On the distribution of the left singular vectors of a random matrix and its applications

E. Bura [a,*], R. Pfeiffer [b]

[a] *Department of Statistics, The George Washington University, Washington, DC 20052, USA*

[b] *Biostatistics Branch, National Cancer Institute, DCEG, 6120 Executive Blvd, EPS/8030, Bethesda, MD 20892-7244, USA*

Abstract

In several dimension reduction techniques, the original variables are replaced by a smaller number of linear combinations. The coefficients of these linear combinations are typically the elements of the left singular vectors of a random matrix. We derive the asymptotic distribution of the left singular vectors of a random matrix that has a normal limit distribution. This result is then used to develop a Wald-type test for testing variable importance in Sliced Inverse Regression (SIR) and Sliced Average Variance Estimation (SAVE), two popular sufficient dimension reduction methods.

Key words: Dimension Reduction, SAVE, SIR, Singular value decomposition.

* Corresponding author.
Email addresses: ebura@gwu.edu (E. Bura), pfeiffer@mail.nih.gov (R. Pfeiffer).

1 The Asymptotic Distribution of the Singular Vectors of an Asymptotically Normal Random Matrix

Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \in R^p$ be $p \times 1$ independent identically distributed random vectors. Let $\widehat{\boldsymbol{\Omega}}_n$ be a $p \times q$ dimensional random matrix that is a function of the $\mathbf{Z}_i$, $i = 1, \ldots, n$. Assume that $\widehat{\boldsymbol{\Omega}}_n$ is asymptotically normally distributed as

$$n^{1/2} \operatorname{vec}(\widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega}) \Longrightarrow N_{pq}(0, \mathbf{V}). \tag{1}$$

Let $d = \operatorname{rank}(\boldsymbol{\Omega})$ with $d \leq \min(p, q)$. The singular value decomposition (SVD) of $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega} = \mathbf{U}^T \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{R}, \tag{2}$$

where the orthogonal matrix $\mathbf{U}^T = (\mathbf{U}_1, \mathbf{U}_0)$ is of order $p \times p$ with $\mathbf{U}_1$: $p \times d$, $\mathbf{U}_0$: $p \times (p - d)$, $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a diagonal matrix containing the descending singular values of $\boldsymbol{\Omega}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$, and $\mathbf{R}^T = (\mathbf{R}_1, \mathbf{R}_0)$ is orthogonal with $\mathbf{R}_1$: $q \times d$, $\mathbf{R}_0$: $q \times (q - d)$. The d left singular vectors $\mathbf{U}_1 = (\mathbf{U}_1^{(1)}, \ldots, \mathbf{U}_d^{(1)})$ of $\boldsymbol{\Omega}$ that correspond to its d non-zero singular values $\lambda_1 \geq \ldots \geq \lambda_d$ span $\boldsymbol{\Omega}$. The projections of $\mathbf{Z}$ onto $\operatorname{span}(\boldsymbol{\Omega})$ are $(Z_1^*, \ldots, Z_d^*) = (\mathbf{U}_1^{(1)T}\mathbf{Z}, \ldots, \mathbf{U}_d^{(1)T}\mathbf{Z})$. Analogously, the SVD of $\widehat{\boldsymbol{\Omega}}_n$ is

$$\widehat{\boldsymbol{\Omega}}_n = \widehat{\mathbf{U}}^T \begin{pmatrix} \widehat{\mathbf{D}}_1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{D}}_0 \end{pmatrix} \widehat{\mathbf{R}} \tag{3}$$

with $\widehat{\mathbf{U}}^T = (\widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_0)$, $\widehat{\mathbf{R}}^T = (\widehat{\mathbf{R}}_1, \widehat{\mathbf{R}}_0)$, where the partition conforms to the SVD of $\boldsymbol{\Omega}$ in (2). The matrices $\widehat{\mathbf{D}}_1 = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$ of dimension $d \times d$, and $\widehat{\mathbf{D}}_0 = \operatorname{diag}(\hat{\lambda}_{d+1}, \ldots, \hat{\lambda}_p)$ of dimension $(p-d) \times (p-d)$ have the singular values of $\widehat{\mathbf{\Omega}}_n$ in descending order on the diagonal.

Theorem 1 *Let $d = rank(\mathbf{\Omega})$ and assume that $\lambda_d > \epsilon > 0$, i.e. the minimum positive singular value of $\mathbf{\Omega}$ is well separated from zero. Then as $n \to \infty$,*

(a) $\widehat{\mathbf{D}}_1 \xrightarrow{p} \mathbf{D}$ *and* $\widehat{\mathbf{D}}_0 \xrightarrow{p} \mathbf{0}$.

