Degrees of Freedom for Combining Regression with Factor Analysis

Patrick O. Perry          Natesh S. Pillai
New York University       Harvard University

October 27, 2013

Abstract

In multivariate regression problems with multiple responses, there often exist unobserved covariates which are correlated with the responses. It is possible to estimate these covariates via factor analytic methods, but calculating unbiased error variance estimates after adjusting for latent factors requires assigning appropriate degrees of freedom to the estimated factors. Many ad-hoc solutions to this problem have been proposed without the backup of a careful theoretical analysis. Using recent results from random matrix theory, we derive an expression for degrees of freedom. Our estimate gives a principled alternative to ad-hoc approaches in common use. Extensive simulation results show excellent agreement between the proposed estimator and its theoretical value. When we apply the methods to a microarray dataset, with 2 estimated latent factors, our estimate assigns between 2.18 and 2.99 degrees of freedom, depending on which response is under consideration.

Key words: Degrees of Freedom, Factor Models, Least Squares, Multiple Regression, Random Matrix Theory, Phase Transition, Eigenvalue.
1. INTRODUCTION

Multivariate response data is prevalent in diverse applications ranging from agriculture to econometrics to psychology (Bock and Gibbons, 1996; Elrod and Keane, 1995; Song and Lee, 2001; Stock and Watson, 2002; Zahn et al., 2007). In these applications, the response variate can be conceived of as a matrix, \( Y = [y_{ij}] \in \mathbb{R}^{n \times m} \); the goal is to explain the variability in the response, and to uncover the relationship between \( Y \) and observed covariates.

Given row and column covariate matrices \( X = [x_{ik}] \in \mathbb{R}^{n \times p} \) and \( Z = [z_{jk}] \in \mathbb{R}^{m \times q} \), one natural model linking the covariates to the response is

\[
Y = XB^T + AZ^T + E,
\]

where \( B = [\beta_{jk}] \in \mathbb{R}^{m \times p} \) and \( A = [\alpha_{ik}] \in \mathbb{R}^{n \times q} \) are unknown coefficient matrices and \( E = [\varepsilon_{ij}] \in \mathbb{R}^{n \times m} \) is a matrix of random errors. The full coefficient matrices are not identifiable, as can be seen by the identity \( XB^T + AZ^T = X(B + ZC)^T + (A - XC^T)Z^T \).

However, for any vector \( s \) orthogonal to the column covariates \( (Z^Ts = 0) \) it is possible to identify \( B^Ts \); similarly, for any vector \( t \) orthogonal to the row covariates it is possible to identify \( A^Tt \).

Unfortunately, model (1) is often inadequate for explaining observed data. It is implausible that all sources of variability have been observed. To this end, one popular approach is to posit existence of \( r \) latent factors, such that

\[
Y = XB^T + AZ^T + UV^T + E,
\]

where \( U = [u_{ik}] \in \mathbb{R}^{n \times r} \) is thought of as a matrix of row scores and \( V = [v_{jk}] \in \mathbb{R}^{m \times r} \) is a matrix of column loadings. Model (2), which combines regression and factor analysis, is known as a bilinear model (Gabriel, 1978).

The bilinear model has appeared in various forms, and it has a long history dat-
ing back to Fisher and Mackenzie (1923), with notable early contributions by Cochran (1943) and Williams (1952). The model was relatively obscure until Tukey, unaware of his predecessors, suggested combining regression and factor analysis in his essay, “The future of data analysis” (1962). This inspired Gollob (1968) and Mandel (1969, 1971) to independently reinvent Williams’ version of the latent factor model. At this point, the models and its variants saw broader adoption. Freeman (1973) surveys the early history, and Bartholomew et al. (2011) give a more recent history.

The special case when $X = 1_{n,1}$ and $Z = 1_{m,1}$ continues to be popular in agronomy, where it is known as the additive main effects with multiplicative interaction (AMMI) model (Crossa et al., 2002; dos S. Dias and Krzanowski, 2003). Other recent work on related latent factor models include papers by van Eeuwijk (1995), Cornelius and Seyed-sadr (1997), Gabriel (1998), West (2003), Hoff (2007), Carvalho et al. (2008), Leek and Storey (2008), Friguet et al. (2009), and Sun et al. (2012).

Usually, the parameters of a bilinear model are estimated via least squares. After this estimation, to perform inference on the coefficients, we need an estimate of the error variance. To this end, a persistent challenge is the assignment of the appropriate “degrees of freedom” to estimates of the factor term. The statistical literature remains divided on this issue:

- **Gollob (1968)** proposed a parameter-counting scheme. The least squares estimate of the first column of $U$, which is orthogonal to $X$, has $n$ components but satisfies $p$ constraints. Similarly, the least squares estimate of the first column of $V$ has $m$ components but satisfies $q$ constraints. The scale of either estimated column can be fixed without affecting the overall fit. Thus, Gollob allocates $(n - p) + (m - q) - 1$ degrees of freedom to the first term of the estimated factor. Similarly, he allocates $(n - p - k) + (m - q - k) - 1$ degrees of freedom to the $(k - 1)^{th}$ estimated factor term.

- **Mandel (1971)** noted that when there are no true factors ($r = 0$), if the elements
of $E$ are independent normal random variables with common variance, then the squared Frobenius norm (sum of squares) of the $k^{th}$ estimated latent factor is distributed as $\lambda_k$, the $k^{th}$ largest eigenvalue of an $(m - q) \times (m - q)$ white Wishart matrix with $n - p$ degrees of freedom. Thus, Mandel proposes allocating $E[\lambda_k]$ degrees of freedom to the $k^{th}$ estimated factor term, which he computes via Monte Carlo simulation.

- More recent approaches do not assume that the elements of $E$ have a common variance, and they use iterative schemes to estimate the factors and the noise variances simultaneously. Essentially, these approaches treat the estimated factor scores $\hat{U}$ like observed covariates $X$. They either treat the factor loadings as fixed effects, allocating $m$ degrees of freedom to the $k^{th}$ estimated factor (Leek and Storey, 2008; Sun et al., 2012), or they treat the factor loadings as random effects which may result in a smaller estimate for the degrees of freedom (Friguet et al., 2009).

In agronomy and psychometrics applications, with smaller sample sizes, Gollob’s estimate is the more popular method (dos S. Dias and Krzanowski, 2003); Mandel’s assumption of no true factors is seen as inappropriate. In genomics applications, the issues of adjusting for degrees of freedom do not receive much attention, likely due to an implicit assumption that with large sample sizes, the adjustment is unimportant.

In this paper, we bring recent developments in random matrix theory to bear on the degrees of freedom problem. We derive conservative estimates for degrees of freedom that are valid when the problem dimensions are large. Even though these estimates rely on asymptotic approximations, we observe them to be accurate for sizes as small as $n = 10$ and $m = 50$. We demonstrate the utility of these estimates, and of the bilinear model in general, via an application to the AGEMAP dataset (Zahn et al., 2007).

The rest of the paper is organized as follows. We motivate the degrees of freedom estimation problem with a genomics dataset presented in Section 2. Then, in Section 3 we reduce the estimation problem to one in which there are no covariates. In Section 4...
we derive analytically an asymptotic expression for the degrees of freedom, which we
verify in Section 5. Next, in Section 6 we propose a conservative degrees of freedom
estimator. We apply this estimator to a regression problem in Section 7 and we close
with a short discussion in Section 8.

2. MOTIVATION

In the AGEMAP genomics study of $M \approx 18,000$ genes measured in $N = 39$ subjects
(mice), researchers are interested in detecting which genes are related to age (Zahn et al.,
2007). For each subject-gene pair $ij$, with $1 \leq i \leq N$ and $1 \leq j \leq M$, they measure $y_{ij}$, the
log-activation in subject $i$ of gene $j$; taken together, these measurements form a response
matrix $Y = [y_{ij}] \in \mathbb{R}^{N \times M}$.

