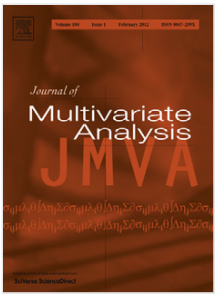
Accepted Manuscript



Predictive power of principal components for single-index model and sufficient dimension reduction

Andreas Artemiou, Bing Li

PII: S0047-259X(13)00076-6
DOI: http://dx.doi.org/10.1016/j.jmva.2013.04.015
Reference: YJMVA 3543

To appear in: *Journal of Multivariate Analysis*

Received date: 3 December 2012

Please cite this article as: A. Artemiou, B. Li, Predictive power of principal components for single-index model and sufficient dimension reduction, *Journal of Multivariate Analysis* (2013), http://dx.doi.org/10.1016/j.jmva.2013.04.015

Predictive power of principal components for single-index model and sufficient dimension reduction

Andreas Artemiou* and Bing Li
Department of Mathematical Sciences, Michigan Technological University and
Department of Statistics, Pennsylvania State University

Abstract

In this paper we demonstrate that a higher-ranking principal component of the predictor tends to have a stronger correlation with the response in single index models and sufficient dimension reduction. This tendency holds even though the orientation of the predictor is not designed in any way to be related to the response. This provides a probabilistic explanation why it is often beneficial to peform regression on principal components — a practice commonly known as principal component regression but whose validity has long been debated. This result is a generalization of earlier results by Li (2007), Artemiou and Li (2009), and Ni (2011), where the same phenomenon was conjectured and rigorously demonstrated for linear regression.

Key words and phrases. Permutation invariance; Principal component analysis; Rotation invariance; Single-index model; Sufficient dimension reduction.

1 Introduction

Principal component analysis has been used for dimension reduction for regression problems ever since its introduction by Pearson (1901) and Hotelling (1933). Let $\boldsymbol{X}$ be a p-dimensional random vector and Y be a random variable. When p is large relative to the sample size n, it is a common practice to regress Y on the first few principal components of $\boldsymbol{X}$ rather than $\boldsymbol{X}$ itself to avoid singularity or

*Corresponding author, aartemio@mtu.edu, 306 Fisher Hall, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, 906-487-2884

ill-conditioned matrix inversion. However, the validity of this tactic, often referred to as the Principal Component Regression (Jolliffe, 1982), has long been debated — questioned by some and defended by others. The gist of the debate is that, since the principal components are extracted solely from the covariance matrix $\mathbf{\Sigma}$ of the predictors $\boldsymbol{X}$, a process in which the response plays no role whatsoever, there seems no reason to think that the first few principal components are any better in predicting the response than the last few principal components. This debate was documented and illuminated in Cook (2007) in the context of Fitted Principal Components. See also Artemiou and Li (2009), Hall and Yang (2010).

Artemiou and Li (2009), inspired by a conjecture by Li (2007), proved the following result: if the response variable Y is not pre-designed to favor any specific orientation of the ellipsoid representing the covariance matrix $\mathbf{\Sigma}$, then, under the linear regression model

$$Y = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X} + \varepsilon, \quad \varepsilon \perp\!\!\!\perp \boldsymbol{X}, \tag{1}$$

with probability greater than a half, the correlation between Y and the ith principal component of $\boldsymbol{X}$ is greater than the correlation between Y and the jth principal component of $\boldsymbol{X}$ for any $i < j$. Here, $\perp\!\!\!\perp$ indicates independence. More specifically, let $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ be the eigenvectors of $\mathbf{\Sigma}$ corresponding to its ith and jth largest eigenvalues, with $i < j$. Then the probability of

$$|\mathrm{corr}(Y, \boldsymbol{v}_i\boldsymbol{X})| > |\mathrm{corr}(Y, \boldsymbol{v}_j\boldsymbol{X})| \tag{2}$$

is always greater than a half, as long as there is no predesigned alignment between $\boldsymbol{\beta}$ and the orientation of the ellipsoid representing the positive definite covariance matrix $\mathbf{\Sigma}$ of $\boldsymbol{X}$.

Using an invariant argument by Arnold and Brockett (1992) and a stronger invariant assumption, Ni (2011) calculated the exact probability for (2) to happen to be

$$(2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\},$$

where λ_i and λ_j are the ith and jth largest eigenvalues of $\mathbf{\Sigma}$. We note that this probability is always greater than or equal to a half and, for $\lambda_i \gg \lambda_j$, it can be arbitrarily close to 1. Ni (2011) also generalized this result in several other directions, but confined his analysis to linear regression.

