Multiple hypothesis testing, adjusting for latent variables

Yunting Sun
Stanford University

Nancy Zhang
Stanford University

Art B. Owen
Stanford University

May 2011

Abstract

In high throughput settings we inspect a great many candidate variables (e.g. genes) searching for associations with a primary variable (e.g. a phenotype). High throughput hypothesis testing can be made difficult by the presence of systemic effects and other latent variables. It is well known that those variables alter the level of tests and induce correlations between tests. It is less well known that dependencies can change the relative ordering of significance levels among hypotheses. Poor rankings lead to wasteful and ineffective followup studies. The problem becomes acute for latent variables that are correlated with the primary variable. We propose a two stage analysis to counter the effects of latent variables on the ranking of hypotheses. Our method, called LEAPP, statistically isolates the latent variables from the primary one. In simulations it gives better ordering of hypotheses than competing methods such as SVA and EIGENSTRAT. For an illustration, we turn to data from the AGEMAP study relating gene expression to age for 16 tissues in the mouse. LEAPP generates rankings with greater consistency across tissues than the rankings attained by the other methods.

1 Introduction

There has been considerable progress in multiple testing methods for high throughput applications. A common example, coming from biology, is testing which of N genes' expression levels correlate significantly with a scalar variable, which we'll call the primary variable. The primary variable may be an experimentally applied treatment or it may be a covariate such as a phenotype. We will use the gene expression example for concreteness, although it is just one of many instances of this problem.

High throughput experiments may involve thousands or even millions of hypotheses. Because N is so large, serious problems of multiplicity arise. For independent tests, methods based on the false discovery rate (Dudoit and van der Laan, 2008) have been very successful. Attention has turned more recently to dependent tests (Efron, 2010).

One prominent cause of dependency among test statistics is the presence of latent variables. For example, in microarray-based experiments, it is well known that samples processed in the same batch are correlated. Batch, technician, and other sources of variation in sample preparation can be modeled by latent variables. Another example comes from genetic association studies, where differences in ancestral history among subjects can lead to false or inaccurate associations. Price et al. (2006) used principal components to extract and correct for ancestral history, in effect modeling the genetic background of the subjects as latent variables. A third example comes from copy number data, where local trends along the genome cause false positive copy number calls (Olshen et al., 2004). Diskin et al. (2008) conducted experiments showing that these local trends correlate with the variation of GC content along the genome, and are caused by differences in the quantity and handling of DNA. These laboratory effects are hard to measure, but can be quantified using a latent variable model. In this paper, we consider latent variables that might even be correlated with the primary variable.

When the primary variable is an experimentally applied treatment, then problematic latent variables are those that are partially confounded with the treatment. Randomization reduces the effects of such confounding but randomization is not always perfectly applied and batch or other effects may be imbalanced with respect to the treatment (Leek et al., 2010).

These latent variables have some severe consequences. They alter the level of the hypothesis tests and they induce correlations among multiple tests. Another consequence, that we find especially concerning, is that the latent variables may affect the rank ordering among the N p-values. When high throughput methods are used to identify candidates for further followup it is important that the highly ranked items contain as many non-null cases as possible.

Our approach to this problem uses a rotated model in which we separate the latent variables from the primary variable. We do this by creating two data sets, one in which both primary and latent variables are present and one in which the primary variables are absent. We use the latter data set to estimate the latent variables and then substitute their estimates into the former. Since each gene has its own effect size in relation to the primary variable, the former model is supersaturated. We conduct inference under the setting where the parameter vector relating the genes to the primary variable is sparse, as is commonly assumed in multiple testing situations. Each non-null hypotheses behaves as an additive outlier, and we then apply an outlier detection method from She and Owen (2011) to find them. We call the method LEAPP, for *l*atent *e*ffect *a*djustment after *p*rimary *p*rojection.

Section 2 presents our notation and introduces LEAPP along with several other related models, including SVA (Leek and Storey, 2008) and EIGENSTRAT (Price et al., 2006), to which we make comparisons. Section 3 shows via simulation that LEAPP generates better rankings of the non-null hypotheses than one would get by either ignoring the latent variables, by SVA, or by EIGENSTRAT. EIGENSTRAT estimates the latent variables (by principal components) without first adjusting for the primary variable. LEAPP outperforms it when the latent variable is weaker than the primary. EIGENSTRAT does well in simulations with weak primary variables, which matches the setting that motivated it. Still it is interesting to learn that it does not extend well to problems with strong primary variables. SVA estimates the primary variable's coefficients without first adjusting for correlation between the primary and latent variables. LEAPP outperforms it when the latent and primary variables are correlated.

Section 4 compares the methods on the AGEMAP data of Zahn et al. (2007). The primary variable there is age. While we don't know the truly non-null genes for this problem, we have a proxy. The data set has 16 subsets, each from one tissue type. We find that LEAPP gives gene lists with much greater overlap among tissues than the gene lists achieved by the other methods. Section 5 contains some theoretical results about the method: The specific rotation matrix used does not affect our answer. For the case of one latent variable and no covariates, LEAPP consistently estimates the latent structure. We also get a bound for the sum of squared coefficient errors when the effects are sparse. Our conclusions are in Section 6. Some background material are given in an appendix.

2 Notation and models

In this section we describe the data available and introduce the parameters and latent variables that arise. Then we describe our LEAPP proposal which is based on a series of reductions from a heteroscedastic multivariate regression including latent factors to a single linear regression problem with additive outliers and known error variance. We also describe EIGENSTRAT and SVA, to which we make comparisons, and then survey several other published methods for this problem.

2.1 Data, parameters, latent variables and tests

The data we observe are a response matrix $Y \in \mathbb{R}^{N\times n}$ and a variable of interest $g \in \mathbb{R}^n$, which we call the primary variable. In an expression problem Y_{ij} is the expression level of gene i for subject j. Very often the primary variable g is a group variable taking just two values, such as ±1 for a binary phenotype, then linearly transformed to have mean 0 and norm 1. The quantity g_j can also be a more general scalar, such as the age of subject j.

We are interested to know which genes, if any, are linearly associated with the variable g. We capture this linear association through the $N \times n$ matrix γg^{T} where γ is a vector of N coefficients. When most genes are not related to g, then γ is sparse.

Often there are covariates X other than g that we should adjust for. The covariate term is βX^{T} where β contains coefficients. The latent variables that cause tests to be mutually correlated are assumed to take an outer product form UV^{T}. Neither U nor V is observed. Finally, there is observational noise with a variance that is allowed to be different for each gene, but assumed to be constant over subjects.

The full data model is

$$Y = \gamma g^{\mathsf{T}} + \beta X^{\mathsf{T}} + UV^{\mathsf{T}} + \Sigma E \tag{1}$$

for variables

$Y \in \mathbb{R}^{N\times n}$	response values
$g \in \mathbb{R}^{n\times 1}$	primary predictor, e.g. treatment, with $g^{\mathsf{T}}g = 1$
$\gamma \in \mathbb{R}^{N\times 1}$	primary parameter, possibly sparse
$X \in \mathbb{R}^{n\times s}$	s covariates (e.g. sex) per subject
$\beta \in \mathbb{R}^{N\times s}$	s coefficients, including per gene intercepts
$U \in \mathbb{R}^{N\times k}$	latent, nonrandom rows (e.g. genes)
$V \in \mathbb{R}^{n\times k}$	latent, independent rows (e.g. subjects)
$E \sim \mathcal{N}(0, I_N \otimes I_n)$	noise, and,
$\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_N)$	standard deviations,

with dimensions

n	number of arrays/subjects
$N \gg n$	number of genes
$s \ll n$	number of covariates, and
$k \ge 1$	latent dimension, often one.

After adjusting for X, the genes are correlated through the action of the latent portion UV^{T} of the model. They may have unequal variances, through both Σ and U. We adopt the normalization $\mathbb{E}(V^{\mathsf{T}}V) = I_k$. It is possible to generalize the model to have a primary variable g of dimension $r \ge 1$ but we focus on the case with $r = 1$.

We pay special attention to the case of $k = 1$ latent variable. When $k = 1$, the dependence between the variable g of interest and the latent variable V can be summarized by a single correlation coefficient $\rho = g^{\mathsf{T}}V/\sqrt{V^{\mathsf{T}}V}$.

Writing (1) in terms of indices yields

$$Y_{ij} = \gamma_i g_j + \beta_i^{\mathsf{T}} X_j + U_i^{\mathsf{T}} V_j + \sigma_i \varepsilon_{ij}, \quad 1 \le i \le N, \quad 1 \le j \le n. \tag{2}$$

Here β_i and U_i are the i'th rows of β and U respectively, while X_j and V_j are the j'th rows of X and V, σ_i is the i'th diagonal element of Σ and ε_{ij} is the ij element of E.

2.2 Model reduction

Our LEAPP proposal is based on a series of reductions described here. First we choose an orthogonal matrix $O \in \mathbb{R}^{n\times n}$ such that $g^\mathsf{T} O^\mathsf{T} = (\eta, 0, 0, \dots, 0) \in \mathbb{R}^{1\times n}$ where $\eta = \|g\| > 0$. Without loss of generality, we assume that the primary predictor has been scaled so that $\eta = 1$. A convenient choice for O is the Householder matrix $O = I_n - 2\kappa\kappa^\mathsf{T}$ where $\kappa = (g - e_1)/\|g - e_1\|_2$.