The researchers have two covariate matrices available. The row covariate matrix,
$X = [x_{ik}] \in \mathbb{R}^{N \times p}$, encodes subject-specific attributes. This matrix has $p = 3$ columns,
for an intercept, the sex, and the age of the subject:

$$
x_{i1} = 1,
$$

$$
x_{i2} = \text{Sex of subject } i \ (\text{Female } = +1, \text{Male } = -1),
$$

$$
x_{i3} = \text{Age of subject } i \ (\text{months}).
$$

The column covariate matrix, $Z = [z_{jl}] \in \mathbb{R}^{M \times q}$, encodes response-specific attributes.
This matrix has $q = 2$ columns, for an intercept and the tissue type of the response:

$$
z_{j1} = 1,
$$

$$
z_{j2} = \text{Tissue of response } j \ (\text{Cerebellum } = +1, \text{Cerebrum } = -1).
$$

To model the associations between the covariates and the response, it is natural to
posit existence of row and column coefficient matrices $A = [\alpha_{il}] \in \mathbb{R}^{N \times q}$ and $B = [\beta_{jk}] \in$
\( \mathbb{R}^{M \times p} \) which link the covariates to the response via the relation

\[
Y = XB^T + AZ^T + E,
\]

where \( E = [ \varepsilon_{ij} ] \in \mathbb{R}^{n \times m} \) is a matrix of mean-zero random errors. The interpretation of \( \beta_j \) is as follows: “holding sex and subject-specific effects constant, increasing age by 1 unit (1 month) is associated with increasing expected log activation of gene j by \( \beta_j \) units.”

Unfortunately, the coefficient matrices are not identifiable. To see this, note that if \( C \in \mathbb{R}^{q \times p} \) is an arbitrary matrix, then \( AZ^T + XB^T = (A + XC^T)Z^T + X(B - ZC)^T \). Replacing \( A \) with \( A + XC^T \) while replacing \( B \) with \( B - ZC \) keeps the sum \( AZ^T + XB^T \) fixed. Thus, it is impossible to identify the elements of \( A \) and \( B \). Despite this non-identifiability, it is possible to identify certain projections of the coefficient matrices. If \( s \) is any vector such that \( Z^Ts = 0 \), then it is possible to identify \( B^Ts \). Similarly, it is straightforward to show that if \( t \) is any vector such that \( X^Tt = 0 \), then it is possible to identify \( A^Tt \). We state the formal result for projections of \( B \) in Proposition 2.1; the result and its proof for \( A \) are analogous.

**Proposition 2.1.** Suppose that \( X \) has full column rank. If \( s \) is a direction vector satisfying \( Z^Ts = 0 \) and if \( B_1 \) and \( B_2 \) are two column coefficient matrices such that \( B_1^Ts \neq B_2^Ts \), then

\[
A_1Z^T + XB_1^T \neq A_2Z^T + XB_2^T
\]

for all row coefficient matrices \( A_1 \) and \( A_2 \) of the appropriate dimensions.

**Proof.** Multiply both sides of the inequality by \( s \). \( \square \)

In the context of AGEMAP, we would like to say that gene \( j \) is associated with age if \( \beta_j \) is nonzero. However, in light of the identifiability issues just raised, this is not a workable definition. Instead, we will say that gene \( j \) is related to age if the age coefficient
for that gene differs from the average age coefficient for all genes of the same tissue type. More precisely, we say that gene \( j \) is related to age if \( \beta \) s\((j)\) = 0, where \( s(j) = (I - H_Z)e_j \), with \( I \) the identity matrix, \( e_j \) the \( j \)th standard basis vector in \( \mathbb{R}^M \) and \( H_Z = \hat{Z}(\hat{Z}^T \hat{Z})^{-1}\hat{Z}^T \). Alternative definitions are possible by using weighted versions of the hat matrix \( H_Z \).

Following Gabriel (1978), we estimate the identifiable components of the coefficient matrices via least squares. We choose estimates \( \hat{A} \) and \( \hat{B} \) to satisfy

\[
(I - H_X)\hat{A} = (I - H_X)YZ(Z^T \hat{Z})^{-1},
\]
\[
(I - H_Z)\hat{B} = (I - H_Z)Y^T X(X^T X)^{-1}.
\]

That is, we find the identifiable components of \( \hat{A} \) by regressing on the row residuals from a column regression of \( Y \) on \( X \); we find the identifiable components of \( \hat{B} \) by regressing on the column residuals from a row regression of \( Y \) on \( Z \). Letting \( \hat{Y} = \hat{A}Z^T + X\hat{B}^T \), the unidentifiable components can be chosen arbitrarily such that

\[
\hat{E} \equiv Y - \hat{Y} = (I - H_X)Y(I - H_Z);
\]

one possibility is to take \( H_Z\hat{B} = H_ZY^T X(X^T X)^{-1} \) and \( H_X\hat{A} = 0 \). When the estimates are chosen in this manner, the next proposition is useful for performing inference on the true column regression coefficients.

**Proposition 2.2.** Suppose that \( X \) is full rank and \( Y = AZ^T + XB^T + E \), where the rows of \( E \) are independent mean-zero multivariate normal random vectors with covariance \( \Sigma \). If \( \hat{A} \) and \( \hat{B} \) are chosen via least squares, and if \( s \) is any vector satisfying \( Z^T s = 0 \), then \( \hat{B}^T s \) and \( s^T \hat{E}^T \hat{E} s \) are independent with

\[
\hat{B}^T s \sim \mathcal{N}(B^T s, s^T \Sigma s \cdot (X^T X)^{-1}),
\]
\[
s^T \hat{E}^T \hat{E} s \sim s^T \Sigma s \cdot \chi^2_{N - p}.
\]
Proof. Define $\beta_s = B^T s$, $\epsilon_s = Es$, and $y_s = Ys = X\beta_s + \epsilon_s$. Set $\hat{\beta}_s = \hat{B}^T s = (X^T X)^{-1} X^T y_s$ and $\hat{\epsilon}_s = \hat{E}s = (I - H_X)y_s$. Finally, note that the elements of $\epsilon_s$ are independent mean-zero normal random variables with variance $s^T \Sigma s$. The results follow from standard linear regression theory (e.g., Draper and Smith (1998)).

The main implication of Proposition 2.2 is that, if $t$ is any vector and $s^T Bt = 0$, then the test statistic

$$T(s, t) \equiv \frac{\sqrt{N - p} \cdot s^T \hat{B}t}{\{s^T \hat{E} \hat{E}^T s \cdot t^T(X^T X)^{-1} t\}^{1/2}}$$

is $t$-distributed with $N - p$ degrees of freedom. This result facilitates hypothesis testing on the components of $B^T s$.

With all this machinery in place, suppose that we want to test whether a particular gene, Mm.71015 (Cerebellum) is related to age. First, we fit $\hat{Y} = \hat{A}Z^T + X\hat{B}^T$ via least squares, and we calculate the residuals $\hat{E} = Y - \hat{Y}$. We set $s = (I - H_Z)e_j$, where $j$ is the index of Mm.71015 (Cerebellum). The estimate (standard error) of the age component of $\hat{B}^T s$ is 0.018 (0.014); the $T$ statistic is 1.36, with 36 degrees of freedom. Apparently, the gene is not significantly related to age.

Visually inspecting the elements of the residual component $\hat{E}s$ reveals a problem with the modeling assumptions (Figure 1). Specifically, our analysis relies on the elements of the regression error component $Es$ being independent mean-zero normal random variables. As evidenced by the multi-model structure in the residuals, the distributional assumptions on the regression errors seem implausible.