(b) $\widehat{\mathbf{R}}_1 \xrightarrow{p} \mathbf{R}_1$ *and* $\widehat{\mathbf{U}}_1 \xrightarrow{p} \mathbf{U}_1$.

where $\mathbf{D}, \mathbf{R}_1, \mathbf{U}_1$ *are defined in (2), and* $\widehat{\mathbf{D}}_0, \widehat{\mathbf{D}}_1, \widehat{\mathbf{R}}_1, \widehat{\mathbf{U}}_1$ *are defined in (3).*

Proof: (a) Let $\widehat{\mathbf{\Omega}}_n\widehat{\mathbf{\Omega}}_n^T = \sum_{i=1}^p \hat{\lambda}_i^2 \widehat{\mathbf{U}}_i \widehat{\mathbf{U}}_i^T$ be the spectral decomposition of the symmetric $p \times p$ matrix $\widehat{\mathbf{\Omega}}_n\widehat{\mathbf{\Omega}}_n^T$. As $\widehat{\mathbf{\Omega}}_n$ is consistent for $\mathbf{\Omega}$ from (1), the continuous transformation $\widehat{\mathbf{\Omega}}_n\widehat{\mathbf{\Omega}}_n^T$ is consistent for $\mathbf{\Omega}\mathbf{\Omega}^T$, and thus consistency of the eigenvalues follows , i.e. $\hat{\lambda}_i^2 \xrightarrow{p} \lambda_i^2$. Tyler (1981, Lemma 2.1) showed that if the eigenvalues of a random matrix have a limit in probability, then this property is preserved under continuous transformations. Hence, since $\lambda_i \geq 0$, $\hat{\lambda}_i \xrightarrow{p} \lambda_i$, for all $i = 1, \ldots, p$, which proves result 1.

(b) The next arguments follow the proof of Proposition 1 in Ratsimalahelo (2003). From the SVD of $\mathbf{\Omega}$ we have $\mathbf{U}\mathbf{\Omega}\mathbf{R}^T = \mathbf{D}$. Also, if we let $\widehat{\mathbf{D}} = \text{diag}(\widehat{\mathbf{D}}_1, \widehat{\mathbf{D}}_2)$, from (3) we have

$$\widehat{\mathbf{U}}\widehat{\mathbf{\Omega}}_n\widehat{\mathbf{R}}^T = \widehat{\mathbf{D}} \tag{4}$$

From (1), $\widehat{\mathbf{\Omega}}_n$ is root n consistent for $\mathbf{\Omega}$. Hence $\widehat{\mathbf{\Omega}}_n = \mathbf{\Omega} + \epsilon\mathbf{W}$, where $\epsilon = n^{-1/2}$ and $\mathbf{W}$ is some matrix conformable with $\widehat{\mathbf{\Omega}}_n$ (Kato, 1982). Then $\mathbf{U}\widehat{\mathbf{\Omega}}_n\mathbf{R}^T = \mathbf{D} + \epsilon\mathbf{U}\mathbf{W}\mathbf{R}^T = \mathbf{D} + \epsilon\mathbf{C}$ where

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{U}_1^T\mathbf{W}\mathbf{R}_1 & \mathbf{U}_1^T\mathbf{W}\mathbf{R}_0 \\ \mathbf{U}_0^T\mathbf{W}\mathbf{R}_1 & \mathbf{U}_0^T\mathbf{W}\mathbf{R}_0 \end{pmatrix}$$

Let $\mathbf{P}$ and $\mathbf{Q}$ be orthogonal transformations up to second order in ϵ, given by

$$\mathbf{P} = \mathbf{I}_p + \epsilon\mathbf{F} + O_p(\epsilon^2) = \mathbf{I}_p + \epsilon \begin{pmatrix} \mathbf{0} & \mathbf{F}_{12} \\ -\mathbf{F}_{21} & \mathbf{0} \end{pmatrix} + O_p(\epsilon^2)$$

with $\mathbf{F}_{12} = -\mathbf{D}^{-1}\mathbf{C}_{21}^T$ and

$$\mathbf{Q} = \mathbf{I}_p + \epsilon\mathbf{B} + O_p(\epsilon^2) = \mathbf{I}_p + \epsilon \begin{pmatrix} \mathbf{0} & \mathbf{B}_{12} \\ -\mathbf{B}_{21} & \mathbf{0} \end{pmatrix} + O_p(\epsilon^2)$$

with $\mathbf{B}_{12} = -\mathbf{D}^{-1}\mathbf{C}_{12}$. Then,

$$\mathbf{P}^T\mathbf{U}\widehat{\mathbf{\Omega}}_n\mathbf{R}^T\mathbf{Q} = \begin{pmatrix} \mathbf{D} + \epsilon\mathbf{C}_{11} & \mathbf{0} \\ \mathbf{0} & \epsilon\mathbf{C}_{22} \end{pmatrix} + O_p(\epsilon^2) \tag{5}$$

