Dimension reduction in regression without matrix inversion

BY R. DENNIS COOK
School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
dennis@stat.umn.edu

BING LI AND FRANCESCA CHIAROMONTE
Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.
bing@stat.psu.edu chiaro@stat.psu.edu

SUMMARY
Regressions in which the fixed number of predictors $p$ exceeds the number of independent observational units $n$ occur in a variety of scientific fields. Sufficient dimension reduction provides a promising approach to such problems, by restricting attention to $d < n$ linear combinations of the original $p$ predictors. However, standard methods of sufficient dimension reduction require inversion of the sample predictor covariance matrix. We propose a method for estimating the central subspace that eliminates the need for such inversion and is applicable regardless of the $(n, p)$ relationship. Simulations show that our method compares favourably with standard large sample techniques when the latter are applicable. We illustrate our method with a genomics application.

Some key words: Central subspace; $\Sigma$-envelope; Singularity of sample covariance.

1. INTRODUCTION

Methods for studying the conditional distribution of a response $Y$, given a predictor vector $X \in \mathbb{R}^p$, include classical linear regression, generalized additive models, projection pursuit (Huber, 1985), inverse regression methods (Li, 1991; Duan & Li, 1991; Cook & Ni, 2005a) and minimum average variance estimation (Xia et al., 2002), as well as machine learning techniques such as support vector machines (Cortes & Vapnik, 1995; Cristianini & Shawe-Taylor, 2000). With the exception of support vector methodology, these methods include some type of dimension reduction for $X$ (see for example, Li & Duan, 1989). In particular, sufficient dimension reduction provides a paradigm for studying intrinsically low-dimensional regressions without requiring a pre-specified parametric or semiparametric model.

A subspace $\text{span}(\alpha)$, the span of the columns of the $p \times r$ matrix $\alpha$, is said to be a dimension reduction subspace if $Y \| X | \alpha^T X$; that is, $X$ and $Y$ are independent conditioning on $\alpha^T X$. The intersection of all such subspaces is itself a dimension reduction subspace for most regressions of practical interest, and is called the central subspace, $S_{Y|X} \subseteq \mathbb{R}^p$. If $\text{span}(\beta) = S_{Y|X}$, then $Y \| X | \beta^T X$, and the regression of $Y$ on $X$ is informationally equivalent to the regression of $Y$ on the sufficient predictor $W = \beta^T X$, in the sense that $Y|X$ and $Y|W$ have the same distribution. In this way, the foundation for sufficient dimension
reduction is in the same spirit as Fisher’s (1922) classical notion of sufficiency; see Cook (1998) for additional background on the central subspace.

There are several reasons why sufficient dimension reduction may be useful in practice, including the possibilities of mitigating the effects of collinearity among the predictors, facilitating model specification by allowing visualization of the regression in low dimensions, and providing a relatively small set of predictors on which to base prediction or interpretation. In addition, our experience with sufficient dimension reduction methods agrees with recent comments by Hand (2006) on classifier technology: regressions based on low-dimensional, global projective summaries may perform in the future as well as or better than local methods whose output is well tuned to the training data.

Existing methods for inference about $S_{Y|X}$ are based on traditional reasoning that requires $n$ to be large relative to $p$. Operational issues occur when $n < p$ or $n \approx p$, because inference methods typically require the inversion of the sample version $\hat{\Sigma}$ of the $p \times p$ predictor covariance matrix $\Sigma$. Since the rank of $\hat{\Sigma}$ is at most $\min(n, p)$, the inversion is impossible if $n < p$, and unstable if $n \approx p$.

We propose a general approach that allows many methods for estimating $S_{Y|X}$ to be adapted to regressions in which $n$ does not dominate $p$. The approach, which requires computation of powers of $\hat{\Sigma}$ instead of $\hat{\Sigma}^{-1}$, is similar to that of iterative Hessian transformations (Cook & Li, 2002, 2004), and our method encompasses partial least squares as a special case. The dimension $d$ of the central subspace is assumed to be selected a priori as a modelling parameter that reflects the likely complexity of the regression, similar in spirit to the selection of the degree of a polynomial mean function.

The following notation will be used repeatedly in our exposition. For positive integers $p$ and $q$, $\mathbb{R}^{p \times q}$ stands for the class of all matrices of dimension $p \times q$. For a matrix $A$ in $\mathbb{R}^{p \times p}$ and a subspace $S$ of $\mathbb{R}^p$, $AS$ stands for the set $\{Ax : x \in S\}$. For an arbitrary matrix $B$ of $p$ rows, $\text{span}(B)$ denotes the subspace of $\mathbb{R}^p$ spanned by the columns of $B$. For a symmetric, positive definite matrix $\Sigma$, the inner product in $\mathbb{R}^p$ defined by $\langle a, b \rangle_\Sigma = a^T \Sigma b$ will be referred to as the $\Sigma$ inner product in $\mathbb{R}^p$; when $\Sigma = I$, the $p \times p$ identity matrix, this inner product will be called the usual inner product. A projection relative to the $\Sigma$ inner product is a projection operator in the inner product space $(\mathbb{R}^p, \langle \cdot, \cdot \rangle_\Sigma)$. If $B \in \mathbb{R}^{p \times q}$, then the projection on to $\text{span}(B)$ relative to the $\Sigma$ inner product has the matrix representation $B(B^T \Sigma B)^+ B^T \Sigma$, where $^+$ indicates the Moore-Penrose inverse. For a symmetric matrix $A$, $A > 0$ means $A$ is positive definite; $A \geq 0$ means $A$ is positive semidefinite.

2. Population-level results and estimation rationale

2.1. Seed matrices

We define a population seed as any matrix $\nu$ such that $\text{span}(\nu) \subseteq \Sigma S_{Y|X}$ and possibly $\text{span}(\nu) = \Sigma S_{Y|X}$. To avoid complications, we will assume the latter coverage condition to hold. The coverage condition, which was reviewed by Cook & Ni (2005a), is common in sufficient dimension reduction and may be reasonable in many applications. Additional discussion of coverage is given in §2.4. If $\Sigma$ is invertible, a seed can be used to obtain the matrix $\beta = \Sigma^{-1} \nu$ whose columns span the central subspace. For instance, the seed for ordinary least squares is the $p \times 1$ covariance vector $\nu = \text{cov}(X, Y)$, and when $d = 1$ the central subspace can be obtained as the span of the least squares vector $\beta = \Sigma^{-1} \nu$. Some methods for estimating $S_{Y|X}$, including least squares, require that $E(X|\beta^T X)$ be a linear function of $\beta^T X$. Hall & Li (1993) showed that, as $p$ increases with $d$ fixed, this linearity
Dimension reduction in regression 571

condition holds to a good approximation in many problems; see also Diaconis & Freedman (1984). Additional discussion of this condition, which is typically regarded as mild, was given by Cook & Ni (2005a). Under the linearity condition,

$$E(X|Y = y) - E(X) \in \Sigma S_{Y|X}$$

(1)

for all y. It follows that $\text{cov}(E(X|Y))$ qualifies as a seed matrix. The conditioning required by (1) cannot normally be performed in practice unless Y is discrete or categorical. A standard practice with a continuous response (Li, 1991) is first to partition the range of Y into h slices indexed by $s = 1, \ldots, h$ and then average (1) over the values of Y in a slice. This yields

$$\xi_s \equiv E(X|J_s(Y) = 1) - E(X) \in \Sigma S_{Y|X}, \quad s = 1, \ldots, h,$$

(2)

where $J_s(Y) = 1$ if Y is in slice s and $J_s(Y) = 0$ otherwise. It follows that the $p \times h$ matrix $\xi = (\xi_1, \ldots, \xi_h)$ qualifies as a seed matrix. However, since h can be large, it will be useful to replace $\xi$ with the $p \times d$ matrix $v$ whose columns are the left singular vectors of $\xi$ corresponding to its d largest singular values. This involves no loss of dimension because $\dim(\text{span}(\xi)) = \dim(\text{span}(v)) = d$. Cook & Ni (2005b) recently proposed a method for recovering intra-slice information from methods based on (2). Their proposal is based on the relation $\text{cov}(X, Y J_s) = \Sigma S_{Y|X}, s = 1, \ldots, h$, which holds under the linearity condition. Let $\zeta$ be the $p \times h$ matrix with columns $\zeta_s = \text{cov}(X, Y J_s)$. Then, following reasoning similar to that given above, the columns of another $p \times d$ seed matrix $v$ are the left singular vectors of $\zeta$ corresponding to its d largest singular values.