From a different angle, Hall and Yang (2010) established, also in the linear regression setting, that regressing Y on to the first few principal components of $\boldsymbol{X}$ achieves the minimax bound of a conditional mean squared error between $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$ and $\hat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{X}$. They allow the predictor to be either a vector or a function, but the relation between $\boldsymbol{X}$ and Y is still intrinsically linear.

In this paper, we extend the probabilistic characterization of the inequality (2) to much more general settings than the linear regression model. For example, consider the single index model

$$Y = f(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}) + \varepsilon, \quad \boldsymbol{X} \perp\!\!\!\perp \varepsilon, \tag{3}$$

where f is an unknown, arbitrary function. See, for example, Powell, Stock, and Stoker (1989), Härdle, Hall, and Ichimura (1993), and Ichimura (1993). Another example is the heteroscedastic single index model

$$Y = f(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}) + g(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X})\varepsilon, \quad \boldsymbol{X} \perp\!\!\!\perp \varepsilon. \tag{4}$$

where f and g are arbitrary functions. Under these models, does the ranking of a principal component of $\boldsymbol{X}$ affects its correlation with the response Y? In other words, do the principal components, which are not designed to predict any response variable, have any predictive power for the response in single index models and beyond, regardless of how that response is related to the single index? If the answer is yes, then it makes sense to perform nonlinear regressions on the principal components of the predictors.

One can put this question in the broader context of unsupervised versus supervised dimension reduction. The PCA is a main tool for unsupervised dimension reduction and one could regard models (3) and (4) as special cases of supervised (or sufficient) dimension reduction. Indeed, consider the following conditional independence relations

$$Y \perp\!\!\!\perp E(Y|\boldsymbol{X})|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \tag{5}$$

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}. \tag{6}$$

Here, $A \perp\!\!\!\perp B|C$ reads "the random elements A and B are conditionally independent given a third random element C". The first relation asserts that the conditional mean $E(Y|\boldsymbol{X})$ depends on $\boldsymbol{X}$ only through the index $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$; it includes the mean regression model (3) as a special case. The second relation asserts that the full

conditional distribution depends on $\boldsymbol{X}$ through the index $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$; it includes the heteroscedastic mean regression model as a special case. The above two relations are, respectively, special cases of sufficient dimension reduction for conditional mean and that for conditional distribution. In sufficient dimension reduction, $\boldsymbol{\beta}$ can be an arbitrary matrix, and the objective is to recover the subspaces spanned by the columns of $\boldsymbol{\beta}$ without the knowledge of the functional forms of the conditional mean or conditional distribution. The subspace spanned by the columns of $\boldsymbol{\beta}$ in (5) is called the central mean subspace; that spanned by the columns of $\boldsymbol{\beta}$ in (6) is called the central subspace. For further information about sufficient dimension reduction, see, for example, Li (1991), Cook (1996, 1998), Cook and Li (2002), Xia, Tong, Li, and Zhu (2002), and Li and Wang (2007).

Intuitively, we can regard supervised dimension reduction as reducing the dimension of $\boldsymbol{X}$ while preserving its relation with a response Y, and principal component analysis as reducing the dimension of $\boldsymbol{X}$ so as to keep those directions that contain most of the variation of $\boldsymbol{X}$. Thus, in this broader context our question becomes: do the variables extracted from the original predictor using unsupervised dimension reduction have the tendency — even if a weak tendency — to be aligned with the variables extracted using supervised dimension reduction? If this relation can be established, then it makes sense to perform unsupervised dimension reduction before supervised dimension reduction as a preprocessing or prescreening step. This would be of practical significance because such prescreenings are commonly used in practice and often work well. See, for example, Chiaromonte and Martinelli (2002), and Li, Kim, and Altman (2010). Our results show that, at least in the situations where the dimension of $\boldsymbol{X}$ can be reduced to 1, the above assertion is true.

The rest of the paper is organized as follows. In Section 2 we give a brief outline of the previous results for linear regression, and point out that the conditions used in Ni (2011) are somewhat stronger than those used in Artemiou and Li (2009): the former involves permutation invariance and the latter involves rotation invariance. In sections and 3 and 4 we generalize the stronger result of Ni (2011) to sufficient dimension reduction for conditional mean and conditional distribution under rotation invariance. In section 5 we make the corresponding generalizations of the weaker result of Artemiou and Li (2009) under permutation invariance. These are followed by a short discussion in section 6.

2 Overview of previous results

Li (2007) argued that, if Y is correlated with $\boldsymbol{X}$ at all, it should be correlated with its first principal component, unless nature's choice of $\boldsymbol{\Sigma}$ has a favored orientation. Artemiou and Li (2009) formulated the notion of "no favorable orientation" rigorously as the following assumption.