An analysis of all $M$ genes further corroborates the evidence of latent structure in the residual matrix $\hat{E}$. If the model were correctly specified, then there should be no apparent row-specific structure in the residual matrix matrix. However, as Figure 2 demonstrates, there are clear clusters in the first two principal component scores computed from $\hat{E}$. One cluster of subjects exhibits low response values across many Cerebrum tissue genes, another cluster exhibits low response values across many Cerebellum tissue genes, and the remaining cluster has medial responses for most genes, regardless of
Figure 1: Residuals Reveal Latent Structure. Residuals from the regression of log activation on sex and age in gene Mm.71015 (Cerebellum). Two clusters of subjects are apparent.
Figure 2: Residual Matrix Reveals Latent Structure. First two principal component scores computed from the residuals after regressing gene response on age and gender. Three clusters of individuals are apparent.

tissue type (Figure 3).

The principal components analysis of the residual matrix hints at the existence of latent subject-specific covariates. It is likely that there is some $N \times r$ matrix $U$ of unobserved subject-specific covariates, and an $M \times r$ matrix $V$ of coefficients such that $Y = AZ^T + XB^T + UV^T + E$. In fact, this is the bilinear model introduced in Section 1.

To make the model identifiable, we require that $U^TX = 0$ and $V^TZ = 0$. Without the identifiability assumption, the least squares estimates of $(I - H_X)A$ and $(I - H_Z)B$ will be biased by $(I - H_X)U(V^TZ)^{-1}Z^T(X^TX)^{-1}$ and $(I - H_Z)VU^TX(X^TX)^{-1}$. In fact, since we never perform interference on $A$, the identifiability assumption on $V$ is inconsequential. The constraint $U^TX = 0$, amounts to a requirement that the latent subject-specific covariates be uncorrelated with the columns of $X$. Even though the identifiability assumption seems strong, making this assumption is less restrictive than assuming that $r = 0$ (that is,
Figure 3: **Mouse Clusters Exhibit Different Response Behaviors.** Scatterplots of age versus log gene activation for ten Cerebrum genes (B) and ten Cerebellum genes (C), with subject colors determined from the clusters identified in Figure 2.
assuming that there are no latent factors which are correlated with the response).

With the estimates $A$ and $B$ the same as in the case with no latent factors, the least squares estimates of $U$ and $V$ can be obtained from the leading $r$ terms of the singular value decomposition of the residual matrix $\hat{E}$. With estimated latent factors having scores $\hat{U}$ and loadings $\hat{V}$, this gives an adjusted residual matrix $\hat{E}_1 = \hat{E} - \hat{U}\hat{V}^T$. Forming the adjusted residual matrix in this way is equivalent to treating $\hat{U}$ like observed row covariates.

We define the degrees of freedom for the latent factors along test direction $s$ as the value $df(s)$ satisfying the equation

$$E[s^T\hat{E}_1^T\hat{E}_1s] = \{N - p - df(s)\} \cdot s^T\Sigma s. \quad (3)$$

If the value of $df(s)$ were known, this would give rise to a natural estimate for the variance along the test direction:

$$\hat{\sigma}^2(s) = \frac{s^T\hat{E}_1^T\hat{E}_1s}{N - p - df(s)}. \quad (4)$$

We devote Sections 3–4 to the correct determination of $df(s)$.

3. REDUCTION TO COVARIATE-FREE CASE

It turns out to be sufficient to consider the case when $p = q = 0$. Consider the model

$$Y = AZ^T + XB^T + UV^T + E,$$

where $Y \in \mathbb{R}^{N \times M}$, $X \in \mathbb{R}^{N \times p}$, $Z \in \mathbb{R}^{M \times q}$, and $UV^T$ has rank $r$. Suppose that identifiability constraints $X^TU = 0$ and $Z^TV = 0$ hold, and that $X$ and $Z$ have full column ranks. Assume that the rows of $E$ are independent mean-zero multivariate normal random vectors with covariance $\Sigma$. Let $s$ be a test direction satisfying $Z^Ts = 0$, and define $\hat{\sigma}^2(s) = s^T\Sigma s$. 

11
Take \( \hat{A}, \hat{B}, \hat{U}, \text{and} \hat{V} \) to be the least squares estimates of the parameters with \( \hat{r} \) estimated latent factors and let \( \hat{E} = Y - (\hat{A}Z^T + X\hat{B}^T + \hat{U}\hat{V}^T) \) be the residual matrix. Define residual degrees of freedom

\[
df_{\text{resid}}(s) = \mathbb{E}(s^T \hat{E}^T \hat{E} s) / \sigma^2(s).
\]

Let \( X = Q_1R \) be the polar decomposition of \( X \); that is, \( Q_1 \in \mathbb{R}^{N \times p} \) is a matrix with orthonormal columns, and \( R \in \mathbb{R}^{p \times p} \) is symmetric and positive definite. Similarly, let \( Z = P_1S \) be the polar decomposition of \( Z \). Choose \( Q_2 \) and \( P_2 \) such that \( Q = [Q_1 \quad Q_2] \) and \( P = [P_1 \quad P_2] \) are orthogonal matrices. Set \( Y_{22} = Q_2^T Y P_2, \ U_2 = Q_2^T U, \ V_2 = P_2^T V, \) and \( E_{22} = U_2^T E V_2 \), so that the reduced model holds:

\[
Y_{22} = U_2 V_2^T + E_{22},
\]

where \( Y_{22} \in \mathbb{R}^{n \times m} \) and \( U_2 V_2^T \) has rank \( r \), with \( n = N - p \) and \( m = M - q \). Note that the rows of \( E_{22} \) are independent mean-zero multivariate normal random vectors with covariance \( \Sigma_{22} = P_2^T \Sigma P_2 \). Define \( s_2 = V_2^T s \) to be the test direction for the reduced model, which satisfies the relation \( s_2^T \Sigma_{22} s_2 = \sigma^2(s) \).

Take \( \hat{U}_2 \) and \( \hat{V}_2 \) to be the least squares estimates from the reduced model for \( Y_{22} \), with \( \hat{r} \) estimated latent factors, and let \( \hat{E}_{22} = E_{22} - \hat{U}_2 \hat{V}_2^T \) be the residual matrix. Define reduced model residual degrees of freedom

\[
df_{\text{resid}}^{(2)}(s_2) = \mathbb{E}(s_2^T \hat{E}_{22}^T \hat{E}_{22} s_2) / \sigma^2(s).
\]

**Theorem 3.1.** Under the above conditions, we have \( \df_{\text{resid}}(s) = \df_{\text{resid}}^{(2)}(s_2) \).

**Proof.** Without loss of generality, redefine \( A \) and \( B \) to reparametrize the model as

\[
Y = AZ^T + XB^T + XZ^T + UV^T + E,
\]
where $X^TA = 0$ and $Z^TB = 0$.

We perform a change of bases and put the model in block form:

$$
\begin{bmatrix}
Q_1^T \\
Q_2^T
\end{bmatrix} Y \begin{bmatrix} P_1 & P_2 \end{bmatrix} =
\begin{bmatrix} Y_{11} & Y_{12} \\
Y_{21} & Y_{22}
\end{bmatrix}
= \begin{bmatrix} R \
A_2 S^T \\
A_2 S^T \
\end{bmatrix}
\begin{bmatrix} R B_2^T \\
U_2 V_2^T
\end{bmatrix}
+ \begin{bmatrix} E_{11} & E_{12} \\
E_{21} & E_{22}
\end{bmatrix},
$$

(5)

where $A_2 = Q_2^T A$ and $B_2 = P_2^T B$ are the identifiable components of the regression coefficients, $U_2 = Q_2^T U$ and $V_2 = P_2^T V$ are the identifiable factor components, and $Y_{kl} = Q_k^T Y P_l$ and $E_{kl} = Q_k^T E P_l$ for $k = 1, 2$ and $l = 1, 2$.