Using O we construct the ***rotated model***

$$\begin{aligned} \widetilde{Y} \equiv YO^\mathsf{T} &= \gamma g^\mathsf{T} O^\mathsf{T} + \beta X^\mathsf{T} O^\mathsf{T} + UV^\mathsf{T} O^\mathsf{T} + \Sigma E O^\mathsf{T} \qquad (3)\\ &\equiv \gamma \widetilde{g}^\mathsf{T} + \beta \widetilde{X}^\mathsf{T} + U\widetilde{V}^\mathsf{T} + \Sigma\widetilde{E}, \qquad (4) \end{aligned}$$

where $\widetilde{g}$, $\widetilde{X}$, $\widetilde{V}$ and $\widetilde{E}$ are rotated versions of their counterparts without the tilde. Notice that $\widetilde{E} = EO^\mathsf{T} \overset{d}{=} E$, because $E \sim \mathcal{N}(0, I_N \otimes I_n)$. By construction, $\widetilde{g}^\mathsf{T} = (1, 0, \dots, 0)$. Therefore the model for $\widetilde{Y}_{ij}$ is different depending on whether $j = 1$ or $j \neq 1$:

$$\begin{aligned} \widetilde{Y}_{i1} &= \beta_i^\mathsf{T} \widetilde{X}_1 + U_i^\mathsf{T} \widetilde{V}_1 + \gamma_i + \sigma_i \varepsilon_{i1}, \qquad \text{and} \qquad (5)\\ \widetilde{Y}_{ij} &= \beta_i^\mathsf{T} \widetilde{X}_j + U_i^\mathsf{T} \widetilde{V}_j + \qquad \sigma_i \varepsilon_{ij}, \qquad j = 2, \dots, n. \qquad (6) \end{aligned}$$

The rotated model concentrates the primary coefficients γ_i in the first column of $\widetilde{Y}$. Our approach is to base tests and estimates of γ_i on equation (5). We need to substitute estimates for unknown quantities σ_i, β_i and U_i in (5). The estimates come from the model in equation (6).

This rotated approach has some practical advantages: First, we do not need to iterate between applying equations (5) and (6). Instead we use (6) once to estimate unknowns U, σ and β and then use (5) once to judge γ_i. Second, the last $n-1$ columns of $\widetilde{Y}$, and hence estimates $\hat{\sigma}$, $\hat{\beta}$, and $\widehat{U}$, are statistically independent of the first column. Third, problems (5) and (6) closely match settings for which there are usable methods as described next.

Assume that estimates $\hat{\sigma}_i$, $\hat{U}_i$, and $\hat{\beta}_i$ from (6) are given. We may then write (5) as

$$\widetilde{Y}_{i1} - \hat{\beta}_i^\mathsf{T} \widetilde{X}_1 = \widehat{U}_i^\mathsf{T} \widetilde{V}_1 + \gamma_i + \hat{\sigma}_i \varepsilon_{i1}. \qquad (7)$$

The right hand side of equation (7) is a regression with measurement errors in the predictors $\widehat{U}_i$, mean-shift outliers γ_i and unequal error variances. We use the Θ–IPOD algorithm of She and Owen (2011), adjusting it to handle unequal σ_i, to estimate γ_i. See Appendix A.1.

We don't know β_i, U_i and σ_i, but we may estimate them from the data for $j \geq 2$. In the process we will also estimate $\widetilde{V}_j$ for $j \geq 2$. Let $\overline{X}$, $\overline{Y}$, $\overline{V}$ and $\overline{E}$ be the last $n-1$ columns of $\widetilde{X}$, $\widetilde{Y}$, $\widetilde{V}$ and $\widetilde{E}$, respectively. Then the model for the last $n-1$ columns of the data is

$$\overline{Y} = \beta \overline{X}^\mathsf{T} + U\overline{V}^\mathsf{T} + \Sigma\overline{E}.$$

We adopt an iterative approach, where to update $\hat{\sigma}_i$ from the other variables, we take

$$\hat{\sigma}_i^2 = \frac{1}{n-1}\sum_{j=2}^{n}(\widetilde{Y}_{ij} - \hat{\beta}_i^{\mathsf{T}}\widetilde{X}_j - \widehat{U}_i^{\mathsf{T}}\widehat{V}_j)^2. \tag{8}$$

Given $\hat{\sigma}_i$ we write the model as

$$Y^* = \beta^*\overline{X}^{\mathsf{T}} + U^*\overline{V}^{\mathsf{T}} + \overline{E} \tag{9}$$

where $Y^* = \widehat{\Sigma}^{-1}\overline{Y}$, $\beta^* = \widehat{\Sigma}^{-1}\beta$, and $U^* = \widehat{\Sigma}^{-1}U$. We use the criss-cross regression of Gabriel and Zamir (1979) (see Appendix A.2) to estimate β^*, U^* and $\overline{V}$ and then multiply those estimates by $\widehat{\Sigma}$ to get $\hat{\beta}$ and $\widehat{U}$. To start the iteration, we set $\hat{\sigma}_i = 1$, for $i = 1, \dots, N$.

To fit our model, we need to choose the rank k for the latent term UV^{T}. We follow Leek and Storey (2008) in using the method of Buja and Eyuboglu (1992), as described in Section 2.3.

Given estimates $\hat{\sigma}_i$ we apply Θ-IPOD to the regression

$$\frac{\widetilde{Y}_{i1} - \hat{\beta}_i^{\mathsf{T}}\widetilde{X}_1}{\hat{\sigma}_i} = \frac{\widehat{U}_i^{\mathsf{T}}}{\hat{\sigma}_i}\widetilde{V}_1 + \frac{\gamma_i}{\hat{\sigma}_i} + \frac{\sigma_i}{\hat{\sigma}_i}\varepsilon_{i1}. \tag{10}$$

We write the final form of this model as

$$\underline{Y}_i = \underline{U}_i^{\mathsf{T}}\widetilde{V}_1 + \underline{\gamma}_i + \underline{\varepsilon}_i, \tag{11}$$

for response $\underline{Y}_i = (\widetilde{Y}_{i1} - \hat{\beta}_i^{\mathsf{T}}\widetilde{X}_1)/\hat{\sigma}_i$, predictors $\underline{U}_i = \widehat{U}_i/\hat{\sigma}_i$, unknown coefficients $\widetilde{V}_1$, additive outliers $\underline{\gamma}_i = \gamma_i/\hat{\sigma}_i$ and errors $\underline{\varepsilon}_i = \varepsilon_{i1}\sigma_i/\hat{\sigma}_i$.

Our statistic for testing $H_{i0} : \gamma_i = 0$ is

$$T_i = \frac{\underline{Y}_i - \underline{U}_i^{\mathsf{T}}\widehat{V}_1}{\hat{\tau}}, \tag{12}$$

where $\widehat{V}_1$ is the Θ-IPOD estimate of $\widetilde{V}_1$ and $\hat{\tau}$ is an estimate of the error variance from (11). We use the median absolute deviation from the median (MAD) to estimate $\hat{\tau}$. For p-values we use use $\Pr(|Z| \geq |T_i|)$ where $Z \sim \mathcal{N}(0,1)$. Candidate hypotheses are ranked from most interesting to least interesting by taking the p values from smallest to largest.

We have emphasized the setting in which γ is a sparse vector. When γ is not a sparse vector, then its large components may not be flagged as outliers because the MAD estimate of τ would be inflated due to contamination by γ. In this case however we can fall back on a simpler approach to estimating τ. The error $\underline{\varepsilon}_i$ has variance $\mathbb{E}(\sigma_i^2/\hat{\sigma}_i^2)$. This variance differs from unity only because of estimation errors in $\hat{\sigma}_i$. We can then use $\tau^2 = 1$. We can account for fitting s regression parameters to the $n-1$ samples in each row of $\overline{Y}$ by taking $\tau^2 = \mathbb{E}((n-1-s)/\chi^2_{n-1-s}) = (n-s-1)/(n-s-3)$. A further approximate adjustment for estimating k latent vectors is to take $\tau^2 = (n-s-k-1)/(n-s-k-3)$. This estimate of τ can be used in (12) for ranking of hypotheses if γ is not suspected to be sparse.

2.3 SVA

We compare our method to the surrogate variable analysis (SVA) method of Leek and Storey (2008). Their iteratively reweighted surrogate variable analysis algorithm adjusts for latent variables before doing a regression. But it does not isolate them.

A full and precise description of SVA appears in the supplementary information and online software for Leek and Storey (2008). Here we present a brief outline. In our notation, their model is

$$Y = \gamma g^{\mathsf{T}} + UV^{\mathsf{T}} + \Sigma E.$$

The SVA algorithm uses iteratively reweighted singular value decompositions (SVDs) to estimate U, V and γ. The weights are empirical Bayes estimates of $\Pr(\gamma_i = 0, U_i \neq 0 \mid Y, g, V)$ from Storey et al. (2005). Their method seeks to remove the primary term γg^{T} by downweighting rows with $\gamma_i \neq 0$. Our method creates columns that are free of the primary variable by rotation.