Also, $\mathbf{U}^T\mathbf{P} = (\mathbf{U}_1 - \epsilon\mathbf{U}_0\mathbf{F}_{21}, \mathbf{U}_0 + \epsilon\mathbf{U}_1\mathbf{F}_{12}) + O_p(\epsilon^2)$ and $\mathbf{R}^T\mathbf{Q} = (\mathbf{R}_1 - \epsilon\mathbf{R}_0\mathbf{B}_{21}, \mathbf{R}_0 + \epsilon\mathbf{R}_1\mathbf{B}_{12}) + O_p(\epsilon^2)$. Comparing (4) with (5) and $\mathbf{U}^T\mathbf{P}$ and $\mathbf{R}^T\mathbf{Q}$, we obtain $\widehat{\mathbf{U}}_1 = \mathbf{U}_1 - \epsilon\mathbf{U}_0\mathbf{F}_{21} + O_p(\epsilon^2)$ and $\widehat{\mathbf{R}}_1 = \mathbf{R}_1 - \epsilon\mathbf{R}_0\mathbf{B}_{21} + O_p(\epsilon^2)$. As both matrices $\mathbf{U}_0$ and $\mathbf{R}_0$ are orthogonal and non-random, they are of order $O_p(1)$. Since $\lambda_d > \epsilon > 0$ by assumption, $\mathbf{D}^{-1}$ is bounded and also of order $O_p(1)$. The latter implies that both $\mathbf{F}_{21}$ and $\mathbf{B}_{21}$ are of the same order. Thus, $\widehat{\mathbf{U}}_1 = \mathbf{U}_1 - O_p(\epsilon) + O_p(\epsilon^2) = \mathbf{U}_1 + O_p(\epsilon)$ and $\widehat{\mathbf{R}}_1 = \mathbf{R}_1 - O_p(\epsilon) + O_p(\epsilon^2) = \mathbf{R}_1 + O_p(\epsilon)$, so that both $\widehat{\mathbf{U}}_1$ and $\widehat{\mathbf{R}}_1$ are root-n consistent for $\mathbf{U}_1$ and $\mathbf{R}_1$, respectively. □

By (2), $\mathbf{\Omega} = \mathbf{U}_1\mathbf{D}\mathbf{R}_1^T$, hence $\mathbf{U}_1 = \mathbf{\Omega}\mathbf{R}_1\mathbf{D}^{-1}$, or equivalently, $\text{vec}(\mathbf{U}_1) = (\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\ \text{vec}(\mathbf{\Omega})$. By the multivariate version of Slutsky's theorem and (1) we obtain

$$n^{1/2}(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p) \operatorname{vec}(\widehat{\mathbf{\Omega}}_n - \mathbf{\Omega}) \Longrightarrow N_{pd}(\mathbf{0}, (\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p) \tag{6}$$

Theorem 1 and (6) lead to the following corollary.

Corollary 1 *The matrix* $\widehat{\mathbf{U}}_1$ *of left singular vectors in (3) is asymptotically normal with*

$$n^{1/2}\ vec(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) \Longrightarrow N_{pd}(\mathbf{0}, \mathbf{\Sigma}_U), \tag{7}$$

where $\mathbf{\Sigma}_U = (\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p)$.

Proof: Let $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{D}}_1\widehat{\mathbf{R}}_1^T\mathbf{R}_1\mathbf{D}^{-1}$. Then

$$n^{1/2} \operatorname{vec}(\widehat{\mathbf{U}}_1\widehat{\mathbf{D}}\widehat{\mathbf{R}}_1^T\mathbf{R}_1\mathbf{D}^{-1} - \mathbf{U}_1) = n^{1/2} \operatorname{vec}(\widehat{\mathbf{U}}_1(\widehat{\mathbf{\Psi}} - \mathbf{I}_p)) + n^{1/2} \operatorname{vec}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1)$$

Theorem 1 implies $\widehat{\mathbf{\Psi}} = \mathbf{I}_p + O_p(n^{-1/2})$ and $\widehat{\mathbf{U}}_1 = \mathbf{U}_1 + O_p(n^{-1/2})$. Focusing on the first term we obtain

$$\begin{aligned} n^{1/2} \operatorname{vec}((\mathbf{U}_1 + O_p(n^{-1/2}))O_p(n^{-1/2})) &= n^{1/2} \operatorname{vec}(\mathbf{U}_1 O_p(n^{-1/2}) + O_p(n^{-1})) \\ &= n^{1/2} \operatorname{vec}(O_p(n^{-1/2}) + O_p(n^{-1})) = n^{1/2} \operatorname{vec}(O_p(n^{-1/2})) = O_p(1) \end{aligned}$$

since $\mathbf{U}_1 = O_p(1)$. Consequently, $n^{1/2} \operatorname{vec}(\widehat{\mathbf{U}}_1\widehat{\mathbf{D}}\widehat{\mathbf{R}}_1^T\mathbf{R}_1\mathbf{D}^{-1} - \mathbf{U}_1) = n^{1/2} \operatorname{vec}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) + O_p(1)$, which together with (6) yield (7). □

It is straightforward to show that $\mathbf{\Sigma}_U$ is consistently estimated by the sample estimate $\widehat{\mathbf{\Sigma}}_U = (\widehat{\mathbf{D}}_1^{-1}\widehat{\mathbf{R}}_1^T \otimes \mathbf{I}_p)\widehat{\mathbf{V}}(\widehat{\mathbf{R}}_1\widehat{\mathbf{D}}_1^{-1} \otimes \mathbf{I}_p)$. Also, when $\mathbf{V}$, the asymptotic covariance matrix of $\widehat{\mathbf{\Omega}}_n$ in (1), is positive definite, so is $\mathbf{\Sigma}_U$. To see this we apply result A4.4 in Seber (1977) to obtain

$$\operatorname{rank}(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes I_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p) = \operatorname{rank}(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p) = \operatorname{rank}(\mathbf{R}_1)p = dp,$$

since $\mathbf{R}_1$ is of order $q \times d$ with orthonormal columns.