Assumption 1 *The covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$ is a random matrix that can be written as $\lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^\intercal + \ldots + \lambda_p \boldsymbol{v}_p \boldsymbol{v}_p^\intercal$ where*

1. *$(\lambda_1, \ldots, \lambda_p)$ are positive, exchangeable random variables;*
2. *$(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ are exchangeable random vectors;*
3. *$(\lambda_1, \ldots, \lambda_p) \perp\!\!\!\perp (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$.*

Intuitively, what is required by this assumption is that the relative positions of eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ can be freely permuted without changing the distribution of $\boldsymbol{\Sigma}$. Under this assumption Artemiou and Li (2009) showed the following result. Suppose linear regression model (1) holds, and

1. $\boldsymbol{\Sigma}$ satisfies Assumption 1;
2. $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ and $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$, $\varepsilon \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) < \infty$;
4. $P \equiv \lambda$, where λ is the Lebesgue measure ($\equiv$ denotes mutual absolute continuity).

Then

$$P\left(\mathrm{corr}^2(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \geq \mathrm{corr}^2(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) > 1/2. \tag{7}$$

$\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the ith and jth eigenvectors of $\boldsymbol{\Sigma}$, and $i > j$.

Under somewhat stronger assumptions, Ni (2011) calculated the exact probability for the inequality in (7) as

$$P\left(\mathrm{corr}^2(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}) \geq \mathrm{corr}^2(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta})\right) = (2/\pi) \arctan\left(\lambda_i/\lambda_j\right)^{\frac{1}{2}}. \tag{8}$$

Ni's result require either $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$ to be invariant under rotation transformations, which will be spelled out precisely in the next section. It is the rotation invariance that makes it possible to determine the probability of the inequality in (8) completely in terms of the relative magnitudes of the eigenvalues.

These results indicate that, as long as the alignment between $\boldsymbol{\beta}$ and the axes of $\boldsymbol{\Sigma}$ is arbitrary at the outset, a linear regression relation has the natural tendency of making the correlation between Y and the first principal component of $\boldsymbol{X}$ larger than that between Y and the second principal component of $\boldsymbol{X}$. In the case that $\boldsymbol{\Sigma}$ is fixed the expectation on the left hand side of the above equality is removed.

3 Predictive power of PCA for the conditional mean model

In this section we extend Ni's result to the setting the conditional mean dimension reduction (5). We first lay out two key assumptions under *either* of which this result holds. The first assumption is the same as that made in Ni (2011, Section 3); the second assumption was used but not explicitly spelled out in Ni (2011). Let $\mathbb{U}^{p\times p}$ be the class of all $p \times p$ orthogonal matrices.

Assumption 2 *The distribution of the random vector $\boldsymbol{\beta}$ is spherically symmetric; that is, for any $\boldsymbol{A} \in \mathbb{U}^{p\times p}$, $\boldsymbol{\beta} \overset{\mathcal{D}}{=} \boldsymbol{A}\boldsymbol{\beta}$.*

A necessary and sufficient condition for this to happen is that the density of $\boldsymbol{\beta}$ depends only on $\|\boldsymbol{\beta}\|$.

Assumption 3 *The random matrix $\boldsymbol{\Sigma}$ is symmetric and invariant under orthogonal transformation; that is, for any $\boldsymbol{A} \in \mathbb{U}^{p\times p}$ we have $\boldsymbol{\Sigma} \overset{\mathcal{D}}{=} \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\top}$. Moreover, all the eigenvalues of $\boldsymbol{\Sigma}$ are distinct and positive.*

Although this assumption was not explicitly mentioned in Ni (2011), the assumption used in the proof of Theorem 1 in that paper seems to be closer to this assumption than that used in Artemiou and Li (2009), as stated in Assumption 1. In fact, Assumption 1 is not sufficient to guarantee the spherical distribution used in the proof of Theorem 1 in Ni (2011).

Note that if $\boldsymbol{\Sigma}$ is invariant under orthogonal transformations, then the eigenvalues of $\boldsymbol{\Sigma}$ are nonrandom constants. Because our results only rely on arbitrary orientations, the relative lengths of the axes of $\boldsymbol{\Sigma}$ are irrelevant and need not be restricted as they were in the previous papers. In the following, when we say $\boldsymbol{\Sigma}$ has

spectrum decomposition $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\mathsf{T}$ we mean $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal elements are eigenvalues of $\boldsymbol{\Sigma}$, with $\lambda_1 \geq \cdots \geq \lambda_p$; $\boldsymbol{V}$ is the orthogonal matrix whose columns are eigenvectors corresponding go $\lambda_1, \ldots, \lambda_p$. The next lemma shows that $\boldsymbol{\Sigma}$ is invariant under orthogonal transformations then the distribution of $\boldsymbol{V}$ is unchanged by multiplying from the left by any orthogonal matrix.

