazan-Mamczarz, et al. AGEMAP: A Gene Expression Database for Aging in Mice. *PLoS Genet*, 3(11):e201, 2007.