2 Application: Assessing variable importance in sufficient dimension reduction methods

We now use the results from section 1 to develop a test for variable contribution in sufficient dimension reduction (SDR) regression. We first briefly review SDR methods.

2.1 Dimension Reduction Methods based on Inverse Regression

Let $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ denote the predictors and Y the outcome variable, either continuous or categorical, in a regression problem. Without loss of generality we assume that the predictors are standardized with $\mathrm{E}(\mathbf{Z}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{Z}_i) = \mathbf{I}_p$, $i = 1, \ldots, p$. Sufficient dimension reduction is based on the idea that $\mathbf{Z}$ can be replaced by a lower-dimensional projection $\mathbf{P}_S\mathbf{Z}$ without loss of information about the conditional distribution of $Y|\mathbf{Z}$. $\mathbf{P}_S$ is the orthogonal projection onto the vector space S in the usual inner product. No pre-specified model for $Y|\mathbf{Z}$ is required. The intersection of all subspaces $S \subset R^p$ that satisfy $F(Y|\mathbf{Z}) = F(Y|\mathbf{P}_S\mathbf{Z})$, where $F(\cdot|\cdot)$ is the conditional distribution function of Y given the second argument, is the *central* subspace $S_{Y|\mathbf{Z}}$ (Cook, 1996, 1998). The dimension $d = \dim(S_{Y|\mathbf{Z}})$ is called the structural dimension of the regression of Y on $\mathbf{Z}$ and can take on any value in the set $\{0, 1, \ldots, p\}$. When $d < p$, the dimension of the regression is reduced. If $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d)$ is a basis for $S_{Y|\mathbf{Z}}$, $P_{\boldsymbol{\eta}}\mathbf{Z}$, or equivalently, the d linear combinations $\boldsymbol{\eta}^T\mathbf{Z} = (\boldsymbol{\eta}_1^T\mathbf{Z}, \ldots, \boldsymbol{\eta}_d^T\mathbf{Z})$ contain all the information in $\mathbf{Z}$ about Y.

The estimation of the central subspace is based on finding a kernel matrix $\mathbf{\Omega}$ so that $\text{span}(\mathbf{\Omega}) \subset S_{\mathbf{Y}|\mathbf{Z}}$. This can be done by first moment methods such as Sliced Inverse Regression (SIR) (Li, 1991) with $\mathbf{\Omega}_{SIR} = \text{Cov}(\text{E}(\mathbf{Z}|\mathbf{Y}))$, or second moment methods such as Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991) with $\mathbf{\Omega}_{SAVE} = \text{E}(\mathbf{I}_p - \text{Cov}(\mathbf{Z}|\mathbf{Y}))^2$. In particular, Cook and Lee (1999) showed that $\mathbf{\Omega}_{SIR} = (\text{E}(\mathbf{Z}|Y = 0) - \text{E}(\mathbf{Z}|Y = 1))$, and $\mathbf{\Omega}_{SAVE} = (\text{E}(\mathbf{Z}|Y = 0) - \text{E}(\mathbf{Z}|Y = 1), \text{Var}(\mathbf{Z}|Y = 0) - \text{Var}(\mathbf{Z}|Y = 1))$, for a binary outcome Y. SAVE is the most comprehensive SDR method (Cook and Lee, 1998). The condition needed for Ω_{SIR} to span subspaces of the central dimension reduction subspace is that $\text{E}(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})$ be linear. For SAVE, in addition to the linearity condition for SIR, the conditional variance $\text{Var}(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})$ needs to be constant.

The projections of the predictors $\mathbf{Z}$ onto the space $\text{span}(\mathbf{\Omega})$ spanned by the SAVE or SIR kernel matrix are $\mathbf{Z}^* = (Z_1^*, \ldots, Z_d^*) = (\mathbf{U}_1^{(1)T}\mathbf{Z}, \ldots, \mathbf{U}_d^{(1)T}\mathbf{Z})$, where $\mathbf{U}_1^{(1)}, \ldots, \mathbf{U}_d^{(1)}$ are the d columns of $\mathbf{U}_1$ in (2). $\mathbf{Z}^*$ contains part or all the information contained in the predictor vector $\mathbf{Z}$ for regressing Y on $\mathbf{Z}$.