The SVA iteration is as follows. First, they fit a linear model without any latent variables, getting estimates $\hat{\gamma}$ and the residual $R = Y - \hat{\gamma} g^{\mathsf{T}}$. Second, they apply the simulation method of Buja and Eyuboglu (1992) to R to estimate the number k of factors, and then take the top k right eigenvectors of R as the initial estimator $\widehat{V}$. Third, they form the empirical Bayes estimates $w_i = \Pr(\gamma_i = 0, U_i \neq 0 \mid Y, g, \widehat{V})$ from Storey et al. (2005). Fourth, based on those weights, they perform a weighted singular value decomposition of the original data matrix Y, where row i is weighted by w_i. The weighted SVD gives them an updated estimator $\widehat{V}$. They repeat steps 3 and 4, revising the weights w_i and then the matrix $\widehat{V}$, until $\widehat{V}$ converges. They perform significance analysis on γ through the multivariate linear regression model

$$Y = \gamma g^{\mathsf{T}} + U\widehat{V} + \Sigma E,$$

where $\widehat{V}$ is treated as known covariates to adjust for the primary effect g.

To estimate the number k of factors in the SVD, they use a simulation method of Buja and Eyuboglu (1992). That algorithm uses Monte Carlo sampling to adjust for the well known problem that the largest singular value in a sample covariance matrix is positively biased. That method has two parameters: the number of simulations employed and a significance threshold. The default significance threshold was 0.1 and the default uses 20 permutations.

2.4 EIGENSTRAT

EIGENSTRAT (Price et al., 2006) was developed to control for differences in ancestry in genetic association studies, where the matrix Y represent the alleles carried by the subjects at the genetic markers (e.g. $Y_{ij} \in \{0, 1, 2\}$ counts the number of one of the alleles). The primary variable can be case versus control, disease status, or other clinical traits.

In our notation, they begin with a principal components analysis approximating Y by $\widehat{U}\widehat{V}^{\mathsf{T}}$ for $\widehat{U} \in \mathbb{R}^{N \times k}$ and $\widehat{V} \in \mathbb{R}^{n \times k}$. Then for $i = 1, \ldots, N$ they test whether $Y_{i,1:n}$ is significantly related to g in a regression including the k columns of $\widehat{V}$, or equivalently whether the partial correlation of $Y_{i,1:n}$ on g adjusted for $\widehat{V}$, is significant. Although the data are discrete and the method resembles one for Gaussian data, the results still clearly obtain latent variables showing a natural connection to the geographical region of the subjects' ancestors.

EIGENSTRAT has an apparent weakness. If the signal γg^{T} is large then its presence will corrupt the estimates of $\widehat{U}$ and $\widehat{V}$. The estimate $\widehat{V}$ will be correlated with the effect g that we are trying to estimate a coefficient for. Indeed, we find in our simulations of Section 3, that EIGENSTRAT performs poorly when the signal is large compared to the latent variable.

EIGENSTRAT also requires the choice of a rank k for the latent term. Price et al. (2006) describe a default choice of $k = 10$. Patterson et al. (2006) apply a spiked covariance model test of Johnstone (2001) using the Tracy-Widom distribution (Tracy and Widom, 1994).

2.5 Other methods

A number of other methods have been proposed for this problem, which we have not included in our numerical comparisons. Here we mention several of them, relating their approaches to the notation of Section 2.1.

Friguet et al. (2009) model their data as $Y = \gamma g^{\mathsf{T}} + UV^{\mathsf{T}} + \Sigma E$. They assume the latent V is normally distributed (independent of E) and that U is nonrandom. They do not assume sparsity for γ. They estimate U, V, γ and Σ by an EM algorithm. They find that using $\widehat{V}$ in an FDR procedure is an improvement compared to a model that does not employ latent variables.

Lucas et al. (2010) take $Y = \beta X^{\mathsf{T}} + UV^{\mathsf{T}} + \Sigma E$ and make extensive use of sparsity priors. They include the primary variable g as one of the columns of X, instead of singling it out as we do. Under their sparsity priors, a coefficient is either 0 or it is $\mathcal{N}(0, \tau^2)$. The probability of a nonzero coefficient is π which in turn has a Beta distribution with a small mean. They apply sparsity priors to the elements of both the coefficient matrix β and the latent variables U. The parameters π and τ are different for each column of β. They use Markov chain Monte Carlo for their inferences.

Allen and Tibshirani (2010) model the data as $Y = \gamma g^{\mathsf{T}} + E$ where $E \sim \mathcal{N}(0, \Sigma \otimes \Gamma)$. That is, the noise covariance is of Kronecker form which models dependence between rows and between columns. Our model has a different variance equal to the sum of two Kronecker matrices, one from UV^{T} and one from ΣE. They estimate Σ and Γ by maximum likelihood with a penalty on the norm of the inverses of Σ and Γ. Their L_1 penalties encourage sparsity in $\widehat{\Sigma}^{-1}$ and $\hat{\Gamma}^{-1}$. They then whiten Y using $\hat{\Gamma}$ and $\widehat{\Sigma}$ and apply false discovery rate methods. They also show that correlations among different columns leads to incorrect estimates of FDR while correlated rows do not much affect the estimates of FDR.

Efron (2007) proposed a method to fit an empirical null to the data to directly account for correlations across arrays. The empirical null method works with estimated Z scores (one per gene) and uses the histogram of those scores to account for the effects of latent variables. This process adjusts significance levels for hypotheses but does not alter their ordering.

Carvalho et al. (2008) consider similar problems but apply a very different formulation. They treat the primary variable (our g) as the response and use the data matrix (our Y) as predictors.

2.6 Rank estimation

The problem of choosing the number k of latent variables is a difficult one that arises for all the methods we used. The Tracy-Widom strategy is derived for the case with $\Sigma = \sigma I_N$ while our motivating applications have heteroscedasticity.

Even for $\Sigma = \sigma I_N$ it is known that the best rank for estimating UV^{T} is not necessarily the true rank. There is a well known threshold strength below which a factor is not detectable and Perry (2009) shows that there is a still higher threshold below which estimating that factor worsens the estimate of UV^{T}. Owen and Perry (2009) present a cross-validatory estimate for the rank k and Perry (2009) shows how to tune it to choose a rank k that gives the best reconstruction as measured by Frobenius norm.

In our numerical comparisons, LEAPP, SVA and EIGENSTRAT were all given the same rank k to use. Sometimes k was fixed at a default value. Other times we used the method of Buja and Eyuboglu (1992).

3 Performance on synthetic data

In this section, we generate data from the model (1) and compare the results from the algorithms to each other, to an oracle which is given the latent variable, and to a raw regression method which makes no attempt to adjust for latent variables.

We choose $s = 0$, omitting the βX^{T} covariate term, so the simulated data satisfy

$$Y = \gamma g^{\mathsf{T}} + UV^{\mathsf{T}} + \Sigma E. \tag{13}$$

Our simulations have $n = 60$ (subjects) and $N = 1000$ (genes). Our primary covariate is a binary treatment vector $g \propto (1, \dots, 1, -1, \dots, -1)$, with equal numbers of 1 and -1, normalized so that $g^{\mathsf{T}} g = 1$.

The vector γ of treatment effects has independent components γ_i taking the values $c > 0$ and 0 with probability $\pi = 0.1$ and $1 - \pi = 0.9$ respectively. We chose c in order to attain specific signal to noise ratios as described below. The matrix Σ is a diagonal with nonzero entries σ_i sampled independently from an inverse gamma distribution: $1/\sigma_i^2 \sim \text{Gamma}(5)/4$. Note that $\mathbb{E}(\sigma_i^2) = 1$.

We use $k = 1$ latent variable that has correlation ρ with g. The latent vector $U = (u_1, \dots, u_N)$ is generated as independent $U(-a, a)$ random variables. We will choose a to obtain specific latent to noise variance ratios. The latent vector V is taken to be $\rho g + \sqrt{1 - \rho^2} W$ where W is uniformly distributed on the set of unit vectors orthogonal to g. That is we sample V so as to have a sample correlation and squared norm that both match their population counterparts.

The model (13) gives Y three components: the signal $\mathcal{S} = \gamma g^{\mathsf{T}}$, the latent structure $\mathcal{L} = UV^{\mathsf{T}}$, and the noise $\mathcal{N} = \Sigma E$. The relative sizes of these components affect the difficulty of the problem. We use Frobenius and spectral norms to describe the size of these matrices.

The noise matrix is constructed so that $\mathbb{E}(\sigma_i^2 \varepsilon_{ij}^2) = \mathbb{E}(\sigma_i^2) = 1$, so that $\mathbb{E}(\|\mathcal{N}\|_F^2) = Nn$. Because the signal and latent matrices have rank 1,

$$\mathbb{E}(\|\mathcal{S}\|_F^2) = \mathbb{E}(\|\mathcal{S}\|_2^2) = \mathbb{E}(\|\gamma\|_2^2) = N\pi c^2, \quad \text{and} \tag{14}$$

$$\mathbb{E}(\|\mathcal{L}\|_F^2) = \mathbb{E}(\|\mathcal{L}\|_2^2) = \mathbb{E}(\|U\|_2^2) = Na^2/3. \tag{15}$$

For our simulation, we specified the ratios

$$\text{SNR} \equiv \pi c^2, \quad \text{and} \quad \text{LNR} \equiv a^2/3$$

and varied them over a wide range. We also use $\text{SLR} = 3\pi c^2/a^2$.

We also varied the level of ρ, the correlation between the latent and primary variables. For each setting of SNR, SLR, LNR and ρ under consideration, we simulated the process 100 times and prepared ROC curves, from the pooled collection of 100,000 predictions.