Lemma 1 *If $\boldsymbol{\Sigma}$ is a random matrix satisfying Assumption 3 with spectrum decomposition $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\mathsf{T}$, then $\boldsymbol{A}\boldsymbol{V} \overset{\mathcal{D}}{=} \boldsymbol{V}$ for any orthogonal matrix $\boldsymbol{A}$.*

Proof. Let $\mathbb{M} = \{\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\mathsf{T} : \boldsymbol{U} \in \mathbb{U}^{p\times p}\}$. Since the diagonal elements of $\boldsymbol{\Lambda}$ are distinct, a matrix $\boldsymbol{\Gamma} \in \mathbb{M}$ uniquely determines the matrix $\boldsymbol{U}$ in $\boldsymbol{\Gamma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\mathsf{T}$. Write this function as $\boldsymbol{U}(\boldsymbol{\Gamma})$. Because $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T} \overset{\mathcal{D}}{=} \boldsymbol{\Sigma}$ we have

$$\boldsymbol{U}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T}) \overset{\mathcal{D}}{=} \boldsymbol{U}(\boldsymbol{\Sigma}).$$

However, because $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\mathsf{T}$, we have

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T} = \boldsymbol{A}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\mathsf{T}\boldsymbol{A}^\mathsf{T}$$

which implies $\boldsymbol{U}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T}) = \boldsymbol{A}\boldsymbol{V}$. Hence

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\mathsf{T}) \overset{\mathcal{D}}{=} \boldsymbol{U}(\boldsymbol{\Sigma}) = \boldsymbol{V},$$

as desired. □

In the following, let $\mathcal{R}^p$ denote the class of Borel sets in $\mathbb{R}^p$. The next two lemmas show that $\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}$ is spherically symmetric if either of Assumption 2 or Assumption 3 is satisfied.

Lemma 2 *If $\boldsymbol{\beta}$ is spherically symmetric and $\boldsymbol{\Sigma}$ is any nonrandom symmetric matrix with spectrum decomposition $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\mathsf{T}$, then $\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}$ is spherically symmetric.*

Proof. Let $\boldsymbol{A} \in \mathbb{U}^{p\times p}$. Because $\boldsymbol{V} \in \mathbb{U}^{p\times p}$, $\boldsymbol{A}\boldsymbol{V}^\mathsf{T} \in \mathbb{U}^{p\times p}$. Hence

$$\boldsymbol{A}(\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}) = (\boldsymbol{A}\boldsymbol{V}^\mathsf{T})\boldsymbol{\beta} \overset{\mathcal{D}}{=} \boldsymbol{\beta} \overset{\mathcal{D}}{=} \boldsymbol{V}^\mathsf{T}\boldsymbol{\beta},$$

which means $\boldsymbol{V}^\mathsf{T}\boldsymbol{\beta}$ is spherically symmetric. □

Lemma 3 *If $\boldsymbol{\Sigma}$ satisfies Assumption 3 with spectrum decomposition $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal$ and $\boldsymbol{\beta}$ is any nonrandom, nonzero vector, then $\boldsymbol{V}^\intercal\boldsymbol{\beta}$ is spherically symmetric.*

Proof. Let $\mathbb{O}_{\boldsymbol{\beta}}$ be the orbit of $\boldsymbol{\beta}$ under transformations in $\mathbb{U}^{p\times p}$; that is,

$$\mathbb{O}_{\boldsymbol{\beta}} = \{\boldsymbol{A}\boldsymbol{\beta} : \boldsymbol{A} \in \mathbb{U}^{p\times p}\} = \{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}\| = \|\boldsymbol{\beta}\|\}.$$

Then, for any $\boldsymbol{b} \in \mathbb{O}_{\boldsymbol{\beta}}$ and $\boldsymbol{A} \in \mathbb{U}^{p\times p}$,

$$\boldsymbol{V}^\intercal\boldsymbol{b} \overset{\mathcal{D}}{=} (\boldsymbol{A}\boldsymbol{V})^\intercal\boldsymbol{b} = \boldsymbol{V}^\intercal(\boldsymbol{A}^\intercal\boldsymbol{b}).$$

Since $\boldsymbol{A}^\intercal$ can be any member of $\mathbb{U}^{p\times p}$, the above implies that the distribution of $\boldsymbol{V}^\intercal\boldsymbol{b}$ is the same for all $\boldsymbol{b} \in \mathbb{O}_{\boldsymbol{\beta}}$. Let $\boldsymbol{\eta}$ be a random vector uniformly distributed on $\mathbb{O}_{\boldsymbol{\beta}}$ and $\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{\Sigma}$. Let $B \in \mathcal{R}^p$. Because $\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{V}$ we have, for any $\boldsymbol{b} \in \mathbb{O}_{\boldsymbol{\beta}}$,

$$P(\boldsymbol{V}^\intercal\boldsymbol{b} \in B) = E\{E[I_B(\boldsymbol{V}^\intercal\boldsymbol{\eta})|\boldsymbol{\eta} = \boldsymbol{b}]\}.$$