For these and other sufficient dimension reduction techniques, a $p \times d$ sample version $\hat{v}$ of the seed can be obtained readily from the sample moments of $(X, Y)$, regardless of the relationship between n and p. Accordingly, we assume that $v \in \mathbb{R}^{p \times d}$. Restricting the column dimension of the seed to be the pre-selected dimension d of $\Sigma S_{Y|X}$ is not necessary for our developments, but it facilitates the methodology when n does not dominate p.

Using $\hat{\Sigma}^{-1}$ to transform the estimated seed so that it spans an estimate of the central subspace may be problematic, unless n dominates p. Our proposal to avoid inverting $\hat{\Sigma}$ provides an alternative way of achieving this transformation.

2.2. $\Sigma$-envelopes

Suppose we know a subspace $\mathcal{M}$ of $\mathbb{R}^p$ that contains $S_{Y|X}$. Then, projecting $\Sigma^{-1}v$ on to $\mathcal{M}$ with respect to the $\Sigma$ inner product results in the column space of $\Sigma^{-1}v$ itself, which is the central subspace we desire. To be more specific, if the columns of R form a basis for $\mathcal{M}$ then $\Sigma^{-1}v = R(R^T \Sigma R)^{-1}R^T v$, and consequently $\Sigma^{-1}$ is not required. Provided that $\dim(\mathcal{M})$ is small relative to n, this greatly ameliorates the inaccuracy incurred in estimating $\Sigma$ using $n < p$ observations.

The first step is therefore to identify a space $\mathcal{M}$ large enough to enclose $S_{Y|X}$, but small enough to allow reasonable estimation based on the available data. Moreover, $\mathcal{M}$ should be chosen so that its identification does not involve the inversion of $\Sigma$. To achieve this, we iteratively transform the seed matrix $v$ by the variance matrix $\Sigma$:

$$R_u \equiv (v, \Sigma v, \ldots, \Sigma^{u-1} v), \quad u = 1, 2, \ldots$$

(3)

As u increases, the $p \times ud$ matrices in (3) span a nondecreasing sequence of nested subspaces which, because of the spectral structure of $\Sigma$, leads us exactly to the type of parsimonious encapsulation of $S_{Y|X}$ that we are seeking. Consider the spectral
decomposition \( \Sigma = \sum_{j=1}^{q} \lambda_j P_j \), \( j = 1, \ldots, q \), where \( \lambda_1 > \cdots > \lambda_q > 0 \) are the distinct nonzero eigenvalues of \( \Sigma \), and \( P_j \) is the orthogonal projection operator on to the \( j \)th eigenspace of \( \Sigma \) relative to the \( \Sigma \) inner product. For an orthogonal projection operator whose range space is an eigenspace of \( \Sigma \), it makes no difference if the operator is defined relative to the \( \Sigma \) inner product or the usual inner product. We define \( P_j \) relative to the \( \Sigma \) inner product only for consistency with other projections employed in our developments.

Next, let \( K \subseteq \{1, \ldots, q\} \) be the indices of the eigenspaces that are not orthogonal to \( S_{Y|X} \), so that \( P_j S_{Y|X} \neq \{0\} \) for \( j \in K \), and let \( k \leq q \) be the cardinality of \( K \). We define the \( \Sigma \)-envelope of \( S_{Y|X} \) as

\[
\mathcal{M}_{Y|X} = \oplus_{j=1}^{q} P_j S_{Y|X} = \oplus_{j \in K} P_j S_{Y|X}.
\]

Clearly \( S_{Y|X} \subseteq \mathcal{M}_{Y|X} \).

Recall that a subspace \( S \) of \( \mathbb{R}^p \) is an invariant subspace of \( \Sigma \) if \( \Sigma S \subseteq S \). This \( S \) is a reducing subspace of \( \Sigma \) if, in addition, \( \Sigma S^\perp \subseteq S^\perp \). Since \( \Sigma \) is symmetric, any invariant subspace of \( \Sigma \) is also a reducing subspace. It can be shown that \( \mathcal{M}_{Y|X} \) equals the intersection of all reducing subspaces of \( \Sigma \) that contain \( S_{Y|X} \). In other words, \( \mathcal{M}_{Y|X} \) is the smallest subspace that contains \( S_{Y|X} \) and that conforms to the eigenstructure of \( \Sigma \), so that it can be constructed without inverting the latter.

The relationship between the spaces \( \text{span}(R_u) \) defined by (3) and the \( \Sigma \)-envelope \( \mathcal{M}_{Y|X} \) is summarized by the following theorem.

**Theorem 1.** Suppose that \( \Sigma \) is positive definite with \( q \) distinct nonzero eigenvalues, \( k \) of which correspond to eigenspaces not orthogonal to \( S_{Y|X} \). Then there exists an integer \( 1 \leq \tilde{u} \leq k \) such that \( \text{span}(R_u) \) is strictly increasing until \( u = \tilde{u} \), and settles upon \( \mathcal{M}_{Y|X} \) thereafter:

\[
\text{span}(R_1) \subset \text{span}(R_2) \subset \cdots \subset \text{span}(R_{\tilde{u}}) = \mathcal{M}_{Y|X} = \text{span}(R_{\tilde{u}+1}) = \text{span}(R_{\tilde{u}+2}) \cdots
\]

where \( \subset \) indicates strict inclusion.

All proofs appear in the Appendix.

Since \( S_{Y|X} \subseteq \mathcal{M}_{Y|X} \), an obvious corollary of Theorem 1 is that the subspaces \( \text{span}(R_u) \), \( u = 1, 2, \ldots \), will grow to contain the central subspace after at most \( \tilde{u} \) iterations; that is, there exists \( u^* \leq \tilde{u} \leq k \) such that \( S_{Y|X} \subseteq \text{span}(R_u) \) for \( u \geq u^* \). Moreover, the theorem states that the sequence \( R_u \) accomplishes the encapsulation of \( S_{Y|X} \) in a parsimonious way, because at each iteration it retrieves new directions in \( \mathcal{M}_{Y|X} \), until no such direction remains to be retrieved.