From Equation (5) it is apparent that the least squares estimates of the coefficients are $\hat{A} = Q_2 S^{-1} Y_{21}$, $\hat{B} = P_2 R^{-1} Y_{12}$, and $\hat{\Gamma} = R^{-1} Y_{11} S^{-T}$. The regression residuals $\hat{E}_0 = Y - \hat{A} Z^T - X \hat{B}^T - X \hat{\Gamma} Z^T$ satisfy

$$
\begin{bmatrix}
Q_1^T \\
Q_2^T
\end{bmatrix} \hat{E}_0 \begin{bmatrix} P_1 & P_2 \end{bmatrix} =
\begin{bmatrix} 0 & 0 \\
0 & U_2 V_2^T + E_{22}
\end{bmatrix}.
$$

Thus, the least squares estimate $\hat{U} \hat{V}^T$ obtained from the rank $\hat{r}$ singular value decomposition of $\hat{E}_0$ is equal to $Q_2 \hat{U}_2 \hat{V}_2^T P_2^T$, where $\hat{U}_2 \hat{V}_2^T$ is the rank $\hat{r}$ singular value decomposition of $Y_{22} = U_2 V_2^T + E_{22}$.

The final residual matrix $\hat{E} = \hat{E}_0 - \hat{U} \hat{V}^T$ is given as $\hat{E} = Q_2 \hat{E}_{22} P_2^T$, where

$$
\hat{E}_{22} = Y_{22} - \hat{U}_2 \hat{V}_2^T.
$$

Hence,

$$
s^T \hat{E}^T \hat{E} s = s_2^T \hat{E}_{22}^T \hat{E}_{22} s_2
$$

and $\text{df}_{\text{resid}}(s) = \text{df}_{\text{resid}}^{(2)}(s_2)$; the proof is finished.
4. DEGREES OF FREEDOM

In light of Theorem 3.1 without loss of generality we will assume that there are no row or column covariates \( p = q = 0 \). Our data generating model has \( r \geq 0 \) true latent factors:

\[
Y = \sqrt{n}UDV^T + E, \quad (6)
\]

with \( U \in \mathbb{R}^{n \times r} \), \( V \in \mathbb{R}^{m \times r} \) having orthonormal columns and a diagonal matrix \( D \in \mathbb{R}^{r \times r} \) with \( [D_{kk}] = \sqrt{\mu_k} \) for \( k = 1, \ldots, r \). We assume that the row vectors of the matrix \( E \) are mean-zero multivariate normal with covariance matrix \( \Sigma \).

The estimates \( \hat{U} \) and \( \hat{V} \) can be obtained from the leading \( \hat{r} \) terms of the singular value decomposition (SVD) of \( Y \). We choose the scaling such that \( \sqrt{n}\hat{U}\hat{D}\hat{V}^T \) comprises the leading \( \hat{r} \) terms of the SVD of \( Y \), where \( \hat{U} \in \mathbb{R}^{n \times \hat{r}} \) and \( \hat{V} \in \mathbb{R}^{m \times \hat{r}} \) have orthonormal columns, and \( \hat{D} \in \mathbb{R}^{\hat{r} \times \hat{r}} \) is diagonal with \( [\hat{D}_{kk}] = \sqrt{\hat{\mu}_k} \) for \( k = 1, \ldots, \hat{r} \). After adjusting for the estimated latent factors, the residual matrix is \( \hat{E} = Y - \sqrt{n}\hat{U}\hat{D}\hat{V}^T \). The residual sum of squares along the test direction \( s \in \mathbb{R}^m \) is given by

\[
\text{RSS}(s) \equiv s^T\hat{E}^T\hat{E}s. \quad (7)
\]

For \( s \in \mathbb{R}^m \), define the degrees of freedom

\[
df(s) = \mathbb{E}\left(n - \frac{\text{RSS}(s)}{s^T\Sigma s}\right) \quad (8)
\]

so that

\[
\mathbb{E}\left(\frac{\text{RSS}(s)}{s^T\Sigma s}\right) = n - \text{df}(s).
\]

**Lemma 4.1.** For the model in (6), the residual sum of squares along a test direction \( s \in \mathbb{R}^m \) is given by

\[
\text{RSS}(s) = n \cdot sE^TEs + 2\sqrt{n} \cdot s^TVDU^TEs + n\left(\sum_{k=1}^{r} \mu_k \cdot (v_k^Ts)^2 - \sum_{k=1}^{\hat{r}} \hat{\mu}_k \cdot (\hat{v}_k^Ts)^2\right). \quad (9)
\]
Proof. The proof is a straightforward computation and is deferred to the Appendix. \qed

4.1 The Noise Case

In this subsection we assume that there are no true latent factors, \( i.e., r = 0 \), and so \( Y = E \). The rows of \( E \) are independently distributed according to \( \mathcal{N}(0, \Sigma) \).

**Theorem 4.2.** Suppose \( \Sigma = \sigma^2 I \) for some \( \sigma > 0 \). If \( \lim_{n \to \infty} \frac{n}{m} = c \in (0, \infty) \), then

\[
\text{df}(s) = \hat{r} \left( 1 + \sqrt{\frac{n}{m}} \right)^2 + o(1). \tag{10}
\]

**Proof.** For ease of exposition, we first assume \( \sigma = 1 \). Applying Lemma 4.1 with \( r = 0 \) yields

\[
\text{RSS}(s) = ns^T EE^T s - n \sum_{k=1}^{\hat{r}} \hat{\mu}_k \cdot (\hat{\vartheta}_k^T s)^2,
\]

and thus

\[
\mathbb{E} \text{RSS}(s) = ns^T s - n \sum_{k=1}^{r} \mathbb{E} (\hat{\mu}_k \cdot (\hat{\vartheta}_k^T s)^2).
\]

The matrix \( EE^T \) is a Wishart matrix with \( n \) degrees of freedom and scale parameter \( \Sigma \). The values \( \hat{\mu}_1, \ldots, \hat{\mu}_{\hat{r}} \) are the \( \hat{r} \) largest eigenvalues of \( (1/n) EE^T \). \textit{Yin et al.} (1988) show under very general conditions that, as \( \lim_{n \to \infty} \frac{n}{m} = c \in (0, \infty) \),

\[
\hat{\mu}_k - (1 + \sqrt{m/n})^2 \xrightarrow{a.s.} 0.
\]

Also, the distribution of \( \hat{\vartheta} \) is invariant under multiplication by any \( m \times m \) orthogonal matrix, hence \( \mathbb{E}[\hat{\vartheta}^T s]^2 = (s^T s)/m \). Thus, if we set

\[
\text{df}_k(s) = (1 + \sqrt{n/m})^2, \quad 1 \leq k \leq \hat{r},
\]
and \( \text{df}(s) = \sum_{k=1}^{p} \text{df}_k(s) \), it follows that

\[
\{ n - \text{df}(s) \} - \mathbb{E}\left[ \frac{\text{RSS}(s)}{s^T s} \right] \to 0
\]

proving the claim for \( \sigma = 1 \). The proof for an arbitrary \( \sigma > 0 \) follows by an identical argument with minor changes.

4.2 The Signal Case

Here we assume that the data are generated according to model (6) with \( r > 0 \) latent factors. Without loss of generality, the matrix \( D \in \mathbb{R}^{r \times r} \) is diagonal with \( [D_{kk}] = \sqrt{\mu_k} \) for \( k = 1, \ldots, r \) and \( \mu_1 > \cdots > \mu_r > 0 \). Let \( \sqrt{n\hat{U}\hat{D}\hat{V}^T} \) be the \( \hat{r} \)-term estimated latent factors obtained from the leading terms of the singular value decomposition of \( Y \), with \( \hat{r} \) not necessarily equal to \( r \). Let \( v_k \) denote the \( k \)-th column of \( V \).