In other words $E\{E[I_B(\boldsymbol{V}^\intercal\boldsymbol{\eta})|\boldsymbol{\eta}]\}$ is constant on the support of $\boldsymbol{\eta}$. Since this is true for all $B \in \mathcal{R}^p$, we have $\boldsymbol{V}^\intercal\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{\eta}$. Hence $P(\mathbf{V}^\intercal\boldsymbol{\beta} \in B) = P(\mathbf{V}^\intercal\boldsymbol{\eta} \in B)$ for all $B \in \mathcal{R}^p$, or equivalently,

$$\boldsymbol{V}^\intercal\boldsymbol{\beta} \overset{\mathcal{D}}{=} \boldsymbol{V}^\intercal\boldsymbol{\eta}. \tag{9}$$

We now show that $\boldsymbol{V}^\intercal\boldsymbol{\eta}$ is spherically symmetric. Because $\boldsymbol{\eta}$ is spherically symmetric and $\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{V}$, $\boldsymbol{\eta}|\boldsymbol{V}$ is also spherically symmetric. By Lemma 2, $\boldsymbol{V}^\intercal\boldsymbol{\eta}|\boldsymbol{V}$ is spherically symmetric. So if we can show $\boldsymbol{V}^\intercal\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{V}$, then $\boldsymbol{V}^\intercal\boldsymbol{\eta}$ is spherically symmetric unconditionally. Let $B \in \mathcal{R}^p$. Then

$$P(\boldsymbol{V}^\intercal\boldsymbol{\eta} \in B|\boldsymbol{V}) = P(\boldsymbol{\eta} \in B|\boldsymbol{V}) = P(\boldsymbol{\eta} \in B),$$

where the first equality holds because $\boldsymbol{\eta}|\boldsymbol{V}$ is spherically symmetric and the second holds because $\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{V}$. Because the above equality is true for all $B \in \mathcal{R}^p$, we have $\boldsymbol{V}^\intercal\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{V}$, as desired. □

In fact, in order for $\boldsymbol{V}^\intercal\boldsymbol{\beta}$ to be spherically symmetric, we can allow both the matrix $\boldsymbol{\Sigma}$ in Lemma 2 and the vector $\boldsymbol{\beta}$ in Lemma 3 to be random as long as we keep $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ independent, as shown by the next corollary.

Corollary 1 *If either of the following conditions are satisfied:*

1. *$\boldsymbol{\beta}$ satisfies Assumption 2 and $\boldsymbol{\Sigma}$ is a random matrix such that $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{\Sigma}$,*

2. *$\boldsymbol{\Sigma}$ satisfies Assumption 3 and $\boldsymbol{\beta}$ is a random vector such that $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{\Sigma}$,*

then $\boldsymbol{V}^\top \boldsymbol{\beta}$ is spherically symmetric.

PROOF. Suppose condition 1 is satisfied. Let $\boldsymbol{\Sigma}_0$ be a fixed matrix in $\mathbb{M}$ with spectrum decomposition $\boldsymbol{V}_0 \boldsymbol{\Lambda} \boldsymbol{V}_0^\top$. Then, by Lemma 2, $\boldsymbol{V}_0^\top \boldsymbol{\beta}$ is spherically symmetric. That is, for any $\boldsymbol{A} \in \mathbb{U}^{p \times p}$ and $B \in \mathcal{R}^p$,

$$P(\boldsymbol{A}\boldsymbol{V}_0^\top \boldsymbol{\beta} \in B) = P(\boldsymbol{V}_0^\top \boldsymbol{\beta} \in B).$$

However, because $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{V}$, we have

$$\begin{aligned} P(\boldsymbol{A}\boldsymbol{V}_0^\top \boldsymbol{\beta} \in B) &= P(\boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{\beta} \in B | \boldsymbol{V} = \boldsymbol{V}_0), \\ P(\boldsymbol{V}_0^\top \boldsymbol{\beta} \in B) &= P(\boldsymbol{V}^\top \boldsymbol{\beta} \in B | \boldsymbol{V} = \boldsymbol{V}_0). \end{aligned}$$

It follows that

$$P(\boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{\beta} \in B | \boldsymbol{V}) = P(\boldsymbol{V}^\top \boldsymbol{\beta} \in B | \boldsymbol{V}).$$

Now take (unconditional) expectation on both sides to obtain

$$P(\boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{\beta} \in B) = P(\boldsymbol{V}^\top \boldsymbol{\beta} \in B),$$

which means $\boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{\beta} \overset{\mathcal{D}}{=} \boldsymbol{V}^\top \boldsymbol{\beta}$. The proof of the assertion under condition 2 is similar. □

Note that Corollary 1 accommodates both Lemma 2 and Lemma 3 because a constant vector or a constant matrix is independent of any random element.