### 2.3. Capturing \( S_{Y|X} \)

To move our development forward, we introduce two additional matrix sequences,

\[
B_u = P_{R_u} \Sigma^{-1} v, \quad \Delta_u = Q_{B_u} B_{u+1}, \quad u = 1, 2, \ldots
\]

where \( P_{(*)} \) and \( Q_{(*)} \) indicate, respectively, the orthogonal projections on to the span of the argument matrix and its complement, relative to the \( \Sigma \) inner product. The column spaces of the \( B_u \)'s are projections of \( S_{Y|X} \) on to \( \text{span}(R_u) \). Once \( \text{span}(R_u) \) becomes large enough to contain \( S_{Y|X} \), \( \text{span}(B_u) \) will coincide with \( S_{Y|X} \) and settle on it thereafter: \( \text{span}(B_u) = S_{Y|X} \) for \( u \geq u^* \). Interestingly, \( \dim(\text{span}(B_u)) \) does not change as the underlying \( \text{span}(R_u) \)'s expand, as Theorem 2 states.
THEOREM 2. The space $\text{span}(B_u) = P_{R_u} S_{Y|X}$ has dimension $d$ for all $u = 1, 2, \ldots$.

One can therefore think of the $\text{span}(B_u)$’s as a sequence of $d$-dimensional subspaces moving within progressively larger linear windows, the $\text{span}(R_u)$’s, on the predictor space, until such windows become large enough that $\text{span}(B_u)$ can settle upon $S_{Y|X}$.

The $\Delta_u$’s in (5) capture the dislocations between contiguous projections, and they provide a way of evaluating whether or not the $\text{span}(B_u)$’s have stopped moving. Since both $\text{span}(B_{u+1})$ and $\text{span}(B_u)$ have dimension $d$, $\text{span}(B_{u+1}) = \text{span}(B_u)$ if and only if $\Delta_u = 0$. With the help again of the $\Sigma$ inner product, one measure of the dislocation is

$$F_u = \text{tr}(\Delta_u^T \Sigma \Delta_u).$$

Clearly, $F_u > 0$ if and only if $\text{span}(B_{u+1}) \neq \text{span}(B_u)$, which in turn implies that $S_{Y|X}$ has not yet been reached. In addition, we have following result.

THEOREM 3. The space $\text{span}(B_u) = S_{Y|X}$ if and only if $F_v = 0$ for all $v > u$.

As a result of the special structures of $B_u$ and $B_{u+1}$, Lemma 1 shows that the increment $\Delta_u$ defined in (5) can be further simplified, which will be useful in later developments.

LEMMA 1. The increment $\Delta_u = B_{u+1} - B_u$.

Since $R_u^T \Sigma R_u > 0$ when $u \leq \tilde{u}$ and $R_u^T \Sigma R_u \geq 0$ otherwise, the inverse of $R_u^T \Sigma R_u$ can be used in the computation of $B_u$ when $u \leq \tilde{u}$, but a generalized inverse is required when $u > \tilde{u}$. As long as $ud$, the dimension of $R_u^T \Sigma R_u$, remains small compared to $n$ and $u$ is not appreciably larger than $\tilde{u}$, these inversions should cause no difficulty in practice. Direct inversion of $\Sigma$ is not required. Consequently, sample versions $\hat{B}_u$ and $\Delta_u$ of our matrices can be produced easily and tracked along the iterations. Eventually, we estimate the central subspace as $\hat{S}_{Y|X} = \text{span}(\hat{B}_{u^*})$, where $u^*$ is taken to be the smallest index after which the statistic

$$\hat{f}_u = n \hat{F}_u = n \text{tr}\left( \hat{\Delta}_u^T \Sigma \hat{\Delta}_u \right)$$

is consistently small; see §3.

If we choose the seed $v = \text{cov}(X, Y)$ then our procedure coincides with partial least squares (Helland, 1990), with the difference that the termination index $u^*$ is estimated based on the specific objective of reconstructing a one-dimensional central subspace. Indeed, the developments described in this article, and the performance of our iterative method on actual and simulated data, see §§4 and 5, provide some insight into the effectiveness of partial least squares.

2.4. Choice of $d$

The results in §§2-2 and 2-3 make no use of the special role that $S_{Y|X}$ plays relative to $Y|X$. In fact, we can use the same rationale to estimate any subspace for which a seed is available. This has implications for the choice of $d$, because in applications we may or may not have $d = \dim(S_{Y|X})$. Suppose that we selected $d < \dim(S_{Y|X})$, and that correspondingly $\beta = \Sigma^{-1} v$ spanned a proper subset $S_{Y|X}$. Then our iterative procedure would result in an estimate of a proper subset of $S_{Y|X}$, namely $\text{span}(\beta)$ itself. Similarly, if $d > \dim(S_{Y|X})$ and $\text{span}(\beta) \supset S_{Y|X}$, we would estimate a proper superset of $S_{Y|X}$. Also, if $d > \dim(S_{Y|X})$ but $\text{span}(\beta) = S_{Y|X}$, then we would still end up with an estimate of the central subspace. We continue to assume that the selected $d$ is equal to $\dim(S_{Y|X})$, unless stated otherwise.
A simple example may help fix ideas. Decompose \((X, Y)\) into \(X|Y\) and \(Y\), and suppose the data follow the inverse model \((X|Y = y) = \mu + \beta Y + \sigma \epsilon\), where \(\beta \in \mathbb{R}^p\) with \(\|\beta\| = 1\), \(\gamma\) is a scalar-valued function of \(y\), and the error \(\epsilon\) is independent of \(Y\) and normally distributed with mean 0 and variance matrix \(I_p\). For this regression, \(S_{Y|X} = \text{span}(\beta)\), with \(d = 1\). The seed for ordinary least squares is \(v = \text{cov}(X, Y) = c\beta\), with \(c = \text{cov}(\gamma, Y)\). Let \(\tau^2 = \text{var}(\gamma)\). Then the following can be verified straightforwardly:

\[
\Sigma = (I_p - \beta \beta^T)\sigma^2 + \beta \beta^T (\sigma^2 + \tau^2),
\]

the columns of \(R_u\) are \(v(\sigma^2 + \tau^2)^j\) for \(j = 0, \ldots, u - 1\), and \(\text{span}(B_u) = \text{span}(\beta) = S_{Y|X}\) for all \(u\).

In this example, \(u^* = k = 1\) and the central subspace is found after the first iteration, because it is spanned by the seed itself. It follows that transforming the seed by \(\Sigma^{-1}\) is not required, and consequently that estimation of \(\Sigma^{-1}\) is not necessary. If the model were known this fact could be found analytically. As implied by the forms of \(R_u\) and \(B_u\), the methodology proposed here allows us to make use of this essential structure of the regression, without knowing the model a priori.

3. Choice of \(u^*\)

Our next task is to determine how to choose an estimator \(\hat{u}^*\) of the termination index \(u^*\). Accurate estimation of the termination index is not crucial. Since the inclusion \(S_{Y|X} \subseteq \text{span}(R_u)\) is guaranteed for \(u \geq u^*\), overestimation of the index, i.e. \(\hat{u}^* > u^*\), will preserve the effectiveness of our iterative approach, provided that \(\hat{u}^* d\) is small compared to \(n\).

When \(d = 1\) or 2, as may often be the case in practice, it may be useful to examine plots of \(Y_i\) versus the estimated sufficient predictor \(\hat{W}_{u,i} = \hat{B}_u^TX_i\), for \(u = 1, 2, \ldots, n\), as illustrated in §4. Application-specific numerical summaries of these plots could also be used. For instance, if \(Y\) is continuous, each plot could be summarized by using the residual sum of squares from a nonparametric fit of \(Y_i\) on \(\hat{W}_{u,i}\), and \(\hat{u}^*\) selected to be near the value of \(u\) at which the residual sum of squares stabilizes. A different index, such as an error rate from crossvalidation, might be required if \(Y\) is binary or categorical.