For any test vector \( s \in \mathbb{R}^m \), write \( s = s_V + s_{V^\perp} \) where \( s_V, s_{V^\perp} \) respectively denote the projections of \( s \) to the column spaces spanned by \( V, V^\perp \). Also recall the degrees of freedom \( \text{df}(s) \) given by (8). We need the following lemma, whose proof is given in the Appendix.

**Lemma 4.3.** The estimate of the \( k \)-th factor can be decomposed as

\[
\hat{\vartheta}_k = \sum_{l=1}^{\hat{r}} \hat{\rho}_{kl} v_l + (1 - \hat{\rho}_k^2)^{1/2} \tilde{\vartheta}_k,
\]

where \( \hat{\rho}_k^2 = \sum_{l=1}^{\hat{r}} \hat{\rho}_{kl}^2 \) and \( \tilde{\VV} = [\tilde{\vartheta}_1 \ldots \tilde{\vartheta}_{\hat{r}}] \) is Haar-distributed over the subspace of the Steifel manifold orthogonal to the column space of \( V \).

For \( k \leq r \), define the quantity

\[
\text{df}_k(s) = n \left( 1 - \frac{m}{\mu_k} - \frac{m}{\mu_k^2} \right) \frac{(v_k^T s)^2}{s^T s} + \left( 1 + \frac{1}{\mu_k} \right)^2 \left( 1 - \frac{(v_k^T s)^2}{s^T s} \right). \tag{11}
\]
Theorem 4.4. Suppose $\Sigma = I$, $\hat{r} = r$. If $n/m = c + o(n^{-1/2})$ for some $c \in (0, \infty)$ and if $\mu_r > c^{-1/2}$, then

$$df(s) = \sum_{k=1}^{r} df_k(s) + o\left(n^{1/2} \sum_{k=1}^{r} \frac{(v_k^T s)^2}{s^T s} + n^{-1/2}\right).$$

(12)

Remark 4.5. Notice that Theorem 4.4 holds for $\Sigma = \sigma^2 I$ for any $\sigma > 0$. For $\sigma \neq 1$, we just need to replace $\mu_k$ with $\mu_k / \sigma^2$ in Equation (11).

Proof. Since $\hat{r} = r$, from Lemma 4.1 we obtain

$$\mathbb{E}\{\text{RSS}(s)\} = n \cdot s^T s + n \cdot \sum_{k=1}^{r} [\mu_k \cdot (v_k^T s)^2 - \mathbb{E}\{\hat{\mu}_k \cdot (\hat{v}_k^T s)^2\}].$$

(13)

Thus

$$df(s) = n - \frac{\mathbb{E}\{\text{RSS}(s)\}}{s^T s} = -(n / s^T s) \cdot \sum_{k=1}^{r} [\mu_k \cdot (v_k^T s)^2 - \mathbb{E}\{\hat{\mu}_k \cdot (\hat{v}_k^T s)^2\}].$$

(14)

We focus our attention on the $k$th summand of the last term. Write

$$\hat{\mu}_k = \bar{\mu}_k + n^{-1/2} Z_k,$$

(15)

$$\hat{\rho}_{kl} = \bar{\rho}_{kl} + n^{-1/2} W_{kl},$$

(16)

where

$$\bar{\mu}_k = (\mu_k + 1) \left(\frac{m}{n\mu_k} + 1\right),$$

(17)

$$\bar{\rho}_{kk}^2 = \frac{1 - \frac{m}{n\mu_k}}{1 + \frac{m}{n\mu_k}},$$

(18)

and $\bar{\rho}_{kl} = 0$ when $k \neq l$.

Theorem 5 of [Onatski (2007)] gives that, if $\mu_k > c^{-1/2}$, then $Z_k$ converges in distribution to a mean-zero normal random variable. Furthermore, since $\mu_k \neq \mu_l$, Theorem 1 of
Onatski (2007) also yields that the vector \((W_{kl}, l = 1, \ldots, r)\) is asymptotically mean-zero multivariate normal with uncorrelated elements. Though not stated explicitly, Onatski’s proof shows that \(Z_k\) and \(W_{kl}\) are asymptotically uncorrelated for \(l = 1, \ldots, r\).

Next, by Lemma 4.3, we have \(\mathbb{E}(\tilde{v}^T s) = 0\) and

\[
\mathbb{E}(\tilde{v}^T s) = \mathbb{E}(\tilde{v}^T_{k} s_{V^\perp}) = \frac{1}{m-r} s^T_{V^\perp} s_{V^\perp}.
\]

By Lemma 4.3 and Equations (15) – (18), we obtain

\[
\mathbb{E}\{\hat{\mu}_k \cdot (\tilde{v}_k^T s)^2\} = \sum_{l=1}^{r} \{\hat{\mu}_k \cdot \tilde{\rho}_{kl}^2 + o(n^{-1/2})\} \cdot (v_l^T s)^2 + \{\hat{\mu}_k \cdot (1 - \tilde{\rho}_{kk}^2) + o(n^{-1/2})\} \cdot \frac{1}{m-r} s^T_{V^\perp} s_{V^\perp}.
\]

Therefore,

\[
-n \left( \mathbb{E}(\mu_k \cdot (v_k^T s)^2) - \mathbb{E}(\hat{\mu}_k \cdot (\tilde{v}_k^T s)^2) \right)
\]

\[
= -n (\mu_k - \hat{\mu}_k \tilde{\rho}_{kk}) (v_k^T s)^2 + o(n^{1/2} \cdot s^T_{V^\perp} s_V) + \mu_k (1 - \tilde{\rho}_{kk}^2) \frac{n}{m} s^T_{V^\perp} s_{V^\perp} + o(n^{-1/2} s^T_{V^\perp} s_{V^\perp})
\]

\[
= s^T s \cdot df_k(s) + o(n^{1/2} s^T_{V^\perp} s_V + n^{-1/2} s^T_{V^\perp} s_{V^\perp}).
\]

The result now follows by summing over \(k\).

**Corollary 4.6.** Under the hypothesis of Theorem 4.4, if we instead suppose \(\hat{r} \neq r\), then

\[
df(s) = \sum_{k=1}^{r} df_k(s) + \text{err}(s) + o \left( n^{1/2} \sum_{k=1}^{r} \frac{(v_k^T s)^2}{s^T s} + n^{-1/2} \right),
\]

where

\[
\text{err}(s) = \begin{cases} 
-n \cdot \sum_{k=\hat{r}+1}^{r} \mu_k \cdot \frac{(v_k^T s)^2}{s^T s} & \text{if } \hat{r} < r, \\
(\hat{r} - r) \left( 1 + \sqrt{\frac{n}{m}} \right)^2 & \text{if } \hat{r} > r.
\end{cases}
\]
Proof. Suppose \( \hat{r} < r \). Then by Equation (14) in the proof of Theorem 4.4 we obtain

\[
\begin{align*}
(s^T s) \cdot df(s) &= -n \cdot \sum_{k=1}^{\hat{r}} \{ \mu_k \cdot (v_k^T s)^2 \} - \sum_{k= \hat{r} + 1}^{r} \{ \mu_k \cdot (v_k^T s)^2 \} \\
&= n \cdot \sum_{k= \hat{r} + 1}^{r} \{ \mu_k \cdot (v_k^T s)^2 \}.
\end{align*}
\]

(21)

From (21) and Theorem 4.4, the claim follows for \( \hat{r} < r \).