We now prove the main result of this section. Consider the following conditional mean dimension reduction model

$$E(Y | \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(Y | \boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{10}$$

This is the same model as (5) except that here we treat $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ as random and that $\boldsymbol{\beta}$ is assumed to be a vector, rather than a matrix.

Theorem 1 *Suppose (10) holds with* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$ *and* $\mathrm{var}(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) < \infty$ *almost surely, and*

1. $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$;

2. $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is a linear function of* $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$;

3. *either Assumption 2 or Assumption 3 are satisfied;*

4. $\mathrm{cov}(Y, \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \neq 0$ *almost surely.*

Then, for $i < j$,

$$P(\mathrm{corr}^2(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(Y, \boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})) = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}].$$

Condition 2 in the theorem is commonly assumed for sufficient dimension reduction. It implies

$$E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \boldsymbol{P}_{\boldsymbol{\Sigma}}^{\mathsf{T}}(\boldsymbol{\beta})\boldsymbol{X}, \tag{11}$$

where $\boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ is the projection matrix $\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}$ relative to the $\boldsymbol{\Sigma}$-inner product. See, for example, Cook (1998). Before proving the theorem we recall another well known fact: if $\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$ are random vectors, then

$$\mathrm{cov}[E(\boldsymbol{U}_1|\boldsymbol{U}_3), \boldsymbol{U}_2] = \mathrm{cov}[\boldsymbol{U}_1, E(\boldsymbol{U}_2|\boldsymbol{U}_3)]. \tag{12}$$

That is, conditional expectation is a self-adjoint operator.

Proof of Theorem 1. By (12),

$$\begin{aligned}\mathrm{cov}(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \mathrm{cov}(Y, E(\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &= \mathrm{cov}(E(Y|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}), \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}).\end{aligned}$$

Applying (10) and then (12), we rewrite the right hand side as

$$\mathrm{cov}(E(Y|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}), \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathrm{cov}(Y, \boldsymbol{v}_i^{\mathsf{T}}E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})|\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

Apply (11) to further rewrite the right hand side as

$$\begin{aligned}\mathrm{cov}(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{P}_{\boldsymbol{\Sigma}}^{\mathsf{T}}(\boldsymbol{\beta})\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{P}_{\boldsymbol{\Sigma}}^{\mathsf{T}}(\boldsymbol{\beta})\mathrm{cov}(\boldsymbol{X}, Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &= \lambda_i\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}\mathrm{cov}(\boldsymbol{X}, Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}).\end{aligned} \tag{13}$$

In the meantime, because $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$, we have $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{\Sigma}$. Hence

$$\mathrm{var}(\boldsymbol{v}_i\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathrm{var}(\boldsymbol{v}_i\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{v}_i = \lambda_i. \tag{14}$$

Combine (13) and (14) to obtain

$$\text{corr}^2(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \frac{\lambda_i^2(\boldsymbol{v}_i^\intercal\boldsymbol{\beta})^2(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-2}[\boldsymbol{\beta}^\intercal\text{cov}(\boldsymbol{X}, Y|\boldsymbol{\beta}, \boldsymbol{\Sigma})]^2}{\text{var}(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma})\lambda_i}.$$

It follows that

$$\frac{\text{corr}^2(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\text{corr}^2(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})} = \frac{\lambda_i(\boldsymbol{v}_i^\intercal\boldsymbol{\beta})^2}{\lambda_j(\boldsymbol{v}_j^\intercal\boldsymbol{\beta})^2}.$$

Hence

$$\begin{aligned} P\left(\text{corr}^2(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \text{corr}^2(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) \\ = P((\boldsymbol{v}_i^\intercal\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\intercal\boldsymbol{\beta})^2 > \lambda_j/\lambda_i) \\ = P(-(\lambda_i/\lambda_j)^{\frac{1}{2}} < \boldsymbol{v}_j^\intercal\boldsymbol{\beta}/\boldsymbol{v}_i^\intercal\boldsymbol{\beta} < (\lambda_i/\lambda_j)^{\frac{1}{2}}). \end{aligned}$$