Lacking an application-specific measure, \(\hat{u}^*\) might be chosen around the value of \(u\) at which the \(d\) canonical correlations between \(\hat{W}_{u,i}\) and \(\hat{W}_{u+1,i}\) stabilize and remain near 1. Alternatively, the statistic \(\hat{f}_u\) introduced in (7) is simply a comparison of estimated sufficient predictor values for adjacent values of \(u\), \(\hat{f}_u = \sum_{i=1}^n \|\hat{W}_{u+1,i} - \hat{W}_{u,i}\|^2\) computed using the usual inner product, and is proposed as a generic measure of agreement. In the remainder of this section we provide a calibration for \(\hat{f}_u\) by considering hypotheses of the form

\[
H_{0,u} : \text{span}(B_{u+1}) = \text{span}(B_u) \quad \text{versus} \quad H_{1,u} : \text{span}(B_{u+1}) \neq \text{span}(B_u), \quad u = 1, 2, \ldots
\]

In particular, we use a large-sample approximation to derive reference distributions for \(\hat{f}_u\). Although approximations based on sample-size asymptotics are generally problematic when \(n\) does not dominate \(p\), in our context we have found them to be reasonably accurate, even when \(n < p\). Recalling that formal inference about \(u^*\) is not required, we have found that the large-sample \(p\)-values for \(H_{0,u}\) provide a useful way of selecting \(\hat{u}^*\), that might then be confirmed by using graphical methods or canonical correlations.

Let \(E_u(\cdot)\) denote the sample averaging operator and let \((\cdot)^\circ\) denote the Fréchet derivative operator with respect to the joint distribution of \((X, Y)\).
THEOREM 4. If $\Delta_u = 0$, then

$$\hat{\Delta}_u^T \hat{\Sigma} \hat{\Delta}_u = E_n (\Sigma^{1/2} \Delta_u^*)^T E_n (\Sigma^{1/2} \Delta_u^*) + O_p (n^{-3/2}),$$

where the leading term is $O_p(n^{-1})$.

The full expression of the Fréchet derivative $\Delta_u^*$, which is rather lengthy but straightforward from the point of view of code implementation, is given in the Appendix.

Let $\text{vec}(A)$ denote the vector obtained by stacking the columns of a generic matrix $A$. Then, under the null hypothesis $H_{0,u}$,

$$\hat{f}_u = n \operatorname{tr}(\hat{\Delta}_u^T \hat{\Sigma} \hat{\Delta}_u) = n E_n \{ \text{vec}^T (\Sigma^{1/2} \Delta_u^*) \} E_n \{ \text{vec}(\Sigma^{1/2} \Delta_u^*) \} + O_p (n^{-1/2}).$$

Since $n^{1/2} E_n \{ \text{vec}(\Sigma^{1/2} \Delta_u^*) \}$ converges in distribution to a normal with mean 0 and $pd \times pd$ covariance matrix $\Omega_u = E \{ \text{vec}(\Sigma^{1/2} \Delta_u^*) \text{vec}^T(\Sigma^{1/2} \Delta_u^*) \}$, we can create an approximate reference distribution for $\hat{f}_u$ by linearly combining $pd$ independent $\chi^2_1$ random variables. The coefficients in this linear combination are the eigenvalues of $\Omega_u$, which need to be estimated.

As shown in the Appendix, the Fréchet derivative of $\Delta_u$ is a functional $\Delta_u^* = \Delta_u^*(X, Y, \Phi)$, where $\Phi$ is the joint distribution of $(X, Y)$. This functional can be estimated as $\hat{\Delta}_u^*(X, Y, \Phi_n)$, where $\Phi_n$ is the empirical distribution based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Thus, $\Omega_u$ is estimated by

$$\hat{\Omega}_u = E_n \left[ \text{vec}(\Sigma^{1/2} \Delta_u^*(X, Y, \Phi_n)) \text{vec}^T(\Sigma^{1/2} \Delta_u^*(X, Y, \Phi_n)) \right] = \frac{1}{n} \sum_{i=1}^n \text{vec}(\Sigma^{1/2} \Delta_u^*(X_i, Y_i, \Phi_n)) \text{vec}^T(\Sigma^{1/2} \Delta_u^*(X_i, Y_i, \Phi_n)).$$

Substituting $\Phi_n$ for $\Phi$ amounts to replacing the expectation operator $E(\cdot)$ in the Fréchet derivative $\Delta_u^*$ by the sample averaging operator $E_n(\cdot)$. It is easy to see that $\hat{\Omega}_u = \Omega_u + O_p(n^{-1/2})$. The reference distribution for $\hat{f}_u$ is therefore approximated by the distribution of $\hat{\Delta}_u^* = \sum_{i=1}^{pd} \hat{\omega}_i K_i$, where $K_1, \ldots, K_{pd}$ are independent $\chi^2_1$ random variables and the eigenvalues $\hat{\omega}_1, \ldots, \hat{\omega}_{pd}$ of $\hat{\Omega}_u$ are treated as fixed constants. Many approaches are available for calculating quantiles of this type of distribution. One approach is to approximate them by simulation. An alternative approach, see for example Bentler & Xie (2000), is to use Sattherthwaite’s chi-squared approximation. Let $\hat{f}_u = m \hat{f}_u / \operatorname{tr}(\hat{\Omega}_u)$, where the degrees of freedom $m$ are given by the nearest integer to $\operatorname{tr}(\hat{\Omega}_u)^2 / \operatorname{tr}(\hat{\Omega}_u^{-2})$. Then $\hat{f}_u$ is approximately distributed as a $\chi^2_m$. An advantage of Sattherthwaite’s approximation is that it does not require the eigenvalues of the matrix $\hat{\Omega}_u$, which may have large dimensions. In the application described in the next section, we have used both simulation and Sattherthwaite’s approximations. They gave similar $p$-values, and allowed us to select successfully the termination index $u^*$.

It should be emphasized that in our context we should ideally pursue an asymptotic analysis where $p \to \infty$ as $n \to \infty$. The above development is an ad hoc solution because it treats $p$ as fixed. While it works well for our application in §4, it is indeed possible that such fixed-$p$ asymptotics may perform poorly when $p \to \infty$ as $n$ increases; see Ledoit & Wolf (2002). The rigorous asymptotic development under such circumstances remains an open problem.
4. Discriminating Regulatory Elements

With increasing availability of sequenced genomes, a large amount of research has been devoted to computational methods for the prediction of regulatory elements, which are stretches of the nuclear DNA whose function is to determine the activation of genes. These methods identify putative locations that can then be tested experimentally for regulatory function. In particular, recently developed algorithms (Elnitski et al., 2003; Kolbe et al., 2004; Taylor et al., 2006) predict human regulatory elements based on patterns in their alignments with other species. These are supervised classification algorithms trained using positive and negative instances. The positive training set comprises alignments of experimentally validated regulatory elements, and the negative one comprises alignments of a random sample of DNA regions believed to be nonfunctional.

For our application, we consider training sets obtained from alignments of five mammals: human, chimpanzee, mouse, rat and dog. These alignments are strings of so-called alignment columns. Each alignment column has five entries, one for each species, containing one of the four nucleotide symbols $A, C, G, T$ or a symbol signifying ‘gap’. Thus, there are a total of $5^5 - 1$ possible alignment columns, the ‘minus 1’ occurring because one cannot have a column composed of five gaps. This number is greatly reduced by pooling alignment columns with a procedure that exploits phylogenetic relations among the species (Taylor et al., 2006). The procedure produces an alphabet of six symbols, each encoding a group of alignment columns. An alignment can then be ‘spelled’ in this reduced alphabet. The data are pre-processed so that the strings in the two training sets all have approximately the same length, and so that the two training sets are balanced in size; each contains 257 strings, for a total of $n = 514$ independent observational units.