Suppose to the contrary that \( \hat{r} > r \). A computation similar to the above yields

\[
\begin{align*}
(s^T s) \cdot err(s) &= n \cdot \sum_{k=r+1}^{\hat{r}} \{ \hat{\mu}_k \cdot (v_k^T s)^2 \}.
\end{align*}
\]

(22)

Theorem 1 of [Onatski (2007)] gives that

\[
\hat{\mu}_k = \left( 1 + \sqrt{\frac{m}{n}} \right)^2 + o_p(1)
\]

and \( \mathbb{E} ((v_k^T s)^2) = s^T s / m \). Summing over \( k \) yields the claim and the proof is finished. \( \Box \)

**Remark 4.7.** The requirement that \( \mu_k > c^{-1/2} \) in Theorem 4.4 and Corollary 4.6 is not artificial; there indeed is a phase transition in the asymptotic behavior of the eigenvalues at \( \mu_k = c^{-1/2} \) (see Baik et al. (2005) and Onatski (2007)). Consequently, Theorem 4.4 and Corollary 4.6 do not apply if some \( \mu_k \) is below the phase transition \( \mu_k \leq c^{-1/2} \). Following an argument similar to the proof of the \( \hat{r} > r \) case of Corollary 4.6, we conjecture that when \( \mu_k \leq c^{-1/2} \), the degree of freedom term \( df_k(s) \) should be defined as

\[
df_k(s) = \left( 1 + \sqrt{\frac{n}{m}} \right)^2 - n \mu_k \frac{(v_k^T s)^2}{s^T s}.
\]

5. SIMULATION STUDY

We perform a number of confirmatory simulations to verify the theory in Section 4. In these simulations, we vary the number of rows, \( n \), over the set \{5, 10, 50, 100\} and we vary the number of columns, \( m \), over the set \{5, 10, 50, 100, 500, 1000, 5000, 10000\}. We take the test direction \( s \) to be the first standard basis vector \( s = (1, 0, \ldots, 0) \) in \( \mathbb{R}^m \).
For a given set of simulation parameters, we perform 10,000 replicates of the following procedure:

1. Generate data from the model with \( r \) latent factors, \( Y = \sqrt{n} U D V^T + E \), where the elements of \( E \) are independent mean-zero normal variates with variance \( \sigma^2 = 1 \). Matrices \( U \) and \( V \) have orthonormal columns, while \( D \) is diagonal with \( (D)_{kk}^2 = \mu_k \) for \( k = 1, \ldots, r \). In each set of simulations, we fix \( D \) and \( V \), and we generate a uniform random \( U \) for each simulation replicate.

2. Fit the bilinear model with \( \hat{r} \) latent factors via least squares, \( \hat{Y} = \sqrt{n} \hat{U} \hat{D} \hat{V}^T \). Compute the residual matrix \( \hat{E} = Y - \hat{Y} \).

3. Calculate the residual sum of squares along the test direction, \( RSS(s) = s^T \hat{E}^T \hat{E} s \), and the observed degrees of freedom along this direction, \( df(s) = n - RSS(s)/\sigma^2 \).

We estimate \( E\{df(s)\} \) as the average value of \( df(s) \) over all replicates of the simulation; we also compute the standard error of the estimate via the central limit theorem. Finally, we compare the theoretical degrees of freedom estimate to the simulation-based estimate.

5.1 Noise Case
In the noise case, we simulate with no true latent factors (\( r = 0 \)), and we fit with one estimated latent factor using \( \hat{r} = 1 \). The theoretical degrees of freedom are computed from Theorem 4.2. As can be seen in Figure 4, the theory fits well with the simulations when the problem dimensions are large, say for \( n \geq 2500 \) (smaller problem dimensions are excluded from the figure).

5.2 Signal Case
For the signal case, the degrees of freedom depend on the signal strength and true factors. We simulate \( r = 1 \) true latent factor with signal strength \( \mu \) varying over the set \( \{1.0, 1.5, 3.0, 21.0\} \). We consider four choices of the factor loading vector \( v \):
Figure 4: Theoretical degrees of freedom for the null case (solid red line) agree with the empirical estimates (blue points and circles). Circle radius shows one standard error of the estimates along the $y$-axis.
Ones $v = (1/\sqrt{m}, \ldots, 1/\sqrt{m})$;

Basis $v = (1, 0, \ldots, 0)$;

Perp. Ones $v = (0, 1/\sqrt{m-1}, \ldots, 1/\sqrt{m-1})$;

Perp. Basis $v = (0, 1, 0, \ldots, 0)$.

In all cases, $v$ is a unit vector. In the “Perp.” cases, $v$ is orthogonal to the test direction $s$.

The asymptotic degrees of freedom in each of the four cases are as follows:

Ones

$$df(s) = \begin{cases} 
1 + n/m + \sigma^2/\mu & \text{if } \mu > \sigma^2\sqrt{m/n}, \\
(1 + \sqrt{n/m})^2 - (\mu/\sigma^2) \cdot \sqrt{n/m} & \text{otherwise}.
\end{cases}$$

Basis

$$\frac{df(s)}{n} = \begin{cases} 
1 - m\sigma^2/(n\mu) - m\sigma^4/(n\mu^2) & \text{if } \mu > \sigma^2\sqrt{m/n}, \\
-(\mu/\sigma^2) & \text{otherwise}.
\end{cases}$$

Perp. Ones, Perp. Basis

$$df(s) = \begin{cases} 
1 + (\sigma^2/\mu)^2 & \text{if } \mu > \sigma^2\sqrt{m/n}, \\
(1 + \sqrt{n/m})^2 & \text{otherwise}.
\end{cases}$$

In the “Basis” case, we study $df(s)/n$ instead of $df(s)$ so that the asymptotic limit depends on $n$ only through the ratio $n/m$. Figures 5–8 demonstrate that the asymptotic expressions agree with the theory, even for relatively small sample sizes.
Figure 5: Theoretical and empirical estimates for the signal case agree. Lighter hues correspond to weaker signal strengths.
Figure 6: When the true signal vector is equal to the test direction, the empirical results fit well with the theory for large sample sizes. Agreement is better when the signal strength is above the phase transition.
Figure 7: When the true signal vector is orthogonal to the test direction, degrees of freedom do not depend on $m/n$ when signal strength is above the phase transition.
Figure 8: Behavior is nearly identical for “Perp. Ones” and “Perp. Basis” signal vectors.
6. ESTIMATING DEGREES OF FREEDOM IN APPLICATIONS: A CONSERVATIVE ESTIMATOR

The main result of Section 4 is that the asymptotic degrees of freedom associated with the \( k \)th latent factor is given by

\[
\text{df}_k(s) = \begin{cases} 
  n \left( 1 - \frac{mc^2}{n\mu_k} - \frac{mc^4}{n\mu_k^2} \right) \left( \frac{v_k^Ts}{s^Ts} \right)^2 + \left( 1 + \frac{c^2}{\mu_k} \right) \left( 1 - \frac{v_k^Ts}{s^Ts} \right) & \text{if } \mu_k > \sigma^2 \frac{\sqrt{m/n}}{n}, \\
  (1 + \sqrt{n/m})^2 - n \frac{\mu_k}{v_k} \left( \frac{v_k^Ts}{s^Ts} \right)^2 & \text{otherwise}. 
\end{cases}
\] (23)

This result, while theoretically interesting, is not directly applicable to data analysis. For practical purposes, we need an estimate of \( \text{df}(s) \) which does not depend on unknown quantities.

A plug-in estimator (replacing population quantities \( \mu_k, \sigma^2 \), and \( v_k^Ts \) with the corresponding sample-based quantities) is likely to under-estimate \( \text{df}_k(s) \) since, almost surely, \( \hat{\mu}_k > \mu_k \) and \( \left( \hat{\sigma}_k^T v_k \right)^2 < 1 \). Under-estimating \( \text{df}_k(s) \) leads to smaller estimates of \( \sigma^2(s) \), which in turn leads to higher \( t \)-statistics and more false discoveries.