The methods that we applied are as follows:

true an oracle given UV^{T} which then does regression of $Y - UV^{\mathsf{T}}$ on g,
raw multivariate regression of Y on g ignoring latent variables,
eig EIGENSTRAT of Price et al. (2006),
sva surrogate variable analysis from Leek and Storey (2008), and
lea our proposed LEAPP method.

The ROC curves for one set of conditions are shown in Figure 1. There, the best performance is from the oracle. The next best method is LEAPP. After that comes the raw method making no adjustment for latents, then SVA and finally EIGENSTRAT has the worst performance in this setting.

Because the ROC curves from the simulations have few if any crossings, we can reasonably summarize each one by a single number. We have used the area under the curve (AUC) for a global comparison as well as a precision measure for the quality of the most highly ranked values. That measure is the fraction of truly different genes among the highest ranking H genes. We use $H = 50$.

When $\rho = 0$, EIGENSTRAT, SVA and LEAPP have almost equivalent performance. For $\rho > 0$, the oracle always had the highest AUC and LEAPP was always second. The ordering among the other three methods varied. Sometimes EIGENSTRAT was the best of those three, and other times SVA was best of those three.

Figure 2 shows a heatmap of the improvement in AUC for LEAPP versus SVA. The improvements are greatest when ρ is large. This is reasonable because SVA is not designed to account for correlation between the latent and primary variables. At each correlation level, the greatest differences arise when SNR is small and LNR is about 2.

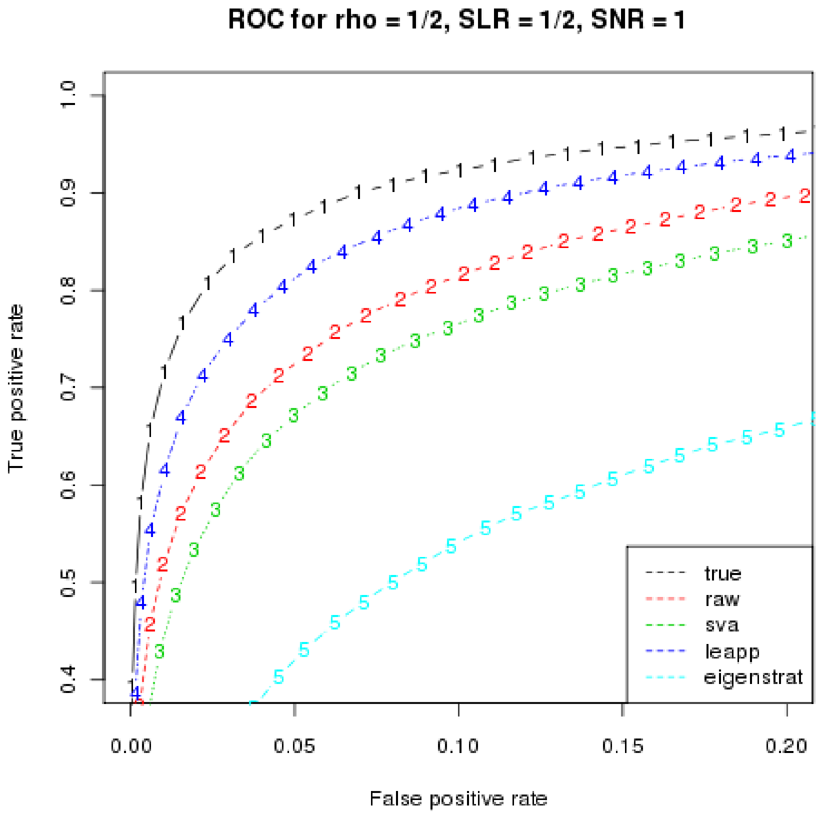

Figure 1: This figure shows the knee of the ROC curves for a simulation with $\rho = 1/2$, SLR= $1/2$ and SNR=1. The best (highest) results are for an oracle that was given the latent variables. The second best are for the proposed LEAPP method. A raw method making no adjustment gives ROCs just barely larger than SVA. EIGENSTRAT did quite poorly in this setting. The relative performance for SVA, EIGENSTRAT and the raw method were different in other settings.

Figure 3 shows the improvement in AUC for LEAPP versus EIGENSTRAT. The improvements are largest when the primary effect is large.

The improvements versus SVA are smaller than those versus EIGENSTRAT. To judge the practical significance of the improvement we repeated some of these simulations for SVA, increasing n until SVA achieved the same AUC that LEAPP did. Sometimes SVA required only 2 more observations (one treatment and one control) to match the AUC of LEAPP. Sometimes it was unable to match the AUC even given double the sample size, that is $n = 120$ observations instead of $n = 60$. Not surprisingly, the advantage of LEAPP is greatest when

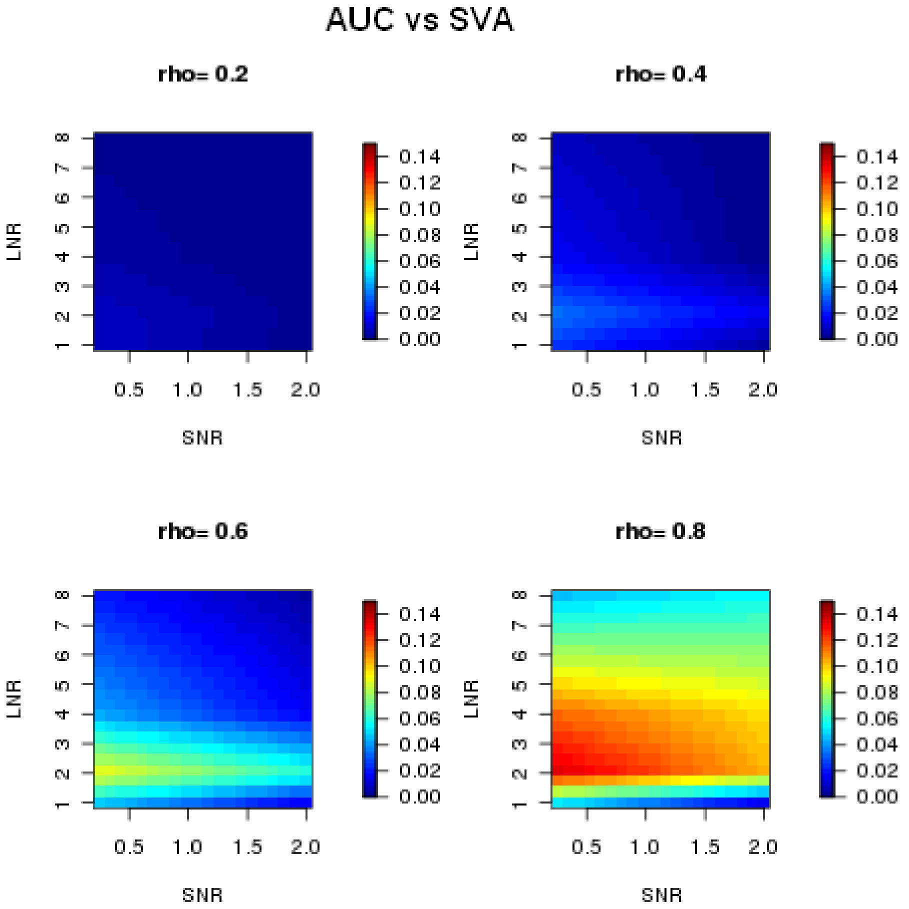

Figure 2: This figure shows the improvement in AUC for LEAPP relative to SVA. Here ρ is the correlation between the primary and latent variables. The signal to noise ratio and latent to noise ratio are described in the text. The color scheme encodes $(\mathrm{AUC}_{\mathbf{lea}} - \mathrm{AUC}_{\mathbf{sva}})/\mathrm{AUC}_{\mathbf{sva}}$.

the latent variable is most strongly correlated with the primary.

Table 1 shows a feature of this problem that we also see in the Figures. The improvement over SVA is quite small when LNR = 0.5. A small enough latent effect becomes undetectable, both methods suffer and there is little difference. Similarly a very large latent effect (LNR = 8) is easy to detect by both methods. The largest differences arise for medium sized latent effects.

High throughput methods are often used to identify candidates for future followup investigation. In that case we value high precision for the most highly ranked hypotheses. Figure 4 shows the improvement of LEAPP over SVA,

Conditions		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
SNR	LNR	n	%	n	%	n	%
2	0.5	66	10	66	10	62	3
2	1	68	13	92	53	120	100
2	2	66	10	74	23	114	90
2	4	62	3	66	10	88	47
2	8	62	3	66	10	72	20
1	0.5	64	7	64	7	62	3
1	1	66	10	90	50	120	100
1	2	64	7	76	27	120	100
1	4	64	7	66	10	90	50
1	8	62	3	66	10	76	27
0.5	0.5	64	7	64	7	62	3
0.5	1	66	10	84	40	120	100
0.5	2	66	10	78	30	110	83
0.5	4	66	10	68	13	88	47
0.5	8	62	3	68	13	72	20

Table 1: This table shows the number of samples required for SVA to attain the same AUC that LEAPP attains with $n = 60$ samples. For example with SNR $= 2$ and LNR $= 0.5$, and $\rho = 0.25$, SVA requires 66 samples or 10% more. The entries of 100% denote settings where the increase needed was $\geq 100\%$.

as measured by precision. Figure 5 shows the improvement of LEAPP over EIGENSTRAT, as measured by precision.