We then consider the $p = 6^4 = 1296$ patterns of length 4 in the 6-symbol alphabet, and compute the frequencies of such ‘words’, $X \in \mathbb{R}^p$ say, for each training string. A binary response $Y$ is defined as $Y = 1$ for regulatory elements and 0 for neutral DNA, and the task is to perform dimension reduction for the regression of $Y$ on $X$, based on the 514 available observations. Since the response is binary we restrict attention to a one-dimensional central subspace spanned by $\beta \in \mathbb{R}^p$, and select a seed corresponding to Fisher’s linear discriminant:

$$\hat{\nu} = \frac{1}{\|\bar{x}_1 - \bar{x}_0\|} (\bar{x}_1 - \bar{x}_0).$$

This seeding works most effectively under the linearity condition mentioned in §2.1. In our application, $p = 1296$ and $d = 1$, so we regard the linearity condition as quite reasonable.

To select the index $u^*$ we compute the $p$-values of $\tilde{f}_u$ using Satterthwaite’s chi-squared approximation, as shown in Table 1. The table shows that the first two dislocations are significant but the following two are not, with substantially higher $p$-values. Thus we select $u^*$

<table>
<thead>
<tr>
<th>$u$</th>
<th>1 to 2</th>
<th>2 to 3</th>
<th>3 to 4</th>
<th>4 to 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_u$</td>
<td>52.19</td>
<td>14.77</td>
<td>6.97</td>
<td>3.97</td>
</tr>
<tr>
<td>$m$</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>$p$-values</td>
<td>$2.734 \times 10^{-11}$</td>
<td>0.022</td>
<td>0.22</td>
<td>0.55</td>
</tr>
</tbody>
</table>

$\tilde{f}_u$, Satterthwaite’s approximation; $m$, degrees of freedom.
\[ \hat{u}^* = 3 \] as termination index. As a check on Sattherthwaite's approximation, we also run simulations to determine \( p \)-values based on the distribution of \( D_u \). Using 1000 simulated realizations of the corresponding random variable, we estimate the \( p \)-values to be 0·00, 0·037, 0·180 and 0·51. These agree well with the Sattherthwaite \( p \)-values.

Figure 1 contains smoothed histograms of the sufficient predictor \( W = \hat{B}^T_x \) for each value of \( Y \) in iterations \( u = 1, 2, 3 \) and 10. They clearly show how the separation between the two classes increases up to iteration 3 and stabilizes afterwards, providing support for the choice of \( \hat{u}^* \) based on \( f_u \).

5. SIMULATION RESULTS

In this section, we use simulations to demonstrate the effectiveness of \( \text{span}(\hat{B}_{u^*}) \) as an estimator of \( S_{Y|X} = \text{span}(\beta) \). To complement our application, we focus on continuous responses. We simulate location regressions with additive normal errors, \( d = 1 \) and \( \beta \in \mathbb{R}^p \):

\[ Y|X \sim g(\beta^T X) + \varepsilon \varepsilon, \varepsilon \varepsilon \sim N(0, 1) \] and \( X \sim N(\mu, \Sigma) \). Since translations or rotations of the predictor vector do not affect our methodology, we take \( \mu = 0_p \) and \( \Sigma = \text{diag}(\lambda) \), with eigenvalues \( \lambda = (\lambda_1, \ldots, \lambda_p)^T \) in nonincreasing order. Our simulation scenarios are
identified by $p, g, \beta$ and $\lambda$. For the predictor dimension we take $p(1) = 10$ and $p(2) = 500$. The regression function is either $g_1(\beta^T X) = \beta^T X$ or $g_2(\beta^T X) = \beta^T X + (\beta^T X)^2/10$. We consider three choices for $\beta$, each normalized to length 1: $\beta(1)$, in which all elements are equal to $p^{-1/2}$, $\beta(2)$, in which the first and last 20% of the elements are equal to $(0.4p)^{-1/2}$ and all other elements are 0, and $\beta(3)$, in which the first and last elements are equal to $2^{-1/2}$ and all other elements are 0. We consider two choices for $\lambda$: $\lambda(1) = (2/3)(2, \ldots, 2, 1, \ldots, 1)^T$, which induces $q = 2$ eigenspaces of equal multiplicity, and $\lambda(2) = (2/p + 1)(p, (p - 1), \ldots, 1)^T$, which induces $q = p$ eigenspaces. We combine $\beta(1)$ and $\beta(2)$ with $\lambda(1)$, and $\beta(3)$ with $\lambda(2)$, so that the number of predictor eigenspaces that are nonorthogonal to the central subspace is $k = 2$ in all the simulations. We therefore know that $\text{span}(B_u) = S_{Y|X}$ for $u \geq 2$. Correspondingly, we expect $\hat{u}^*$ to be small, and easily identified if we extend iterations to, say, $u_{\text{max}} = 10$.

The error standard deviation parameter $\tau$ is set to 0.3 for all simulations. As a result of how the $\beta$’s and $\lambda$’s are defined, in all scenarios $\var(\beta^T X) = 1$, so that $\tau$ is 30% of the signal standard deviation in the linear regression cases. In summary, we consider $2 \times 2 \times 3 = 12$ simulation scenarios, each of which is sampled 100 times, with a sample size of $n = 100$. In the six scenarios with $p = 10$ we have $n > p$, while in the six scenarios with $p = 500$ we have $n < p$.

We record the performance of three different estimators: the iterative estimator, $\hat{\beta}_{\text{It}} = \hat{B}_{u^*}$, with seed given by $\hat{v} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})/n$: the ordinary least squares vector for the regression of $Y$ on the first three principal components of $X$, $\hat{\beta}_{\text{PC}(3)}$; and the ordinary least squares vector for the regression of $Y$ on the first ten principal components of $X$, $\hat{\beta}_{\text{PC}(10)}$. In the simulations with $p = 10$ the latter coincides with the full ordinary least squares vector $\beta_\infty = \hat{\Sigma}^{-1}\hat{v}$.

Applying a standard sufficient dimension reduction technique after reduction to a few leading principal components of the predictors is an intuitive and common way of overcoming a shortage of observations. However, in our simulations the central subspace is not aligned with the leading principal component directions and thus we are intentionally testing the performance of $\hat{\beta}_{\text{It}}$ in situations in which the estimators $\hat{\beta}_{\text{PC}(\ell)}$, for $\ell < p$, ought not to work effectively.

For a generic sample and estimator vector $\hat{\beta}$, we measure estimation accuracy by the absolute value of the sample correlation $c(\cdot, \cdot)$ between $\hat{\beta}^T X$ and $\beta^T X$. For each simulation, the average and standard deviation of $|c(\hat{\beta}; \beta)|$ across the 100 samples are computed for each of the three estimators, and for a random vector as a benchmark. Averages are reported in Table 2.

The performance of the iterative estimator is outstanding. In the six scenarios with $p = 10$, the average absolute correlations were around 0.985, with standard deviations in the range 0.002–0.008, while, in the six scenarios with $p = 500$, the correlations were around 0.955 with standard deviations in the 0.007–0.010 range. In terms of comparisons, $\hat{\beta}_{\text{It}}$ did consistently much better than $\hat{\beta}_{\text{PC}(3)}$, and on par with $\hat{\beta}_{\text{PC}(10)} = \hat{\beta}_\infty$ when $p = 10$. When $p = 500$, $\hat{\beta}_{\text{It}}$ outperforms by far both $\hat{\beta}_{\text{PC}(3)}$ and $\hat{\beta}_{\text{PC}(10)}$, which achieve average absolute correlations around 0.25 and 0.45, respectively.