To estimate the kernel matrices for discrete outcomes Y, the conditional expectations and covariances in $\mathbf{\Omega}_{SIR} = \text{Cov}(\text{E}(\mathbf{Z}|\mathbf{Y}))$ and $\mathbf{\Omega}_{SAVE} = \text{E}(\mathbf{I}_p - \text{Cov}(\mathbf{Z}|\mathbf{Y}))^2$ are computed over the discrete levels of Y. If Y is continuous, it is replaced by a discretized version $\tilde{Y}$, resulting from partitioning the observed range of Y into H fixed, non-overlapping slices. The sample moment estimate of the SIR kernel matrix of the standardized predictors is $\widehat{\mathbf{\Omega}}_{SIR} = \sum_{h=1}^{H} f_h \bar{\mathbf{Z}}_h \bar{\mathbf{Z}}_h^T$, where $\bar{\mathbf{Z}}_h$ denotes the intraslice mean of the standardized predictors and $f_h = n_h/n$ is the fraction of observations falling into slice h (Li, 1991). For SAVE, the sample moment estimate is $\widehat{\mathbf{\Omega}}_{SAVE} = \sum_{h=1}^{H} f_h (\mathbf{I}_p - \widehat{\mathbf{\Sigma}}_{z|h})^2$, where $\widehat{\mathbf{\Sigma}}_{z|h}$ is the intraslice covariance estimate of $\mathbf{Z}$ (Cook and Weisberg, 1991).

To infer the structural dimension d, a test statistic based on the singular values of $\widehat{\Omega}_{SIR}$ or $\widehat{\Omega}_{SAVE}$ is typically used (Li, 1991; Cook, 1998; Cook and Lee, 1999; Bura and Cook, 2001). The estimation of d is carried out by sequentially testing $H_0 : d = k$ against $H_a : d > k$, starting at $k = 0$, which corresponds to independence of Y and $\mathbf{Z}$, and adding unit increments to k until H_0 cannot be rejected at a preset α level.

2.2 Asymptotic distribution of $\widehat{\boldsymbol{\Omega}}_{SIR}$ *and* $\widehat{\boldsymbol{\Omega}}_{SAVE}$

Recall that for continuous Y, $\widehat{\boldsymbol{\Omega}}_{SIR} = \sum_{h=1}^{H} f_h \bar{\mathbf{Z}}_h \bar{\mathbf{Z}}_h^T$, where $\bar{\mathbf{Z}}_h$ denotes the intra slice mean of the standardize predictors and $f_h = n_h/n$ is the fraction of observations falling into slice h. Let $\hat{\mathbf{Z}}_n = (\bar{\mathbf{Z}}_1\sqrt{f_1}, \ldots, \bar{\mathbf{Z}}_H\sqrt{f_H})$, to obtain $\widehat{\boldsymbol{\Omega}}_n = \hat{\mathbf{Z}}_n \hat{\mathbf{Z}}_n^T$. By the multivariate central limit theorem,

$$n^{1/2}\text{vec}(\widehat{\boldsymbol{\Omega}}_{SIR} - \boldsymbol{\Omega}_{SIR}) \Longrightarrow N_{pd}(\mathbf{0}, \mathbf{V})$$

where $\mathbf{V} = \sum_h p_h \mathbf{V}_h$, is a $p^2 \times p^2$ matrix with $\mathbf{V}_h = \text{Var}[\, \text{vec}(\mathbf{Z} - \boldsymbol{\mu}_{z|h})(\mathbf{Z} - \boldsymbol{\mu}_{z|h})^T | Y = h]$ and $p_h = P(Y \text{ falls in slice } h)$.

For continuous Y, $\widehat{\boldsymbol{\Omega}}_{SAVE} = \sum_{h=1}^{H} f_h (\mathbf{I}_p - \widehat{\boldsymbol{\Sigma}}_{z|h})^2$, where $\widehat{\boldsymbol{\Sigma}}_{z|h}$ is the intraslice covariance of $\mathbf{Z}$. If we let $\mathbf{K}_n = ((\mathbf{I}_p - \widehat{\boldsymbol{\Sigma}}_1)\sqrt{f_1}, \ldots, (\mathbf{I}_p - \widehat{\boldsymbol{\Sigma}}_H)\sqrt{f_H})$ denote a $p \times (pH)$ matrix, we have $\widehat{\boldsymbol{\Omega}}_{SAVE} = \widehat{\mathbf{K}}_n \widehat{\mathbf{K}}_n^T$. By the multivariate central limit theorem,

$$n^{1/2}\text{vec}(\widehat{\mathbf{K}}_n - \mathbf{K}) \Longrightarrow N_{pd}(\mathbf{0}, \mathbf{V}_K). \tag{8}$$

The $p^2 H \times p^2 H$ matrix $\mathbf{V}_K$ is provided by Yin (2005). Let $\boldsymbol{\Delta} = \frac{\partial(\mathbf{K}\mathbf{K}^T)}{\partial \mathbf{K}} =$

$(\frac{\partial(\mathbf{K}\mathbf{K}^T)}{\partial k_{ij}})$, $i = 1, \ldots, p$, $j = 1, \ldots, m$, denote the $p^2 \times (p^2 H)$ matrix of partial derivatives of $\mathbf{K}\mathbf{K}^T$. By applying the multivariate version of the delta-method and (8) we obtain

$$n^{1/2}\text{vec}(\widehat{\mathbf{K}}_n\widehat{\mathbf{K}}_n^T - \mathbf{K}\mathbf{K}^T) = n^{1/2}\text{vec}(\widehat{\mathbf{\Omega}}_{SAVE} - \mathbf{\Omega}_{SAVE}) \Longrightarrow N(\mathbf{0}, \mathbf{\Delta}\mathbf{V}_K\mathbf{\Delta}^T).$$

For binary Y the asymptotic normality of $\widehat{\mathbf{\Omega}}_{SIR}$ and $\widehat{\mathbf{\Omega}}_{SAVE}$ is easily obtained from results in the appendix of Cook and Lee (1999).