We propose a conservative estimator for \( \text{df}(s) \). First, from (23), we have the upper bound

\[
\text{df}_k(s) \leq n \left( 1 - \frac{mc^2}{n\mu_k} - \frac{mc^4}{n\mu_k^2} \right) \left( \frac{v_k^Ts}{s^Ts} \right)^2 + (1 + \sqrt{n/m})^2. 
\]

Next, we note that

\[
\left( \hat{\sigma}_k^T s \right)^2 \geq \tilde{\rho}_{kk}^2 \left( v_k^Ts \right)^2 + O_P(n^{-1/2}) \geq \left( 1 - \frac{mc^4}{n\mu_k^2} \right) \left( v_k^Ts \right)^2 + O_P(n^{-1/2}).
\]

Therefore, the estimator

\[
\text{df}_k(s) = n \frac{\left( \hat{\sigma}_k^Ts \right)^2}{s^Ts} + (1 + \sqrt{n/m})^2
\] (24)

is asymptotically greater than \( \text{df}_k(s) \).

Even though the estimator is conservative, the difference \( \text{df}_k(s) - \text{df}_k(s) \) is small
in regimes of practical interest, when $\mu_k$ is well above the phase transition, \textit{i.e.}, $\mu_k \gg \sigma^2 \sqrt{m/n}$.

7. APPLICATION TO AGEMAP

As an application of these methods, we fit the bilinear model to the AGEMAP dataset described in Section 2. For each gene, our goal is to assess the relationship between log activation and age after adjusting for observed and latent subject-specific covariates. For gene $j$, we take test direction $s_j = (I - H_Z)e_j$, where $e_j \in \mathbb{R}^m$ denotes the $j$th basis vector. Using a bilinear model to adjust for observed and latent subject-specific covariates, we perform a test on $[B^T s_j]_3$, the identifiable component of the age coefficient for gene $j$.

We first regress gene response on the observed covariates (subject age and sex; gene tissue type). An investigation of the residuals from this bilinear multiple regression fit reveals that two latent factors explain 51.3% of the residual variance (Table 1). After adding these two estimated latent factors to the regression model, there is no obvious low-dimensional structure in the residuals.

Table 1: \textbf{First Two Factors Explains Most of Residual Variance}. Residual variance explained by each principal component. A large proportion of the total variance is explained by the first two components.

| Factor | Resid. Var. % | Resid. Var. % | Factor | Resid. Var. % | Resid. Var. % | Factor | Resid. Var. % | Resid. Var. % |
|--------|---------------|---------------|--------|---------------|---------------|--------|---------------|---------------|
| 1      | 37.1          | 62.9          | 13     | 1.4           | 17.8          | 25     | 0.7           | 6.5           |
| 2      | 14.2          | 48.7          | 14     | 1.2           | 16.6          | 26     | 0.7           | 5.8           |
| 3      | 5.8           | 42.9          | 15     | 1.2           | 15.5          | 27     | 0.7           | 5.1           |
| 4      | 4.3           | 38.7          | 16     | 1.1           | 14.3          | 28     | 0.7           | 4.5           |
| 5      | 3.7           | 34.9          | 17     | 1.0           | 13.3          | 29     | 0.6           | 3.8           |
| 6      | 3.4           | 31.5          | 18     | 1.0           | 12.3          | 30     | 0.6           | 3.2           |
| 7      | 2.8           | 28.6          | 19     | 0.9           | 11.4          | 31     | 0.6           | 2.6           |
| 8      | 2.2           | 26.5          | 20     | 0.9           | 10.5          | 32     | 0.6           | 2.1           |
| 9      | 2.0           | 24.4          | 21     | 0.8           | 9.6           | 33     | 0.6           | 1.5           |
| 10     | 1.9           | 22.5          | 22     | 0.8           | 8.8           | 34     | 0.5           | 1.0           |
| 11     | 1.8           | 20.7          | 23     | 0.8           | 8.0           | 35     | 0.5           | 0.5           |
| 12     | 1.5           | 19.2          | 24     | 0.8           | 7.3           | 36     | 0.5           | 0.0           |

We obtain conservative degree of freedom estimate $\hat{df}_k(s_j)$ for the $\hat{r} = 2$ estimated
Figure 9: **Latent Factor Model Leads to Different Conclusions.** Gene-specific regression coefficient $t$ statistics for Age under the ordinary regression model and the latent factor model with $\hat{r} = 2$ estimated factors. There are 496 coefficients which are significant at level 0.001 in the latent factor model but not in the ordinary regression model; there is 1 coefficient significant at this level in the ordinary regression model but not the latent factor model.

latent factors using the estimator (24). We use the estimate to derive a gene-specific error variance estimate

$$\hat{\delta}^2(s_j) = \text{RSS}(s_j) / \{n - \hat{df}(s_j)\},$$

with $\hat{df}(s_j) = \sum_{k=1}^{\hat{r}} \hat{df}_k(s_j)$ and $n = N - p$. This, in turn, can be used to compute a test statistic for $[B^T s_j]_3$.

After adjusting for latent factors, there are 514 age coefficients out of 17,864 which are significant at level 0.001. Without the latent factor adjustment, we would find only 19 genes to be significant at that level. Figure 9 shows the test statistics from the model.
with no estimated factors (\( \hat{r} = 0 \)) and the model with (\( \hat{r} = 2 \)). For most genes (85%), adjusting for latent factors results in a larger test statistic. Adjusting for 2 latent factors uses between 2.18 and 2.99 degrees of freedom, depending on the gene.

8. DISCUSSION

We have shown how to adjust for latent sources of variability in multivariate regression problems by proposing a simple degrees of freedom assignment for estimated latent factors. Our estimate gives a principled alternative to ad-hoc approaches in common use. We have thus bridged the gap between theory and practice in this context by proposing a conservative estimate for the degrees of freedom and rigorously establishing the underpinning theory. Although our estimator is conservative, it is close to the exact theoretical value in regimes of common interest, with many responses and strong latent signals. Moreover, it is quite simple to apply, and thus ideal for routine use.

In order to perform a rigorous theoretical analysis, we have made two main simplifying assumptions. First, we have assumed that the regression errors are normally-distributed. Second, we have assumed that the noise covariance is a multiple of the identity. In light of many universality results in random matrix theory \cite{Pillai:2013,Benaych-Georges:2011}, the first assumption (normality) can likely be weakened. The second assumption is harder to tackle analytically, but we believe our results hold as long as the eigenvalues of the error covariance matrix are small relative to the latent signal strength. A rigorous analysis of the extent to which this assumption can be weakened is an area for further research.

APPENDIX

Proof of Lemma 4.1

By construction,

\[
\hat{E}^T \hat{E} = (Y - \hat{Y})^T (Y - \hat{Y}).
\]
Since the factors were estimated from the singular value decomposition of \( Y \), they are orthogonal to the residual matrix. That is, \( \hat{E}^T \hat{U} = 0 \) and \( \hat{E} \hat{V}^T = 0 \) and hence \( \hat{E}^T \hat{Y} = 0 \). Thus,

\[
\hat{E}^T \hat{E} = YY^T - 2Y^T \hat{Y} + \hat{Y}^T \hat{Y} \\
= YY^T - (\hat{Y} + \hat{E})^T \hat{Y} + \hat{Y}^T \hat{Y} \\
= YY^T - \hat{Y}^T \hat{Y} \\
= YY^T - n \hat{\theta} \hat{D}^2 \hat{V}^T
\]

and

\[
\text{RSS}(s) \equiv s^T YY^T s - n s^T \hat{\theta} \hat{D}^2 \hat{V}^T s. \quad (A.1)
\]

Now the result follows from expanding the terms, and using the identity

\[
s^T \hat{\theta} \hat{D}^2 \hat{V} s = \sum_{k=1}^{r} \mu_k \cdot (v_k^T s)^2
\]

along with an analogous expansion for \( s^T \hat{\theta} \hat{D}^2 \hat{V}^T s \).