The average correlations in Table 2 also indicate that $\hat{\beta}_{\text{It}}$ does better relative to the least-squares-type estimators when the predictor covariance has $q = 2$ eigenspaces and, as to be expected, when $n < p$. Additional simulations, results not shown, confirmed that the performance of $\hat{\beta}_{\text{It}}$ relative to $\hat{\beta}_{\text{PC}(\ell)}$ and $\hat{\beta}_{\text{PC}(p)} = \hat{\beta}_\infty$ improves as the number of eigenspaces $q$ decreases. Additional simulations also confirmed that the relative performance improves
of $\Sigma$, which is the sample covariance matrix. By finding the optimal projection, the iterative estimator can host a nondegenerate projection of $Y$ on $\Sigma$ when $n > p$. We also have evidence that it may have the edge over standard estimators generally when $n$ is not much larger than $p$. The proposed estimator seems particularly advantageous in at least two situations: (i) when only a few eigenspaces of $\Sigma$ host a nondegenerate projection of $S_{Y|X}$, and (ii) when $\Sigma$ has a small number of eigenspaces. Situation (i) is implicitly assumed in two-stage reduction strategies that bypass the $n < p$ predicament by applying standard sufficient dimension reduction techniques to the first few principal components of $X$; see for example Chiaromonte & Martinelli (2002) and Li & Li (2004). However, in the methodology we propose, the sufficient predictors need not be generated from the leading principal components.

Situation (ii) may characterize the regulatory elements application in §4, since word frequencies vary in a highly structured way because words share sub-words. This covariance is reflected in the eigenstructure of $\Sigma$, which presents two dominant and well-separated eigenvalues ($\hat{\lambda}_1, \hat{\lambda}_2$), followed by three smaller eigenvalues of similar sizes ($\hat{\lambda}_3, \hat{\lambda}_4, \hat{\lambda}_5$), followed by a long tail of yet smaller and slowly decaying eigenvalues. The first forty eigenvalues are shown as solid circles in Fig. 2. This suggests that the underlying $\Sigma$ may be

Table 2. Simulation study. Average absolute correlations between $\hat{\beta}^T X$ and $\beta^T X$

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Iterative</th>
<th>PC(3)</th>
<th>PC(10) = OLS</th>
<th>Random</th>
<th>Iterative</th>
<th>PC(3)</th>
<th>PC(10)</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1, \beta_1, q = 2$</td>
<td>0.994</td>
<td>0.641</td>
<td>0.996</td>
<td>0.315</td>
<td>0.961</td>
<td>0.223</td>
<td>0.420</td>
<td>0.089</td>
</tr>
<tr>
<td>$g_1, \beta_2, q = 2$</td>
<td>0.994</td>
<td>0.669</td>
<td>0.996</td>
<td>0.273</td>
<td>0.956</td>
<td>0.232</td>
<td>0.415</td>
<td>0.091</td>
</tr>
<tr>
<td>$g_1, \beta_3, q = p$</td>
<td>0.984</td>
<td>0.911</td>
<td>0.996</td>
<td>0.329</td>
<td>0.952</td>
<td>0.291</td>
<td>0.510</td>
<td>0.097</td>
</tr>
<tr>
<td>$g_2, \beta_1, q = 2$</td>
<td>0.994</td>
<td>0.671</td>
<td>0.996</td>
<td>0.278</td>
<td>0.963</td>
<td>0.229</td>
<td>0.418</td>
<td>0.102</td>
</tr>
<tr>
<td>$g_2, \beta_2, q = 2$</td>
<td>0.994</td>
<td>0.694</td>
<td>0.996</td>
<td>0.292</td>
<td>0.958</td>
<td>0.227</td>
<td>0.419</td>
<td>0.087</td>
</tr>
<tr>
<td>$g_2, \beta_3, q = p$</td>
<td>0.984</td>
<td>0.910</td>
<td>0.996</td>
<td>0.389</td>
<td>0.954</td>
<td>0.296</td>
<td>0.498</td>
<td>0.094</td>
</tr>
</tbody>
</table>

OLS, ordinary least squares; PC(\ell), OLS regression of $Y$ on the first $\ell$ principal components of $X$, for $\ell = 3, 10$.

Table 3. Simulation study. Average absolute correlations between $\hat{\beta}^T X$ and $\beta^T X$, for $p = 98, n = 100$

<table>
<thead>
<tr>
<th>Iterative</th>
<th>PC(3)</th>
<th>PC(10)</th>
<th>PC(98)</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterative</td>
<td>0.970</td>
<td>0.332</td>
<td>0.564</td>
<td>0.959</td>
</tr>
</tbody>
</table>

OLS, ordinary least squares; PC(\ell), OLS regression of $Y$ on the first $\ell$ principal components of $X$, for $\ell = 3, 10, 98$. as the sample size $n$ decreases, even when $n$ exceeds $p$. For instance, with the linear regression function, all predictors active, $q = 2$ and fixing the predictor dimension just below the sample size, $p = 98 < n = 100$, we obtained the results shown in Table 3. Note that here the iterative estimator outperforms the full least squares vector, albeit by a small amount.

6. DISCUSSION

Our simulation results indicate that the proposed estimator performs well when $n < p$, and on a par with standard estimators requiring inversion of $\Sigma$ when $n > p$. We also have evidence that it may have the edge over standard estimators generally when $n$ is not much larger than $p$. The proposed estimator seems particularly advantageous in at least two situations: (i) when only a few eigenspaces of $\Sigma$ host a nondegenerate projection of $S_{Y|X}$, and (ii) when $\Sigma$ has a small number of eigenspaces. Situation (i) is implicitly assumed in two-stage reduction strategies that bypass the $n < p$ predicament by applying standard sufficient dimension reduction techniques to the first few principal components of $X$; see for example Chiaromonte & Martinelli (2002) and Li & Li (2004). However, in the methodology we propose, the sufficient predictors need not be generated from the leading principal components.

Situation (ii) may characterize the regulatory elements application in §4, since word frequencies vary in a highly structured way because words share sub-words. This covariance is reflected in the eigenstructure of $\Sigma$, which presents two dominant and well-separated eigenvalues ($\hat{\lambda}_1, \hat{\lambda}_2$), followed by three smaller eigenvalues of similar sizes ($\hat{\lambda}_3, \hat{\lambda}_4, \hat{\lambda}_5$), followed by a long tail of yet smaller and slowly decaying eigenvalues. The first forty eigenvalues are shown as solid circles in Fig. 2. This suggests that the underlying $\Sigma$ may be
partitioned into $q = 3$ or $4$ eigenspaces. To gain insight into such a possibility, we simulated normal data of the same dimension, $p = 1296$, and size, $n = 514$, as the regulatory elements data, with a covariance matrix partitioned on $4$ eigenspaces, and distinct eigenvalues given by $\hat{\lambda}_1, \hat{\lambda}_2$, the average of $\hat{\lambda}_3, \hat{\lambda}_4$ and $\hat{\lambda}_5$, and the average of $\hat{\lambda}_j$, $j = 6, \ldots, 1296$. The first $40$ eigenvalues of the sample covariance matrix for these simulated data are shown as triangles in Fig. 2, and present a good agreement with the eigenvalues of $\hat{\Sigma}$. As a consequence, we do not need to assume that the central subspace resides within the span of the dominant variability directions of $X$ for our iterative approach to work effectively. Indeed, the iterative method settles on to a central subspace estimate after only $\hat{u}^* = 3$ iterations; see Table 1.