2.3 Testing for variable contribution to the SAVE and SIR projections

Variable selection is important in model building, especially when the dimension of input variables is large, as the inclusion of many noise variables may impact the performance of a model. In this section, we evaluate the contributions of predictors to the SAVE or SIR linear combinations. For example, if the predictor Z_k does not contain any information for predicting Y, then $\mathbf{U}_{1k}^{(1)}, \ldots, \mathbf{U}_{dk}^{(1)}$, i.e. its contribution to the SDR predictors $(Z_1^*, \ldots, Z_d^*) = (\mathbf{U}_1^{(1)T}\mathbf{Z}, \ldots, \mathbf{U}_d^{(1)T}\mathbf{Z})$ should be zero.

Given the asymptotic normality of the estimated kernel matrices, $\widehat{\mathbf{\Omega}}_{SIR}$ and $\widehat{\mathbf{\Omega}}_{SAVE}$, we can use Corollary 1 to derive a formal test for the coefficients of $\mathbf{U}_1$. First, though, we need to overcome a technical difficulty, namely that the asymptotic distributions of singular values and their corresponding eigenvectors depend on the multiplicity of the singular values of the limit matrix.

The singular value decomposition provides orthonormal bases for the range (column) and null space of $\mathbf{\Omega}$. That is, $\mathcal{S}(\mathbf{\Omega}) = \mathcal{S}(\mathbf{U}_1) = \mathcal{S}(\mathbf{\Omega}\mathbf{\Omega}^T)$ and $\mathcal{N}(\mathbf{\Omega}) = \mathcal{S}(\mathbf{R}_0)$. Also, $\mathbf{\Omega}\mathbf{\Omega}^T = \sum_{i=1}^d \lambda_i^2 \mathbf{U}_i\mathbf{U}_i^T = \sum_{i=1}^d \lambda_i^2 \mathbf{P}_i$, where $\mathbf{P}_i = \mathbf{U}_i\mathbf{U}_i^T$

is the projection onto the subspace spanned by $\mathbf{U}_i$. The d left singular vectors $\mathbf{U}_1, \ldots, \mathbf{U}_d$ of $\boldsymbol{\Omega}$ that correspond to its d non-zero singular values $\lambda_1 \geq \ldots \geq \lambda_d$ are the d eigenvectors corresponding to the d eigenvalues $\lambda_1^2 \geq \ldots \geq \lambda_d^2$ of $\boldsymbol{\Omega}\boldsymbol{\Omega}^T$. If λ_i has multiplicity one, then $\text{rank}(\mathbf{P}_i) = 1$. If the multiplicity of λ_i is $m_i > 1$, then $\mathbf{P}_i = \sum_{j=1}^{m_i} \mathbf{U}_j^{(i)} \mathbf{U}_j^{(i)T}$ where $\mathbf{U}_j^{(i)}$ are the m_i eigenvectors that correspond to the multiple eigenvalue λ_i^2 and $\text{rank}(\mathbf{P}_i) = m_i$. Thus, unless the singular value corresponding to an eigenvector has multiplicity one, eigenvectors are not themselves distinguishable and we must focus on their corresponding eigenspaces. When the asymptotic distribution of the estimator matrix is normal, the limit matrix is symmetric and the multiplicities of its eigenvalues are known, Tyler (1981) computed the asymptotic distribution of the estimated eigenspace $\widehat{\mathbf{P}}_i = \sum_{j=1}^{m_i} \widehat{\mathbf{U}}_j^{(i)} \widehat{\mathbf{U}}_j^{(i)T}$ to be normal. Dossou-Gbete and Pousse (1991) generalized this result and computed the asymptotic distribution for all eigenelements (eigenvalues, eigenvectors and eigenspaces) of a sequence of random symmetric matrices also assuming known multiplicities of the eigenvalues of the limit matrix.

In practice, the multiplicity of singular values is unknown. We circumvent this obstacle by using (7) to test a hypothesis of the form $\mathbf{C}\text{vec}(\mathbf{U}_1) = \mathbf{0}$ for a prespecified $r \times pd$ matrix $\mathbf{C}$ of zeroes and ones. The rank of $\mathbf{C}$, r, equals the number of the elements of $\mathbf{U}_1$ set to zero. The entries of $\mathbf{C}$ set elements of $\text{vec}(\mathbf{U}_1)$ to zero to allow the assessment of the contribution of any given subset of the individual predictors $Z_1, \ldots, Z_p$ to the SDR predictors $Z_1^*, \ldots, Z_d^*$ simultaneously. For example, if $d = 2$ and we wish to test whether X_1 has no contribution to $\mathbf{Z}^*$, then $C_{11} = C_{2,p+1} = 1$, and all other entries are zero.