Because $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$, we have $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{\Sigma}$. This combined with Assumption 2 is the condition 1 of Corollary 1; it combined with Assumption 3 is condition 2 of Corollary 1. In either case

$$\boldsymbol{V}^\intercal\boldsymbol{\beta} = (\boldsymbol{v}_1^\intercal\boldsymbol{\beta}, \ldots, \boldsymbol{v}_p^\intercal\boldsymbol{\beta})^\intercal$$

is spherically symmetric. Hence, by Arnold and Brocket (1992), $\boldsymbol{v}_j^\intercal\boldsymbol{\beta}/\boldsymbol{v}_i^\intercal\boldsymbol{\beta}$ has a Cauchy distribution, which entails

$$P(-(\lambda_i/\lambda_j)^{\frac{1}{2}} < \boldsymbol{v}_j^\intercal\boldsymbol{\beta}/\boldsymbol{v}_i^\intercal\boldsymbol{\beta} < (\lambda_i/\lambda_j)^{\frac{1}{2}}) = (2/\pi)\arctan[(\lambda_j/\lambda_i)^{\frac{1}{2}}],$$

as desired. □

The above result can be extended to vector-valued responses as follows. Let $\boldsymbol{Y}$ be a q-dimensional random vector and assume

$$E(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(\boldsymbol{Y}|\boldsymbol{\beta}^\intercal\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{15}$$

Then, for any $\boldsymbol{\alpha} \in \mathbb{R}^q$, we have

$$E(\boldsymbol{\alpha}^\intercal\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(\boldsymbol{\alpha}^\intercal\boldsymbol{Y}|\boldsymbol{\beta}^\intercal\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

We can now apply Theorem 1 to $(\boldsymbol{\alpha}^\intercal\boldsymbol{Y}, \boldsymbol{X})$ to obtain the following generalization.

Corollary 2 *Suppose relation (15) holds with* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$ *and*

$$E(\boldsymbol{Y}^\mathsf{T}\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) < \infty$$

almost surely. Then, under conditions 1, 2, and 3 in Theorem 1 we have, for $i < j$ *and any* $\boldsymbol{\alpha} \in \mathbb{R}^p$ *such that*

$$\mathrm{cov}(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{Y}, \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \neq 0 \quad \textit{almost surely},$$

we have

$$P\left(\mathrm{corr}^2(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{Y}, \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{Y}, \boldsymbol{v}_j^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}].$$

4 Predictive power of PCA for conditional distribution

We now turn to the general sufficient dimension reduction problem

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \tag{16}$$

where $\boldsymbol{Y}$ and $\boldsymbol{X}$ are both random vectors as defined in the last section. This is the same model as (6) except that here $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are treated as random.

Theorem 2 *Suppose relation (16) holds with* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$*. Then, under the conditions 1, 2, and 3 in Theorem 1, for* $i < j$ *and any* $f : \mathbb{R}^p \to \mathbb{R}$ *satisfying*

$$\mathrm{var}(f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\Sigma}) < \infty, \quad \mathrm{cov}(f(\boldsymbol{Y}), \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \neq 0 \ \textit{almost surely},$$

we have

$$P\left(\mathrm{corr}^2(f(\boldsymbol{Y}), \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(f(\boldsymbol{Y}), \boldsymbol{v}_j^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}].$$

Proof. Since the proof is similar to that of Theorem 1, here we only highlight the difference. Because $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, we have

$$E[f(\boldsymbol{Y})|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}] = E[f(\boldsymbol{Y})|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}].$$

Using this fact and derivations parallel to the proof of Theorem 1, we find

$$\mathrm{corr}^2(f(\boldsymbol{Y}), \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \frac{\lambda_i^2(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^2(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-2}[\boldsymbol{\beta}^\mathsf{T}\mathrm{cov}(\boldsymbol{X}, f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\Sigma})]^2}{\mathrm{var}(f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\Sigma})\lambda_i}.$$

The rest of the proof follows similarly. □

5 Weaker form of the inequalities

As we have noticed in Section 2, the inequality obtained by Ni (2011) is stronger than that by Artemiou and Li (2009), and the conditions used for the two results are also different. In the previous two sections we have generalized Ni's result to sufficient dimension reduction for conditional mean and conditional distribution. In this section we demonstrate the weaker form of the inequality to sufficient dimension reduction under weaker conditions than Assumptions 3. These weaker assumptions are essentially the same as Assumption 1 but we are able to remove some unnecessary ingredient from that assumption. Let $\mathbb{V}^{p\times p}$ be the class of all $p \times p$ permutation matrices. Recall that a permutation matrix is any matrix that can be obtained by permuting the rows (or columns) of an identity matrix, that $\mathbb{V}^{p\times p}$ is a finite group, and that $\mathbb{V}^{p\times p} \subseteq \mathbb{U}^{p\times p}$.