4 AGEMAP data

The AGEMAP study (Zahn et al., 2007) investigated age-related gene expression in mice. Ten mice at each of four age groups were investigated. From these 40 mice, samples were taken of 16 different tissues, resulting in 640 microarray data sets. A small number of those 640 microarrays were missing. From each microarray 8932 probes were sampled. Perry and Owen (2010) found that many of the tissues in this dataset exhibited strong latent variables. Their approach assumed that the latent variables were orthogonal to the treatment.

Our underlying assumption is that aging should have at least mildly consistent results from tissue to tissue. That should in turn show up as overlap in gene lists computed from multiple tissues, whereas a noisier estimation method should tend to have less overlap among tissues.

For any two tissues, we can measure the overlap between their sets of highly

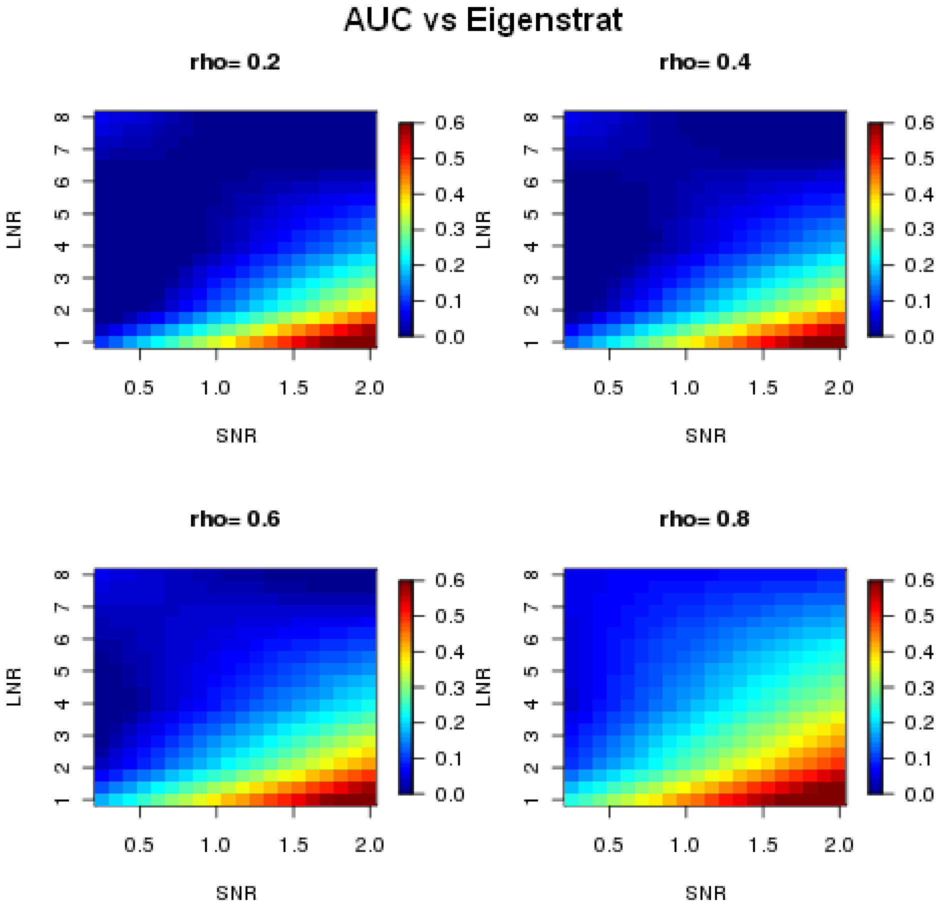

Figure 3: This figure shows the improvement in AUC for LEAPP relative to EIGENSTRAT. The simulation conditions are as described in Figure 2. The color scheme encodes $(\mathrm{AUC}_{\mathbf{rot}} - \mathrm{AUC}_{\mathbf{eig}})/\mathrm{AUC}_{\mathbf{eig}}$.

ranked genes. For two sets A and B, their resemblance (Broder, 1997) is

$$\mathbf{res}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where $|\cdot|$ denoted cardinality. Given two tissues and a significance level α we can compute the resemblance of the genes identified as age-related in the tissues. Resemblance is then a function of α. Plotting the numerator $|A \cap B|$ versus the denominator $|A \cup B|$ as α increases we obtain curves depicting the strength of the overlap.

In our setting with 16 tissues there are $\binom{16}{2} = 120$ resemblances to consider.

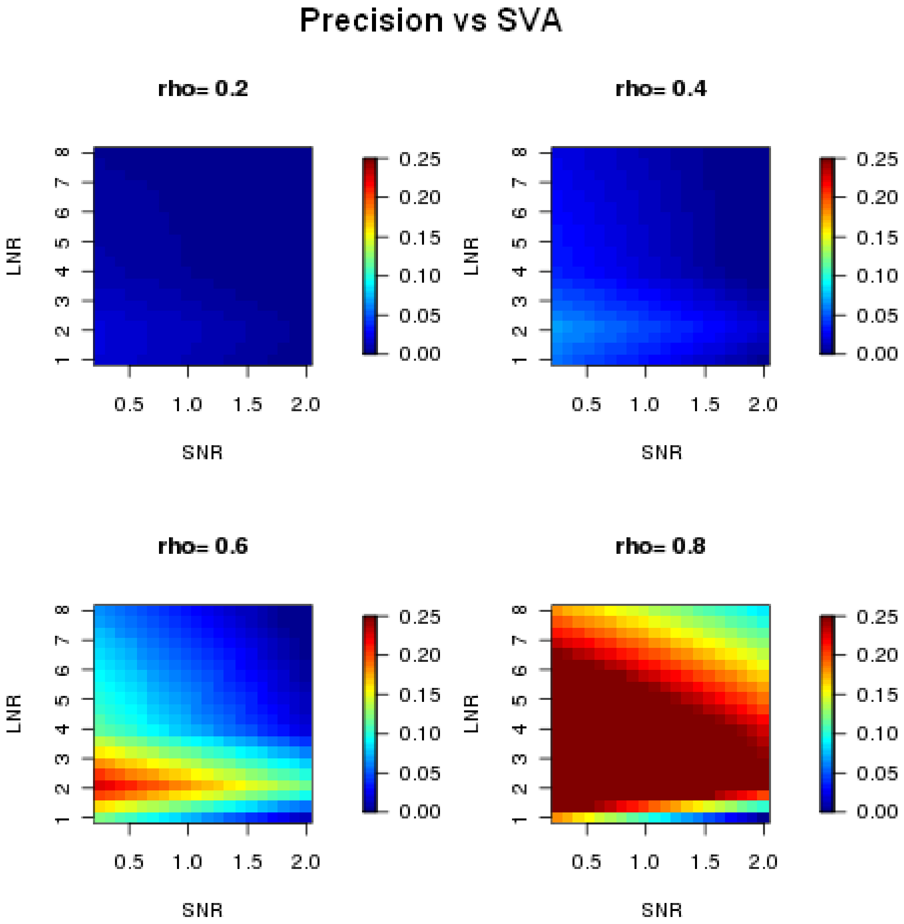

Figure 4: This figure shows the improvement in precision for LEAPP relative to SVA. Precision is the fraction of truly affected genes among the top $H = 50$ ranked genes. The simulation conditions are as described in Figure 2. The color scheme encodes $(\text{PRE}_{\mathbf{lea}} - \text{PRE}_{\mathbf{sva}})/\text{PRE}_{\mathbf{sva}}$.

To keep the comparison manageable as well as to pool information from all tissues, we computed the following quantities

$$I_\alpha = \sum_{1 \le j < j' \le 16} |A_j^\alpha \cap A_{j'}^\alpha|, \quad \text{and} \quad U_\alpha = |\cup_{j=1}^{16} A_j^\alpha|, \tag{16}$$

where A_j^α is the set of statistically significant genes at level α for tissue j. We can think of I_α/U_α as a pooled resemblance. We would like to see large I_α at each given level of U_α.