Even if most of the eigenvalues of $\Sigma$ are distinct, they may be amenable to clustering, and replacing each group of similar eigenvalues with a single representative number may provide a good approximation in many practical instances. Selecting an appropriate $u^*$ in our iterative procedure can therefore be viewed as implicitly ‘smoothing’ the eigenvalues of $\Sigma$, and thereby damping the inaccuracy caused by estimating too many parameters.

In this article we have addressed regressions that are intrinsically low-dimensional. Regardless of the fixed value of $p$, and of its relationship to the sample size $n$, the rationale is that a relatively small number of composite predictors $\beta^T X$ will suffice to capture the regression information, so that $Y \mid X \mid \beta^T X$. This model-free conditional independence requirement distinguishes sufficient dimension reduction from other reductive approaches. The central subspace $S_{Y \mid X}$ serves as the foundation for sufficient dimension reduction theory and methodology, which has proven its value in many large-sample regression settings over the past decade. The definition of $S_{Y \mid X}$ has nothing to do with the sample size $n$, but it suggests that methodology may be influenced more by the relationship between $n$ and $d = \dim(S_{Y \mid X})$ than by the relationship between $n$ and $p$. Our experience indicates that the methodology proposed here will be useful in application, provided $n$ is large relative to the dimension of the $\Sigma$-envelope, which lies between $d$ and $p$. These include very high-dimensional, ‘under-sampled’ regressions which admit a low-dimensional informative core. However, in its current version, our methodology cannot tackle contemporary
‘large-\(p\)-small-\(n\)’ regressions in which information accumulates as \(p\) grows. Indeed, other techniques such as support vector machines (Cortes & Vapnik, 1995) can be used to extract information from the data without reducing the predictors. It may be possible to adapt sufficient dimension reduction to large-\(p\)-small-\(n\) regressions by allowing \(d\) to grow with \(p\).

ACKNOWLEDGEMENT

Research for this article was supported by grants from the U.S. National Science Foundation, and by a grant from the U.S. National Institutes of Health. We would like to thank Svitlana Tyekucheva for help with data preparation, and the referees for helpful suggestions.

APPENDIX

Proofs

Proof of Theorem 1. The assertion of the theorem is equivalent to the following two statements combined: (i) span\((R_u)\) \(\subseteq\) \(M_{Y|X}\) for \(u < k\) and span\((R_v)\) = \(M_{Y|X}\) for \(u \geq k\); (ii) if, for some \(u\), span\((R_{u+1})\) = span\((R_u)\), then span\((R_v)\) = span\((R_u)\) for all \(v > u\). We now prove these two statements separately.

Proof of statement (i). Let \(\lambda_1 > \cdots > \lambda_q > 0\) be the distinct eigenvalues of \(\Sigma\), and let \(P_1, \ldots, P_q\) be the projections on to the corresponding eigenspaces. Write \(M_j = P_j v\). Then, by the coverage condition, \(M_{Y|X} = (M_1, \ldots, M_k)\). Also, recall that \(M_j \neq 0\) only for \(k\) eigenspaces. For simplicity, assume these to be the first \(k\) eigenspaces. Since \(\Sigma = \sum_{j=1}^q \lambda_j P_j\) and \(I_p = \sum_{j=1}^q P_j v = \sum_{j=1}^k M_j\), for any \(u = 1, 2, \ldots\) we have

\[
\Sigma^u = \sum_{j=1}^q \lambda_j^u P_j v = \sum_{j=1}^k \lambda_j^u M_j, \quad v = I_p v = \sum_{j=1}^q P_j v = \sum_{j=1}^k M_j.
\]

The matrix \(R_u\) can therefore be rewritten as

\[
R_u = \begin{pmatrix}
\sum_{j=1}^k M_j, & \sum_{j=1}^k \lambda_j M_j, & \ldots, & \sum_{j=1}^k \lambda_j^{u-1} M_j
\end{pmatrix}.
\]

This can be re-expressed as the matrix product \(R_u = (M_1 \ldots M_k)(V_u \otimes I_d)\), where the \((i, j)\)th element of \(V_u\) is \(\lambda_i^{j-1}\), \(i = 1, \ldots, k\), \(j = 1, \ldots, u\). It follows that span\((R_u)\) \(\subseteq\) span\((M_1 \ldots M_k)\) for any \(u\). If \(u = k\) then det\((V_k)\) is the Vandermonde determinant of \(k\) variables, which is nonzero because \(\lambda_1, \ldots, \lambda_k\) are distinct and nonzero. Hence, when \(u = k\), \(V_k \otimes I_d\) is nonsingular and span\((R_u)\) = span\((M_1 \ldots M_k)\). When \(u > k\), \(V_u\) can be written as \((V_k, U)\) for some \(k \times (u - k)\) matrix \(U\). Hence, \(R_u = (M_1 \ldots M_k)(V_k \otimes I_d, U \otimes I_d)\) and span\((R_u)\) = span\((M_1 \ldots M_k)\), as desired.

Proof of statement (ii). This statement is equivalent to the following: if, for some \(u\), span\((\Sigma^u v)\) \(\subseteq\) span\((R_u)\), then span\((\Sigma^v v)\) \(\subseteq\) span\((R_u)\) for all \(v > u\). This is a generalization of a theorem in Cook & Li (2002), which concerned a single \(p \times 1\) vector instead of the \(p \times d\) matrix \(v\). If span\((\Sigma^u v)\) \(\subseteq\) span\((R_u)\), then there are \(d \times d\) matrices \(C_1, \ldots, C_u\) such that \(\Sigma^u = v C_1 + \cdots + \Sigma^{u-1} v C_u\). Consequently

\[
\Sigma^{u+1} = \Sigma(v C_1 + \cdots + \Sigma^{u-1} v C_u) = \Sigma(v C_1 + C_2 C_u) + \cdots + \Sigma^{u-1} v (C_{u-1} + C_u C_u).
\]

This implies that span\((\Sigma^{u+1} v)\) \(\subseteq\) span\((R_u)\). By induction, span\((\Sigma^v v)\) \(\subseteq\) span\((R_u)\) for all \(v > u\), as desired. \(\square\)
The matrices expressions for the particular seed (8) used in our regulatory elements application. However, because span \( v \) and there are at least \( span \) removed, which leads to the desired equality. □

Proof of Lemma 1. For notational convenience let \( R_u = \text{span}(R_u) \), and recall that \( B_u \) is the projection of \( \Sigma^{-1} v \) on to \( R_u \), and \( R_u \subseteq R_{u+1} \). We therefore have

\[
B_u = P_{\mathbb{R}_u} \Sigma^{-1} v = P_{\mathbb{R}_u} P_{\mathbb{R}_{u+1}} \Sigma^{-1} v = P_{\mathbb{R}_u} B_{u+1}.
\]

Hence, \( P_{\mathbb{R}_u} B_{u+1} = P_{\mathbb{R}_u} B_{u+1} \). The proof will be complete if we can show \( P_{\mathbb{R}_u} B_{u+1} = P_{\mathbb{R}_u} B_{u+1} \). To see this equality, rewrite the right-hand side as

\[
P_{\mathbb{R}_u} B_{u+1} = P_{\mathbb{R}_u} P_{\mathbb{R}_{u+1}} P_{\mathbb{R}_u} B_{u+1} = P_{\mathbb{R}_u} P_{\mathbb{R}_{u+1}} P_{\mathbb{R}_u} B_{u+1}.
\]

However, because \( \text{span}(P_{\mathbb{R}_u} B_{u+1}) \subseteq \mathcal{R}_u \), on the right-hand side the projection operator \( P_{\mathbb{R}_u} \) can be removed, which leads to the desired equality.