Relation (7) yields

$$n^{1/2}\,\mathbf{C}\;\mathrm{vec}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) \Longrightarrow N_r(\mathbf{0}, \mathbf{C}\boldsymbol{\Sigma}_U\mathbf{C}^T) \tag{9}$$

The next lemma provides consistent estimates for the asymptotic variance-covariance structure in (9).

Lemma 1 *Let* $\mathbf{A} = \mathbf{C}\boldsymbol{\Sigma}_U\mathbf{C}^T = \mathbf{C}(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p)\mathbf{C}^T$*. The sample estimate* $\widehat{\mathbf{A}} = \mathbf{C}\widehat{\boldsymbol{\Sigma}}_U\mathbf{C}^T = \mathbf{C}(\widehat{\mathbf{D}}_1^{-1}\widehat{\mathbf{R}}_1^T \otimes \mathbf{I}_p)\widehat{\mathbf{V}}(\widehat{\mathbf{R}}_1\widehat{\mathbf{D}}_1^{-1} \otimes \mathbf{I}_p)\mathbf{C}^T$ *is consistent for* $\mathbf{A}$*. If* $\mathbf{V}$ *is positive definite then* $\widehat{\mathbf{A}}^{-1}$ *is consistent for* $\mathbf{A}^{-1}$*. If* $\mathbf{V}$ *is positive semi-definite then* $\widehat{\mathbf{A}}^{+}$ *is consistent for* $\mathbf{A}^{+}$ *where* $+$ *indicates the Moore-Penrose inverse.*

Proof: As $\widehat{\mathbf{V}}$ is a moment-based estimator it is consistent for $\mathbf{V}$. By Theorem 1, $\widehat{\mathbf{D}}^{-1}$ and $\widehat{\mathbf{R}}_1$ are consistent, and hence $\mathbf{C}\widehat{\boldsymbol{\Sigma}}_U\mathbf{C}^T = \mathbf{C}(\widehat{\mathbf{D}}_1^{-1}\widehat{\mathbf{R}}_1^T \otimes \mathbf{I}_p)\widehat{\mathbf{V}}(\widehat{\mathbf{R}}_1\widehat{\mathbf{D}}_1^{-1} \otimes \mathbf{I}_p)\mathbf{C}^T$ is also consistent. Similarly, the inverse is a continuous function hence the second result when $\mathbf{V}$ is positive definite. Also then, as the matrix $\mathbf{R}_1$ is orthogonal with full column rank, $(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p) \otimes (\mathbf{R}_1\mathbf{D} \otimes \mathbf{I}_p) = \mathbf{I}_d \otimes \mathbf{I}_p$. Hence

$$[(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p)]^{+} = (\mathbf{D}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}^{+}(\mathbf{R}_1\mathbf{D} \otimes \mathbf{I}_p)$$

Since $\widehat{\mathbf{V}}$ is consistent for $\mathbf{V}$, the Moore-Penrose generalized inverse of $\widehat{\mathbf{V}}$ is consistent for the Moore-Penrose generalized inverse of $\mathbf{V}$ (Lemma 2.2 in Tyler, 1981). The latter and Theorem 1 yield $(\widehat{\mathbf{D}}_1\widehat{\mathbf{R}}_1^T \otimes \mathbf{I}_p)\widehat{\mathbf{V}}^{+}(\widehat{\mathbf{R}}_1\widehat{\mathbf{D}}_1 \otimes \mathbf{I}_p) \xrightarrow{p} (\mathbf{D}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}^{+}(\mathbf{R}_1\mathbf{D}_z \otimes \mathbf{I}_p)$ or, equivalently, $\widehat{\mathbf{A}}^{+} \xrightarrow{p} \mathbf{A}^{+}$. □

A general Wald-type test for H_0: $\mathbf{C}\mathrm{vec}(\mathbf{U}_{z1}) = 0$ is given in the next theorem.

Theorem 2 *Let* $\mathbf{C}$ *be an* $r \times pd$ *matrix of rank* r*,* $\boldsymbol{\theta} = vec(\mathbf{C}vec(\mathbf{U}_1))$*, and* $\widehat{\boldsymbol{\theta}} = vec(\mathbf{C}vec(\widehat{\mathbf{U}}_1))$*, both* $rpd \times 1$ *vectors. Also, let* $\mathbf{A} = \mathbf{C}\boldsymbol{\Sigma}_U\mathbf{C}^T$ *and* $\widehat{\mathbf{A}} =$