Figure 6 plots I_α versus U_α for the methods we are comparing. To make a

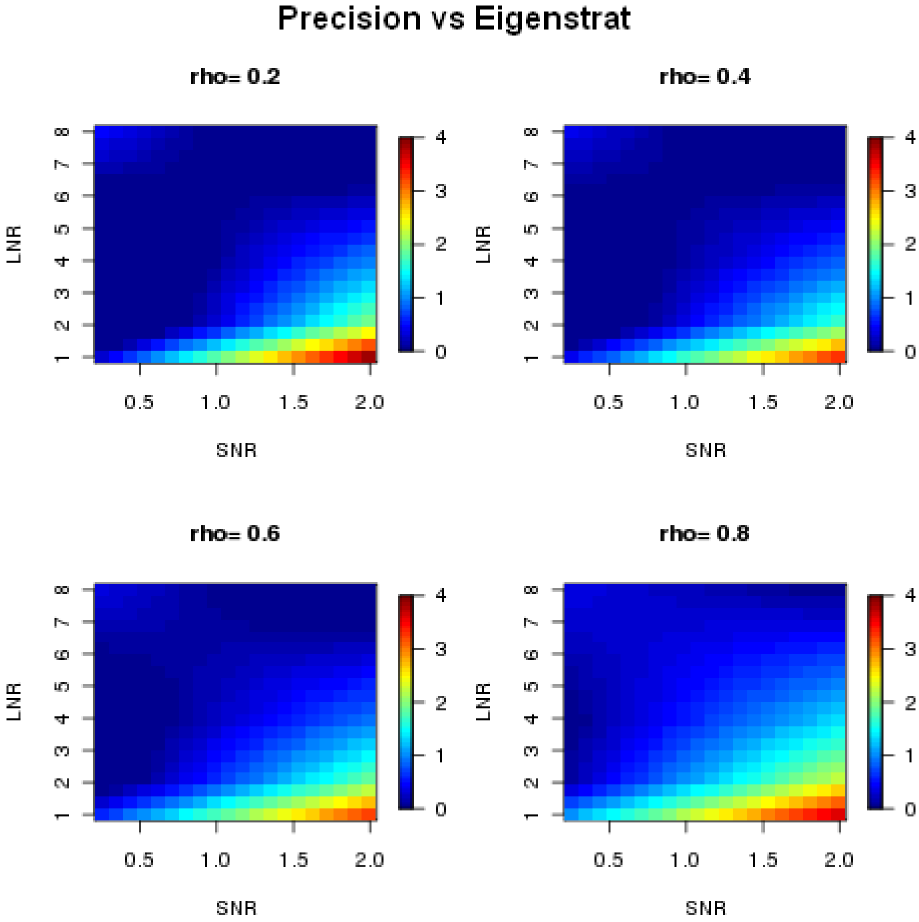

Figure 5: This figure shows the improvement in precision for LEAPP relative to EIGENSTRAT. Precision is the fraction of truly affected genes among the top $H = 50$ ranked genes. The simulation conditions are as described in Figure 2. The color scheme encodes $(\mathrm{PRE}_{\mathbf{rot}} - \mathrm{PRE}_{\mathbf{eig}})/\mathrm{PRE}_{\mathbf{eig}}$.

precise comparison we arranged for each method that estimated latent structure to employ the same estimate for the rank of the latent component. That rank is either 1, 2, 3, or the value chosen by the method of Buja and Eyuboglu (1992). At any rank LEAPP generates the most self-consistent gene lists over almost the entire range. EIGENSTRAT is usually second. SVA beats a raw method that makes no adjustments. LEAPP retains its strong performance when the rank is chosen from the data while the other two methods become poorer in that case.

Resemblance across tissues could also be high if there exists latent variables strongly correlated with age which are repeated across tissues. For example, consider a scenario where all tissues from young mice are in one batch, and all tissues from elder mice are in a different batch. If there are strong batch biases, then “age-related” genes would be reported by the raw method, and the same genes would be ranked high across all tissues. However, note that raw performs the worst of all methods in Figure 6, which gives some reassurance that the high resemblance of the other methods is due to successful removal of latent variables.

Given what we have learned from simulations, the relative performance of EIGENSTRAT and SVA gives us some insight into this data. Since EIGENSTRAT has done well, it is more likely that the signal is not very strong. Since SVA has done poorly, it is more likely that the latent variables in this data are correlated with age. There is also the possibility that they are correlated with sex (the covariate). Our simulations did not include a covariate.

5 Theory

In this section we prove some properties of our approach to testing many hypotheses in the presence of latent variables. We focus on a simpler version of the model that is more tractable:

$$Y = \gamma g^{\mathsf{T}} + UV^{\mathsf{T}} + \sigma E \tag{17}$$

where $g \in \mathbb{R}^{n\times 1}$ with $\|g\| = 1$ as before, $U \in \mathbb{R}^{N\times k}$ is nonrandom $V \in \mathbb{R}^{n\times k}$ has IID rows with $\mathbb{E}(V^{\mathsf{T}}V) = I_k$, known rank k and $E \sim \mathcal{N}(0, I_N \otimes I_n)$. Compared to the full model (1), equation (17) has no covariate term βX^{T}, has constant variance $\Sigma = \sigma I_N$.

This simplification allows us to apply results from the literature to our model. It removes the Monte Carlo based rank estimation step and the alternation between estimating Σ and using the estimate $\widehat{\Sigma}$. When $k = 1$, the primary to latent correlation is $\rho = g^{\mathsf{T}}V/\sqrt{V^{\mathsf{T}}V}$.

Our algorithm requires the choice of a rotation matrix O such that $Og = e_1$. There are multiple possibilities for this matrix. We show that our algorithm is invariant to the choice of O.

Theorem 1. *Let Y follow the model* (17)*. Then our estimates of U and γ do not depend on the rotation O used as long as $Og = e_1$.*

Proof. Let $O_0 \in \mathbb{R}^{n\times n}$ be a fixed orthogonal matrix with $O_0 g = e_1$. Suppose $O \neq O_0$ is any other orthogonal matrix in $\mathbb{R}^{n\times n}$ with $Og = e_1$. Then $O = PO_0$ for an orthogonal matrix P. Now $e_1 = Og = PO_0 g = Pe_1$ and so the first column of P is e_1. Therefore

$$P = \begin{pmatrix} 1 & \mathbf{0}_{n-1}^{\mathsf{T}} \\ \mathbf{0}_{n-1} & P^{\star} \end{pmatrix}, \tag{18}$$

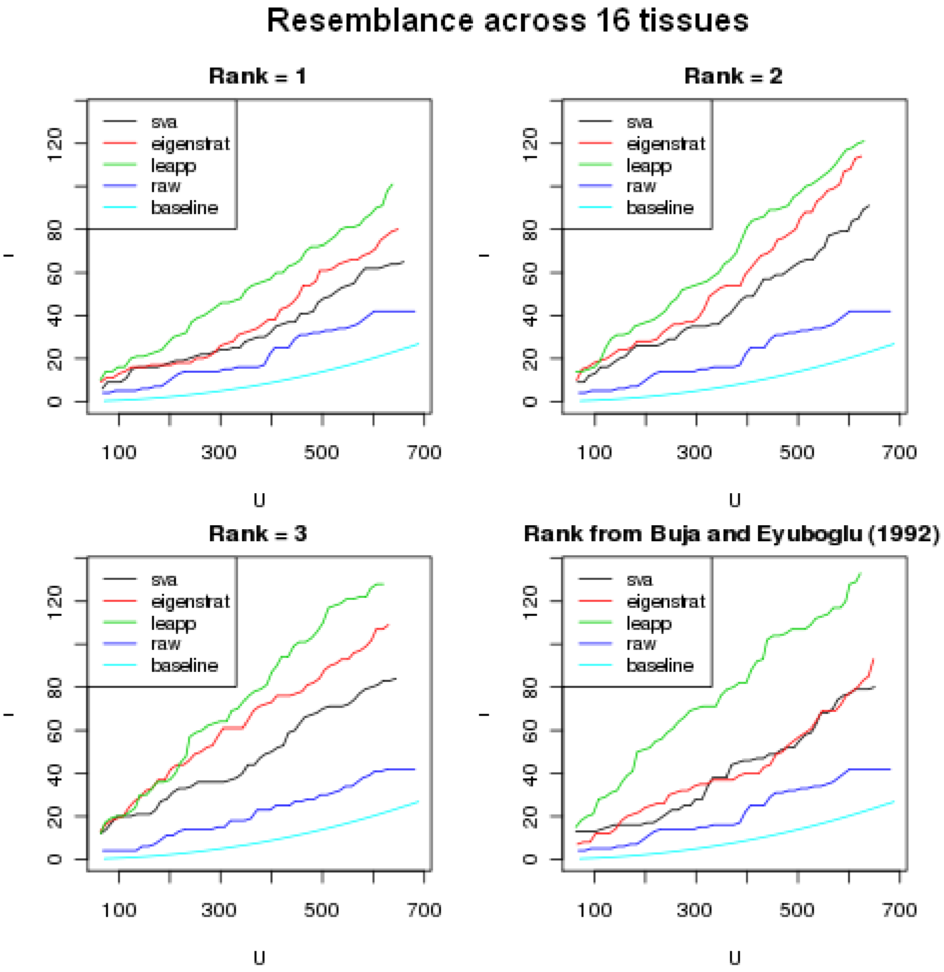

Figure 6: This figure shows the resemblance among significant gene sets from 16 tissues in the AGEMAP study. We plot I_α versus U_α (from equation (16)) increasing α from 0 until $U_\alpha = 700$. The greatest self-consistency among lists is from LEAPP. EIGENSTRAT is second best. The baseline curve is computed assuming that the rankings for all 16 tissues are mutually independent.

for an orthogonal matrix $P^\star \in \mathbb{R}^{(n-1)\times(n-1)}$. Let $\widetilde{Y}^0 = YO_0^\mathsf{T}$ and $\widetilde{Y} = YO^\mathsf{T}$. Then $\widetilde{Y} = YO^\mathsf{T} = YO_0^\mathsf{T}P^\mathsf{T} = \widetilde{Y}^0P^\mathsf{T}$. From (18), we know that the first column of $\widetilde{Y}$ equals that of $\widetilde{Y}^0$.