**Proof of Lemma 4.3** Suppose that \( Y = UDV^T + E \), where the rows of \( E \) are independent mean-zero multivariate normal random vectors with covariance matrix \( \Sigma = \sigma^2 I \). Let \( Y = \hat{U} \hat{D} \hat{V}^T \) be the singular value decomposition of \( Y \). Set \( \mathbf{V}_1 = V \) and choose \( \mathbf{V}_2 \) such that \( \mathbf{V} = [\mathbf{V}_1 \quad \mathbf{V}_2] \) is an orthogonal matrix. The matrix \( \hat{\mathbf{V}} \) can be decomposed as \( \hat{\mathbf{V}} = \mathbf{V}_1 \hat{\mathbf{V}}_1 + \mathbf{V}_2 \hat{\mathbf{V}}_2 \), where \( \hat{\mathbf{V}}_l = \mathbf{V}_l^T \hat{\mathbf{V}} \) for \( l = 1, 2 \). The claim will follow if we show that the distribution of \( \hat{\mathbf{V}}_2 \) is invariant under multiplication on the left by any orthogonal matrix, i.e., if \( O \hat{\mathbf{V}}_2 \overset{d}{=} \hat{\mathbf{V}}_2 \) for every orthogonal \( O \).

Set \( E_l = E \mathbf{V}_l^T \) for \( l = 1, 2 \). Note that \( E_2 O^T \overset{d}{=} E_2 \). Write \( Y = (UD + E_1) \mathbf{V}_1^T + E_2 \mathbf{V}_2^T \). Set \( Y' = (UD + E_1) \mathbf{V}_1^T + E_2 O^T \mathbf{V}_2^T \) and let \( Y' = \hat{U}' \hat{D}' \hat{V}'^T \) be the singular value decomposition of \( Y' \). Since \( Y' \overset{d}{=} Y \), it must follow that \( \mathbf{V}_2^T \hat{V}' \overset{d}{=} \hat{\mathbf{V}}_2 \). In fact, \( \hat{\mathbf{V}}' = \mathbf{V}_1 \hat{\mathbf{V}}_1 + \mathbf{V}_2 O \hat{\mathbf{V}}_2 \) by
construction since $Y' (V_1 \hat{V}_1 + V_2 O \hat{V}_2) = Y \hat{V}$. Therefore, $V_2^T \hat{V} = O \hat{V}_2$ and thus $O \hat{V}_2 \overset{d}{=} \hat{V}_2$, finishing the proof.

ACKNOWLEDGEMENTS

The authors thank Art Owen for suggesting the research problem and Paul Bourgade for helpful discussions. Natesh Pillai is partially supported by the National Science Foundation under Grant No. DMS-1107070.

REFERENCES

Baik, J., Ben Arous, G., and Péché, S. (2005), “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices,” *Ann. Probab.*, 33, 1643–1697.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011), *Latent Variable Models and Factor Analysis: A Unified Approach*, Wiley, 3rd ed.

Benaych-Georges, F. and Nadakuditi, R. R. (2011), “The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices,” *Adv. Math.*, 227, 494–521.

Bock, R. D. and Gibbons, R. D. (1996), “High-dimensional multivariate probit analysis,” *Biometrics*, 52, 1183–1194.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Want, Q., and West, M. (2008), “High-dimensional sparse factor modeling: Applications in gene expression genomics,” *J. Am. Stat. Assoc.*, 103, 1438–1456.

Cochran, W. G. (1943), “The comparison of different scales of measurement for experimental results,” *Ann. Math. Stat.*, 14, 205–216.

Cornelius, P. L. and Seyedsadr, M. S. (1997), “Estimation of general linear-bilinear models for two-way tables,” *J. Stat. Comput. Simul.*, 58, 287–322.
Crossa, J., Cornelius, P. L., and Yan, W. (2002), “Biplots of linear-bilinear models for studying crossover genotype × environment interaction,” *Crop Sci.*, 42, 619–633.

dos S. Dias, C. T. and Krzanowski, W. J. (2003), “Model selection and cross validation in additive main effect and multiplicative interaction models,” *Crop Sci.*, 43, 865–873.

Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, Wiley, 3rd ed.

Elrod, T. and Keane, M. P. (1995), “A factor-analytic probit model for representing the market structure in panel data,” *J. Marketing Res.*, 32, 1–16.

Fisher, R. A. and Mackenzie, W. A. (1923), “Studies in crop variation. II. The manurial response of different potato varieties,” *J. Agr. Sci*, 13, 311–320.

Freeman, G. H. (1973), “Statistical methods for the analysis of genotype-environment interactions,” *Heredity*, 31, 339–354.

Friguet, C., Kloareg, M., and Causeur, D. (2009), “A factor model approach to multiple testing under dependence,” *J. Am. Stat. Assoc.*, 104, 1406–1415.

Gabriel, K. R. (1978), “Least squares approximation of matrices by additive and multiplicative models,” *J. Roy. Stat. Soc. B*, 40, 186–196.

Gabriel, R. K. (1998), “Generalised bilinear regression,” *Biometrika*, 85, 689–700.

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

Hoff, P. (2007), “Model averaging and dimension selection for the singular value decomposition,” *J. Am. Stat. Assoc.*, 102, 674–685.

Leek, J. T. and Storey, J. D. (2008), “A general framework for multiple testing dependence,” *Proc. Natl. Acad. Sci. USA*, 105, 18718–18723.
Mandel, J. (1969), “The partitioning of interaction in analysis of variance,” *J. Res. Nat. Bur. Stand.*, 73B, 309–328.

— (1971), “A new analysis of variance model for non-additive data,” *Technometrics*, 13, 1–18.

Onatski, A. (2007), “Asymptotics of the principal components estimator of large factor models with weak factors and IID Gaussian noise,” Unpublished manuscript.

Pillai, N. S. and Yin, J. (2013), “Universality of covariance matrices,” *Ann. Appl. Probab.*, to appear.

Song, X.-Y. and Lee, S.-Y. (2001), “Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations,” *Brit. J. Math. Stat. Psy.*, 54, 237–263.

Stock, J. H. and Watson, M. W. (2002), “Macroeconomic forecasting using diffusion indexes,” *J. Bus. Econ. Stat.*, 20, 147–162.

Sun, Y., Zhang, N. R., and Owen, A. B. (2012), “Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data,” *Ann. Appl. Stat.*, 6, 1664–1688.

Tukey, J. W. (1962), “The future of data analysis,” *Ann. Math. Stat.*, 33, 1–67.

van Eeuwijk, F. A. (1995), “Multiplicative interaction in generalized linear models,” *Biometrics*, 51, 1017–1032.

West, M. (2003), “Bayesian factor regression models in the “Large p, Small n” paradigm,” in *Bayesian Statistics 7*, eds. Bernardo, J. M., Bayarri, M. J., Dawid, A. P., Berger, J. O., Heckerman, D., Smith, A. F. M., and West, M., Oxford University Press, pp. 723–732.

Williams, E. J. (1952), “The interpretation of interactions in factorial experiments,” *Biometrika*, 39, 65–81.
Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. R. (1988), “On the limit of the largest eigenvalue of the large dimensional sample covariance matrix,” *Probab. Theory Rel.*, 78, 509–521.

Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeraratna, 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., Wood, W. H., Zonderman, A. B., Kim, S. K., and Becker, K. G. (2007), “AGEMAP: A gene expression database for aging in mice,” *PLoS Genet.*, 3, 2326–2337.