$\mathbf{C}\widehat{\boldsymbol{\Sigma}}_U\mathbf{C}^T$ *be its sample-based estimate. Suppose* $\boldsymbol{\theta} = \mathbf{0}$*. Then,*

a. If $\mathbf{V}$ *is positive definite,*

$$\widehat{T} = n\,\hat{\boldsymbol{\theta}}^T\widehat{\mathbf{A}}^{-1}\hat{\boldsymbol{\theta}} \Longrightarrow \chi^2(r) \tag{10}$$

where $\widehat{\mathbf{A}}$ *is a consistent estimate of* $\mathbf{A}$*.*

b. If $\mathbf{V}$ *is positive semidefinite with rank*$(\mathbf{V}) \geq r$ *and rank*$(\widehat{\mathbf{V}}) \xrightarrow{p}$ *rank*$(\mathbf{V})$*,*

$$\widehat{T} = n\,\hat{\boldsymbol{\theta}}^T\widehat{\mathbf{A}}^{+}\hat{\boldsymbol{\theta}} \Longrightarrow \chi^2(r) \tag{11}$$

where $\widehat{\mathbf{A}}^{+}$ *is a consistent estimate of* $\mathbf{A}^{+}$*.*

Proof: a. Since $\mathbf{V}$ is positive definite, Lemma 1 yields $\text{rank}(\mathbf{A}) = r = \text{rank}(\mathbf{C})$. This in conjunction with (9) yield (10) (see section 2 in Moore (1977)).

b. The asymptotic covariance matrix of $\widehat{\mathbf{U}}_1$ has rank

$$\begin{aligned}\text{rank}(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p) &\leq \min\{\text{rank}(\mathbf{V}), \text{rank}(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\}\\ &= \min\{\text{rank}(\mathbf{V}, pd)\}\end{aligned}$$

If $\text{rank}(\mathbf{V}) \geq pd$, then $(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p)$ is full rank and hence positive definite which implies that $\text{rank}(\mathbf{A}) = \text{rank}(C) = r$. Since $\mathbf{A}$ is invertible, this yields (11) as in a. Note that in this case $\widehat{\mathbf{A}}^{+} = \widehat{\mathbf{A}}^{-1}$ and $\mathbf{A}^{+} = \mathbf{A}^{-1}$. If $\text{rank}(\mathbf{V}) < pd$, the matrix $(\mathbf{D}^{-1}\mathbf{R}_1^T \otimes \mathbf{I}_p)\mathbf{V}(\mathbf{R}_1\mathbf{D}^{-1} \otimes \mathbf{I}_p)$ is positive semidefinite (see ex. 3.4 in Rao, 1973) with rank that of $\mathbf{V}$. Since $\text{rank}(\mathbf{V}) \geq r$ we have $\text{rank}(\mathbf{A}) = r$. Lemma 1 and Theorem 2 in Moore (1977) now yield (11). □

As an aside observation, we note that the SAVE asymptotic test for dimension by Cook and Lee (1999) implicitly assumes that $\mathbf{V}$ is positive definite. In effect, for the test to be valid one has to assume that $\text{rank}(\mathbf{V}) \geq \text{rank}(\mathbf{R}_0 \otimes \mathbf{U}_0) =$

$(p - d + 1)(p - d)$.

We also want to point out that the asymptotic framework for the test of marker contribution assumes that the structural dimension d is known. Future work will entail adjustments of the test for estimating the dimension.

References

Bura, E. and Cook, R. D., 2001. Extending SIR: The Weighted Chi-Square Test. J. Amer. Statist. Assoc. 96, 996–1003.

Cook, R. D., 1998. Regression Graphics: Ideas for studying regressions through graphics. Wiley, New York.

Cook, R. D. and Lee, H., 1999. Dimension reduction in regressions with a binary response. J. Amer. Statist. Assoc. 94, 1187–1200.

Cook, R. D. and Li, B., 2002. Dimension reduction for the conditional mean in regression. Ann. Statist. 30, 455–474.

Cook, R. D. and Weisberg, S., 1991. Discussion of Li (1991). J. Amer. Statist. Assoc. 86, 328–332.

Dossou-Gbete, S. and Pousse, A., 1991. Asymptotic study of eigenelements of a sequence of random selfadjoint operators. Statistics 22, 479–491.

Kato, T., 1982. A Short Introduction to Perturbation Theory for Linear Operators. Springer-Verlag, New York.

Li, K. C., 1991. Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86, 316–342.

Magnus, J. R. and Neudecker, H., 1979. The commutation matrix: some properties and applications. Ann. Statist. 7, 381–394.

Moore, D. S., 1977. Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. J. Amer. Statist. Assoc. 72, 131–137.

Muirhead, R. J., 1982. Aspects of Multivariate Statistical Theory. Wiley, New York.

Rao, C. R., 1973. Linear Statistical Inference and Its Applications, 2nd Edition. Wiley, New York.

Ratsimalahelo, Z., 2003. Rank test based on matrix perturbation theory. From *Economics Working Paper Archive at WUSTL*, http://econpapers.repec.org/paper/wpawuwpem/0306008.htm.

Seber, G. A. F., 1977. Linear Regression Analysis. Wiley, New York.

Tyler, D. E., 1981. Asymptotic inference for eigenvectors. Ann. Statist. 9, 725-736.

Yin, X., 2005. Tests of Dimension with Sliced Average Variance Estimation. Technical Report Stat 2005-13, from *Technical Reports, Department of Statistics, University of Georgia*, http://www.stat.uga.edu/research/techreports.html