Let the last $n-1$ columns of $\widetilde{Y}$ and $\widetilde{Y}^0$ be $\overline{Y}$ and $\overline{Y}^0$ respectively. Then from (18), we notice that $\overline{Y} = \overline{Y}^0 P^{\star\mathsf{T}}$. As $P^\star$ is an orthogonal matrix, the left singular vectors of $\overline{Y}$ are the same as those of $\overline{Y}^0$ up to a sign change. Similarly the singular values of $\overline{Y}^0$ match those of $\overline{Y}$. These left singular vectors (multiplied by the corresponding singular values) are the columns of $\widehat{U}$ and $\widehat{U}^0$ respectively. Therefore $\widehat{U} = \widehat{U}^0 S$ where $S = \mathrm{diag}(\pm 1, \dots, \pm 1) \in \mathbb{R}^{k\times k}$.

The outlier detection algorithm uses $\widetilde{Y}_{i1}$ and $\widehat{U}$ (for rotation O) or $\widetilde{Y}^0_{i1}$ and $\widehat{U}^0$ (for rotation O^0). Because $\widetilde{Y}_{i1} = \widetilde{Y}^0_{i1}$ the only change can be through the choice of $\widehat{U}$ or $\widehat{U}^0 = \widehat{U}S$. The matrix $\widehat{U}$ enters (7) through $\widehat{U}^\mathsf{T}\widetilde{V}_1 = (\widehat{U}^0)^\mathsf{T} S\widetilde{V}_1$. Changing the sign of the j'th column of $\widehat{U}$ results in a sign change for the j'th estimated coefficient in $\widetilde{V}_1$, so that $\hat{\gamma} = \hat{\gamma}^0$. □

It is not hard to extend the proof of Theorem 1 to account for the βX^T term. The criss-cross regression begins by computing $\hat{\beta}$ from sums of squares and cross-products. Those sums of squares and cross-products are invariant under the rotation.

The following theorem provides a sufficient condition for our estimate $\widehat{U}$ to consistently estimate U. We study the case where the data are generated with $k = 1$ and the model is also estimated using the correct rank $k = 1$. Then as long as the latent factor U is large enough compared to the noise level, we will be able to detect and estimate U fairly well. Our size measure $\|U\|_2^2(1-\rho^2)/n$ takes account of the correlation. With a higher ρ, more of the latent factor is removed from $\overline{Y}$.

We measure error by the cosine $\Phi(\widehat{U}, U) = \widehat{U}^\mathsf{T} U/(\|\widehat{U}\|_2\|U\|_2)$ of the angle between $\widehat{U}$ and U. The estimate $\widehat{U}$ is determined only up to sign. Replacing $\widehat{U}$ by $-\widehat{U}$ causes a change from $\widehat{V}$ to $-\widehat{V}$ and leaves the model unchanged. We only need $\max(\Phi(\widehat{U}, U), \Phi(-\widehat{U}, U)) = |\Phi(\widehat{U}, U)| \to 1$ for consistency.

Theorem 2. *Let Y follow the model* (17) *with $k = 1$ and $\|U\|_2^2(1-\rho^2)/n \to \infty$ and $N(n)/n \to c \in (0, \infty)$ as $n \to \infty$. Let $\widehat{U}$ be our estimator for U using $k = 1$. Then $|\Phi(\widehat{U}, U)| \to 1$ as $n \to \infty$ with probability* 1.

Proof. In this setting our estimator $\widehat{U}$, from criss-cross regression, is the top left singular vector of $\overline{Y}$, multiplied by the top singular value. Write $V = \rho g + \sqrt{1-\rho^2}W$, where $W \perp g$. By construction $\widetilde{g} = Og = (1, 0, \cdots, 0)$, $\widetilde{W} = OW = (0, \overline{W}^\mathsf{T})^\mathsf{T}$, and $\|\overline{W}\| = 1$. Therefore

$$\overline{Y} = \sqrt{1-\rho^2}\, U\overline{W}^\mathsf{T} + \Sigma\overline{E}.$$

Define

$$\widetilde{\mu} = \frac{\min(\mathrm{Sp}(\lim_{n\to\infty}(\frac{1}{n}(1-\rho^2)UU^\mathsf{T})))}{\max \mathrm{Sp}(\lim_{n\to\infty}(\frac{1}{n}\overline{E}\,\overline{E}^\mathsf{T}))}$$

where $\mathrm{Sp}(A)$ is the set of non zero eigenvalues of the matrix A. Proposition 9 of Harding (2009) yields

$$\sqrt{\frac{1-c/\widetilde{\mu}^2}{1+c/\widetilde{\mu}}} \le |\Phi(\widehat{U}, U)| \le 1.$$

Now $\|U\|_2^2(1-\rho^2)/n \to \infty$ implies that $\widetilde{\mu} \to \infty$ and so $|\Phi(\widehat{U}, U)| \to 1$. □

Next we give conditions for the final step of LEAPP to accurately estimate γ, that is, for $\|\hat{\gamma} - \gamma\|_2$ to be small. To do this we combine methods used in random matrix theory from Bai (2003) with methods used in compressed sensing in Candes and Randall (2006).

In our simulations we found little difference between robust and non-robust versions of the Θ-IPOD algorithm. This is not surprising, since our simulations did not place nonzero γ_i preferentially at high leverage points (extreme u_i). For our analysis we replace the robust Θ-IPOD algorithm by the Dantzig selector for which strong results are available.

Our algorithm was designed assuming that the primary variable g is not too strongly correlated with the latent variable V. In our analysis we also impose a separation between the effects γ and the latent quantity U. Specifically, we assume that γ is sparse and that U is not.

The vector x is s-sparse if it has at most s nonzero components. Following Candes and Randall (2006), we define the sequences $a_s(A)$ and $b_s(A)$ as the largest and smallest numbers (respectively) such that

$$a_s(A)\|x\|_2 \le \|Ax\|_2 \le b_s(A)\|x\|_2$$

holds for all s-sparse x.

Theorem 3. *Suppose that Y follows the model* (17) *with $k = 1$, a fixed correlation $\rho \in (-1, 1)$ between g and V, and an s-sparse vector γ. Assume that $N/n \to c \in (0, \infty)$, $V^{\mathsf{T}}V \xrightarrow{p} 1$, and $(Nn)^{-1}\|U\|_2^2 \to \sigma_u^2 > 0$ hold as $n \to \infty$. Let our estimated U be $\widehat{U}$ and set $U^\star = \widehat{U}/\|\widehat{U}\|_2$. Writing $|U^\star_{(1)}| \ge |U^\star_{(2)}| \ge \cdots \ge |U^\star_{(N)}|$ for the ordered components of $U^\star$, assume that there is a constant $0 < B < 1$ such that*

$$\sum_{i=1}^{2s} (U^\star_{(i)})^2 + \frac{1}{2}\sum_{i=1}^{3s} (U^\star_{(i)})^2 \le B.$$

Then the Dantzig estimator $\hat{\gamma}$, which minimizes

$$\|\hat{\gamma}\|_1 \quad \textit{subject to} \quad \|(I - U^\star U^{\star\mathsf{T}})(\widetilde{Y}_1 - \hat{\gamma})\|_\infty \le \sigma\sqrt{\log N}$$

satisfies

$$\|\hat{\gamma} - \gamma\|_2^2 \le \frac{16\sigma^2 s \log(N)}{(1-\rho^2)(1-B)^2}.$$

Proof. As we saw in the proof of Theorem 2, the last $n - 1$ columns of the rotated data are

$$\overline{Y} = \sqrt{1-\rho^2}\,U\overline{W}^{\mathsf{T}} + \overline{E}$$

where $\|\overline{W}\|_2 = 1$ and $\overline{E}_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Our setting is a special case of that in Theorem 2(i) of Bai (2003): we have just one latent variable and no time series structure. From that theorem,

$$\xi_i \equiv \widehat{U}_i - U_i \xrightarrow{d} \mathcal{N}\Big(0, \frac{\Psi_i}{(1-\rho^2)Q^2}\Big) \quad i = 1, \ldots, N \quad \text{as} \quad n \to \infty$$

where

$$\Psi_i \equiv \mathrm{Var}\left(\sum_{j=1}^{n} \overline{W}_j \overline{E}_{ij}\right) = \sigma^2, \quad \text{and} \quad Q \equiv \mathrm{plim}_{n\to\infty} \widehat{\overline{W}}^{\mathsf{T}} \overline{W} = 1.$$

In his proof of Theorem 2, Bai (2003) further shows that $\xi_i = \eta_i + O_p(1/\sqrt{n})$, with $\eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2/(1-\rho^2))$. We write the first column of the rotated data as

$$\begin{aligned} y &\equiv U\rho + \gamma + \sigma\epsilon \\ &= \widehat{U}\rho + \gamma + \sigma\epsilon + (U - \widehat{U})\rho \\ &= \widehat{U}\rho + \gamma + \widetilde{\epsilon}, \end{aligned}$$

where $\widetilde{\epsilon}_i = \sigma\epsilon_i + \rho\eta_i + o_p(1)$. Next we set $\tau_i \equiv \sigma\epsilon_i + \rho\eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2/(1-\rho^2))$, which follows because $\sigma\epsilon$ and $U - \widehat{U}$ are independent. Then $\widetilde{\epsilon}_i \xrightarrow{d} \mathcal{N}(0, \sigma^2/(1-\rho^2))$.