Proof of Theorem 3. First, suppose that \( \text{span}(B_u) = \mathcal{S}_{Y|X} \). Then \( \mathcal{S}_{Y|X} \subseteq \text{span}(R_u) \), and consequently \( \mathcal{S}_{Y|X} \subseteq \text{span}(R_v) \) for all \( v \geq u \). This implies that \( \text{span}(B_v) = \mathcal{S}_{Y|X} \), and therefore \( F_v = 0 \), for all \( v \geq u \).

Next, suppose that \( F_v = 0 \) for all \( v \geq u \). By definition \( F_v = 0 \) implies that \( \Delta_v = 0 \), which in turn implies that \( B_{v+1} = P_{B_v} B_{v+1} \). Consequently, \( \text{span}(B_{v+1}) \) is contained in the range of \( P_{B_v} \), which is \( \text{span}(B_v) \). It follows that

\[
\text{span}(B_v) \supseteq \text{span}(B_{v+1}) \supseteq \ldots.
\]

Suppose \( \text{span}(B_u) \neq \mathcal{S}_{Y|X} \). Since \( B_u \) has rank \( d \) by Theorem 2, \( \mathcal{S}_{Y|X} \) cannot be a proper subset of \( \text{span}(B_u) \). Hence, by the above relationship there cannot be any \( R_v, v \geq u \), that contains \( \mathcal{S}_{Y|X} \). However, this contradicts Theorem 1, which implies that \( \text{span}(B_v) = \mathcal{S}_{Y|X} \) for all sufficiently large \( v \).

Proof of Theorem 4. We first give the full expression for \( \Delta_{u}^* \):

\[
\Delta_u = B_u^* - B_u^*,
\]

where

\[
B_u^* = R_u^* (R_u^T \Sigma R_u) (R_u^T v + R_u^T v^*) \\
- R_u (R_u^T \Sigma R_u) (R_u^T v + R_u^T v^* + R_u^T \Sigma R_u + R_u^T \Sigma R_u) (R_u^T \Sigma R_u) R_u^T v,
\]

\[
R_u^* = \{v^*, \Sigma \Sigma^T v + \Sigma \Sigma^T v^*, \ldots, (\Sigma v^*)^T v + \Sigma^T v^* \},
\]

\[
\Sigma^* = X^T X - \Sigma, \quad (\Sigma^* )^T = \sum_{0 \leq i, j \leq -1} \Sigma^T \Sigma^T \Sigma^T, \quad \text{for } \ell = 2, \ldots, v - 1.
\]

The matrices \( v \) and \( v^* \) are specific to the seed used. Towards the end of this proof we will give their expressions for the particular seed (8) used in our regulatory elements application.

We now prove Theorem 4. We assume without loss of generality that \( E(X) = 0 \). Let \( \mathcal{F} \) be a convex set of distributions of \((X,Y)\) that contains all empirical distributions \( \Phi_n \) and the true distribution \( \Phi \). Then \( \Sigma \) and \( \Delta_u \) can be viewed as matrix-valued functionals defined on \( \mathcal{F} \); that is, \( \Sigma: \mathcal{F} \rightarrow \mathbb{R}^{p \times p} \) and \( \Delta_u: \mathcal{F} \rightarrow \mathbb{R}^{p \times d} \). In this notation, we have

\[
\Sigma = \Sigma(\Phi), \quad \hat{\Sigma} = \Sigma(\Phi_n); \quad \Delta_u = \Delta_u(\Phi), \quad \hat{\Delta}_u = \Delta_u(\Phi_n).
\]
Let $\Sigma^*$ and $\Delta_u^*$ be the Fréchet derivatives of $\Sigma(\cdot)$ and $\Delta_u(\cdot)$ at the true distribution $\Phi$. Then the matrices $\hat{\Sigma}$ and $\hat{\Delta}_u$ have the following von-Mises expansions:

$$
\hat{\Sigma} = \Sigma + E_n(\Sigma^*) + O_p(n^{-1}), \quad \hat{\Delta}_u = \Delta_u + E_n(\Delta^*_u) + O_p(n^{-1}).
$$

More details of Fréchet derivatives and von-Mises expansions can be found in Fernholz (1983, Ch. III), Bickel et al. (1993, p. 454) and Serfling (1980, p. 214). Hence $\hat{\Delta}_u^T \hat{\Sigma} \hat{\Delta}_u$ has the expansion

$$
\Delta_u^T \Sigma \Delta_u + E_n(\Delta_u^T \Sigma \Delta_u + \Delta_u^T \Sigma^* \Delta_u + \Delta_u^T \Sigma \Delta_u^*)
+ E_n(\Delta_u^T) E_n(\Sigma^*) E_n(\Delta_u^*) + E_n(\Delta_u^T) E_n(\Sigma \Delta_u^*) + O_p(n^{-1}).
$$

Since $\Delta_u = 0$ under the null hypothesis, the above expansion reduces to

$$
\Delta_u^T \hat{\Sigma} \hat{\Delta}_u = E_n(\Sigma^{1/2} \Delta_u^*)^T E_n(\Sigma^{1/2} \Delta_u^*) + O_p(n^{-1}).
$$

We now derive the Fréchet derivative $\Delta_u^*$. The first equality in (A1) follows from Lemma 1; the third and fifth follow from straightforward Fréchet differentiation. To show the second equality, differentiate $B_u = R_u(R_u^T \Sigma R_u)^{-1} R_u^T v$ to obtain

$$
B_u^* = R_u^*(R_u^T \Sigma R_u)^{-1} R_u^T v + R_u((R_u^T \Sigma R_u)^{-1} \Sigma^* R_u^T) R_u^T v + R_u(R_u^T \Sigma R_u)^{-1} R_u^T v
+ R_u(R_u^T \Sigma R_u)^{-1} R_u^T v^*.
$$

We need to compute $R_u((R_u^T \Sigma R_u)^{-1} \Sigma^* R_u^T)$ for the Fisher linear discriminant, as defined by (8). The population version of (8) is $A = E[XY]/E(Y)$, $B(F) = E[X(1-Y)]/E(1-Y)$. The population version of (8) is

$$
v = (A - B)/\|A - B\|,
$$

where

$$
A = E[XY]/E(Y), \quad B(F) = E[X(1-Y)]/E(1-Y).
$$

Now take the Fréchet derivative:

$$
v^* = \{(\|A - B\|(A^* - B^*) - \|A - B\|^*(A - B))/(\|A - B\|)^2 \}
\quad (A3)
$$

In the above, the Fréchet derivatives $A^*$ and $B^*$ are easily found to be

$$
A^* = [E(Y)]^{-2} [E(Y)(XY - E(XY)) - E(XY)[Y - E(Y)]],
B^* = [E(1 - Y)]^{-2} [(X(1 - Y) - E[X(1 - Y)])E(1 - Y) - E[X(1 - Y)](-Y + E(Y))].
$$
The Fréchet derivative $\|A - B\|^*$ in (A3) is now derived as

$$
\|A - B\|^* = \left[\left( (A - B)^T (A - B) \right)^{1/2} \right]^*
$$

$$
= 1/2 \left( (A - B)^T (A - B) \right)^{-1/2} \left( (A - B)^T (A - B) \right)^*
$$

$$
= \|A - B\|^{-1/2} (A - B)^T (A^* - B^*).
$$

Now substitute $A^*, B^*$ and $\|A - B\|^*$ into (A3) to obtain the expression for $v^*$.

REFERENCES


[Received January 2006. Revised November 2006]