The proof of Theorem 3.1 in Candes and Randall (2006) shows that the Dantzig estimator $\hat{\gamma}$ satisfies

$$\|\hat{\gamma} - \gamma\|_2 \le 4\sqrt{s}\frac{\sigma\sqrt{\log N}}{\sqrt{1-\rho^2}D}$$

where $D = 1 - b_{2s}^2(U^{\star\mathsf{T}}) - b_{3s}^2(U^{\star\mathsf{T}})/2 + a_{3s}^2(U^{\star\mathsf{T}})/2$. As $U^\star$ is a vector, we have for any $h \ge 1$, $a_h^2(U^{\star\mathsf{T}}) \ge 0$ and $b_h^2(U^{\star\mathsf{T}}) \le \sum_{i=1}^{h}(U_{(i)}^\star)^2$. Hence as long as there exists a constant $0 < B < 1$ for which $\sum_{i=1}^{2s}(U_{(i)}^\star)^2 + \frac{1}{2}\sum_{i=1}^{3s}(U_{(i)}^\star)^2 \le B$, then $D \ge 1 - B$ and

$$\|\hat{\gamma} - \gamma\|_2 \le 4\sqrt{s}\frac{\sigma\sqrt{\log N}}{\sqrt{1-\rho^2}(1-B)}$$

and thus our result is proved. Note that the left hand side grows linearly with the sample size n, because we scale $g^{\mathsf{T}}g = 1$, so the term $o_p(1)$ in the $\widetilde{\epsilon}_i$ is negligible. □

6 Conclusions

High throughput testing has performance that deteriorates in the presence of latent variables. Latent variables that are correlated with the treatment variable of interest can severely alter the ordering of p-values. Our LEAPP method separates the latent variable from the treatment variable, making an adjustment possible. We have found in simulations that the adjustment brings about a better ordering among hypotheses than is available from either SVA or EIGENSTRAT. The improvement over SVA is largest when the latent variable is correlated with the primary one. The improvement over EIGENSTRAT is largest when the primary variable has a large effect. On the AGEMAP data we found better consistency among tissues for significance estimated by LEAPP than for either SVA or EIGENSTRAT.

Acknowledgments

This work was supported by grants DMS-0906394 and DMS-0906056, from the U.S. National Science Foundation.

References

Allen, G. I. and Tibshirani, R. J. (2010). Inference with transposable data: modeling the effects of row and column correlations. Technical report, Stanford University, Department of Statistics.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Broder, A. Z. (1997). On the resemblance and containment of documents. In *In Compression and Complexity of Sequences (SEQUENCES97*, pages 21–29. IEEE Computer Society.

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540.

Candes, E. J. and Randall, P. A. (2006). Highly robust error correction by convex programming. *IEEE Transactions on Information. Theory*, 54:2829–2840.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion. *Biometrika*, 94:759–771.

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, 36(19):e126+.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple testing procedures with applications to genetics*. Springer-Verlag, New York.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing and prediction*. Cambridge University Press, Cambridge.

Efron, B. M. (2007). Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104:1406–1415.

Gabriel, K. R. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21:489–498.

Gannaz, I. (2006). Robust estimation and wavelet thresholding in partial linear models. Technical report, University Joseph Fourier.

Harding, M. C. (2009). Structural estimation of high-dimensional factor model. Technical report, Stanford University, Economics.

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327.

Leek, J. T., Scharpf, R. B., Corrada-Bravo, H., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerley, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733–739.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Science*, 105:18718–18723.

Lucas, J. E., Kung, H.-N., and Chi, J.-T. A. (2010). Latent factor analysis to discover pathway-associated putative segmental aneuploidies in human cancers. *PLOS Computational Biology*, 6:e100920:1–15.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557–572.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the SVD and the non-negative matrix factorization. *Annals of applied statistics*. Tentatively accepted.

Patterson, N. J., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12):2074–2093.

Perry, P. O. (2009). *Cross-validation for unsupervised learning*. PhD thesis, Stanford University.

Perry, P. O. and Owen, A. B. (2010). A rotation test to verify latent structure. *Journal of Machine Learning Research*, 11:603–624.

Price, A. L., Patterson, N. J., Plengt, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components ananysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38:904–909.

She, Y. and Owen, A. B. (2011). Outlier identification using nonconvex penalized regression. *Journal of the American Statistical Association*. to appear.

Storey, J. D., Akey, J. M., and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, 3(8):1380–1390.

Tracy, C. and Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):295–327.

Zahn, J., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., Lakatta, E. G., Boheler, K. R., Xu, X., Mattson, M. P., Falco, G., Ko, M. S. H., Schlessinger, D., Firman, J., Kummerfeld, S. K., III, W. H. W., Zonderman, A. B., Kim, S. K., and Becker, K. G. (2007). AGEMAP: A gene expression database for aging in mice. *PLoS Genetics*, 3(11):2326–2337.

Appendix

Our proposal splits the data set into two pieces. We then apply two different methods to those pieces. Those methods are described here.

A.1 Θ–IPOD

Here we describe the Θ–IPOD algorithm of She and Owen (2011). We use the standard regression notation, so that, for example, Y has a usual regression response meaning in this appendix, not the matrix version we use in the body of the article.

The additive outlier model is

$$Y = X\beta + \gamma + \varepsilon,$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}^n$ and ε_i are IID mean zero random variables. In terms of vector indices, the model is

$$Y_i = X_i^{\mathsf{T}}\beta + \gamma_i + \varepsilon_i.$$

The parameter γ_i is the coefficient of a dummy variable that eliminates a potential outlying i'th observation. This formulation was earlier used by Gannaz (2006)

There are $n + p$ parameters. Those in γ need to be strongly regularized to prevent fitting a saturated model.

The first criterion they consider is

$$\|y - X\beta - \gamma\|_2 + \sum_{i=1}^{n} \lambda_i|\gamma_i|$$

with $\lambda_i = \lambda\sqrt{1 - h_i}$, where h_i is the i'th diagonal element of the hat matrix $H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$ and $\lambda \geq 0$ must be determined.

Given γ we get β by regression of $Y - \gamma$ on X. Given β we get γ by thresholding: $\gamma_i = \Theta(y_i - X_i\beta, \lambda_i)$. An L_1 penalty corresponds to

$$\Theta(z, \lambda) = \Theta_{\text{soft}}(z, \lambda) = \begin{cases} 0 & |z| \leq \lambda \\ z - \operatorname{sgn}(z)\lambda & |z| > \lambda. \end{cases}$$

An L_1 penalty corresponding to soft thresholding does not produce robust results. Better results come from hard thresholding:

$$\Theta_{\text{hard}}(z,\lambda) = \begin{cases} 0 & |z| \le \lambda \\ z & |z| > \lambda. \end{cases}$$

The choice of λ is based on a modified BIC from Chen and Chen (2008).

To handle unequal error variances, we write

$$Y_i = X_i^{\mathsf{T}}\beta + \gamma_i + \sigma_i\varepsilon_i,$$

for $\sigma_i > 0$. Making the replacements $\widetilde{Y}_i = Y_i/\sigma_i$, $\widetilde{X}_i = X_i/\sigma_i$ and $\widetilde{\gamma}_i = \gamma_i/\sigma_i$ we get

$$\widetilde{Y}_i = \widetilde{X}_i^{\mathsf{T}}\beta + \widetilde{\gamma}_i + \varepsilon_i.$$

We apply the original IPOD algorithm to the reweighted points and then multiply the estimated γ_i by σ_i.

A.2 Criss-cross regressions

Gabriel and Zamir (1979) studied how to fit regression models of the form

$$Y = X\beta^{\mathsf{T}} + \delta Z^{\mathsf{T}} + UV^{\mathsf{T}} + E$$

where X are measured features of the rows with coefficients in β, Z are measured features of the columns with coefficients in δ and UV^{T} are latent variables (both unknown) and E is an error matrix.

They estimate β, then δ then U and V sequentially. First

$$\hat{\beta} = Y^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1},$$

then

$$\hat{\delta} = (Y - X\hat{\beta})Z(Z^{\mathsf{T}}Z)^{-1}$$

and finally setting $R = Y - X\hat{\beta}^{\mathsf{T}} - \hat{\delta}Z^{\mathsf{T}}$ he computes the SVD $R = U_R D_R V_R^{\mathsf{T}}$ and picks $\widehat{U}$ and $\widehat{V}$ such that $\widehat{U}\widehat{V}^{\mathsf{T}} = U_R D_R V_R^{\mathsf{T}}$. For example the singular values D_R can be absorbed into either the left or right factors, or $D_R^{1/2}$ can be absorbed into each. In our analysis we suppose that $\widehat{V}$ is normalized to be a unit vector.

When the rows have unequal variances we may replace E by ΣE. Our approach to criss-cross regression is then to alternate between fitting criss-cross regression to $\widehat{\Sigma}^{-1}Y$ and estimating Σ by

$$\widehat{\Sigma}^2 = \frac{1}{n}\mathrm{diag}\Big((Y - X\hat{\beta}^{\mathsf{T}} - \hat{\delta}Z^{\mathsf{T}} - \widehat{U}\widehat{V}^{\mathsf{T}})(Y - X\hat{\beta}^{\mathsf{T}} - \hat{\delta}Z^{\mathsf{T}} - \widehat{U}\widehat{V}^{\mathsf{T}})^{\mathsf{T}}\Big).$$

Our setting is somewhat simpler than the general case. We do not have δZ^{T} term. We followed the steps above but with $\delta Z^{\mathsf{T}} = 0$.