Assumption 4 *The random matrix $\boldsymbol{\Sigma}$ is symmetric and positive definite with spectral decomposition $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\mathbf{V}$ has exchangeable columns in the sense that $\boldsymbol{V}\boldsymbol{C} \overset{\mathcal{D}}{=} \boldsymbol{V}$ for all $\boldsymbol{C} \in \mathbb{V}^{p\times p}$ and $\boldsymbol{\Lambda}$ is nonrandom. Moreover, the eigenvalues of $\boldsymbol{\Sigma}$ are all distinct.*

Compared with Assumption 1 used in Artemiou and Li (2009), here we no longer assume the diagonal elements of $\boldsymbol{\Lambda}$ to be exchangeable random variables. The intuition for this relaxation is that it is the relative position of eigenvalues and eigenvectors that matters; in other words, as long as eigenvectors can be permuted freely, we do not need the extra freedom to permute the eigenvalues. The next proposition shows that Assumption 4 is indeed weaker than Assumption 3.

Proposition 1 *Assumption 3 implies Assumption 4.*

Proof. We need to show that, for any $\boldsymbol{C} \in \mathbb{V}^{p\times p}$, $\boldsymbol{V} \overset{\mathcal{D}}{=} \boldsymbol{V}\boldsymbol{C}$. We will show a stronger result:

$$\boldsymbol{V}\boldsymbol{A} \overset{\mathcal{D}}{=} \boldsymbol{V} \quad \text{for any } \boldsymbol{A} \in \mathbb{U}^{p\times p}. \tag{17}$$

Let $\boldsymbol{W}$ be a random matrix supported on $\mathbb{U}^{p\times p}$ such that $\boldsymbol{W} \perp\!\!\!\perp \boldsymbol{V}$ and $\boldsymbol{A}\boldsymbol{W} \overset{\mathcal{D}}{=} \boldsymbol{W}$ for all $\boldsymbol{A} \in \mathbb{U}^{p\times p}$. Let B be a measurable set in $\mathbb{U}^{p\times p}$. Because $\boldsymbol{W} \perp\!\!\!\perp \boldsymbol{V}$,

$$P(\boldsymbol{V}^\top\boldsymbol{w} \in B) = P(\boldsymbol{V}^\top\boldsymbol{W} \in B|\boldsymbol{W} = \boldsymbol{w})$$

for any $\boldsymbol{w} \in \mathbb{U}^{p\times p}$. Because $\boldsymbol{w}^\top \in \mathbb{U}^{p\times p}$, by Lemma 1 $\boldsymbol{w}^\top \boldsymbol{V} \overset{\mathcal{D}}{=} \boldsymbol{V}$, which implies that the distribution of $\boldsymbol{V}^\top \boldsymbol{w}$ is the same for all $\boldsymbol{w} \in \mathbb{U}^{p\times p}$. Hence the right hand side above does not depend on $\boldsymbol{w}$, which implies $\boldsymbol{V}^\top \boldsymbol{W} \perp\!\!\!\perp \boldsymbol{W}$. Moreover, because $\boldsymbol{W} \perp\!\!\!\perp \boldsymbol{V}$ and $\boldsymbol{A}\boldsymbol{W} \overset{\mathcal{D}}{=} \boldsymbol{W}$,

$$P(\boldsymbol{A}^\top \boldsymbol{W} \in B|\boldsymbol{V}) = P(\boldsymbol{W} \in B|\boldsymbol{V}).$$

Taking $\boldsymbol{A}$ to be $\boldsymbol{V}$ we see that

$$P(\boldsymbol{V}^\top \boldsymbol{W} \in B|\boldsymbol{V}) = P(\boldsymbol{W} \in B|\boldsymbol{V}) = P(\boldsymbol{W} \in B).$$

Since the right hand side is nonrandom, we have $\boldsymbol{V}^\top \boldsymbol{W} \perp\!\!\!\perp \boldsymbol{V}$.

Because $\boldsymbol{W} \perp\!\!\!\perp \boldsymbol{V}$ and $\boldsymbol{V}^\top \boldsymbol{W} \perp\!\!\!\perp \boldsymbol{W}$,

$$P(\boldsymbol{V}^\top \in B) = P(\boldsymbol{V}^\top \boldsymbol{W} \in B|\boldsymbol{W} = \boldsymbol{I}_p) = P(\boldsymbol{V}^\top \boldsymbol{W} \in B).$$

This means $\boldsymbol{V}^\top \overset{\mathcal{D}}{=} \boldsymbol{V}^\top \boldsymbol{W}$. Because $\boldsymbol{V}^\top \boldsymbol{W} \perp\!\!\!\perp \boldsymbol{V}$, $\boldsymbol{W} \perp\!\!\!\perp \boldsymbol{V}$, and $\boldsymbol{A}\boldsymbol{W} \overset{\mathcal{D}}{=} \boldsymbol{W}$ for any $\boldsymbol{A} \in \mathbb{U}^{p\times p}$, we have

$$\begin{aligned} P(\boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{W} \in B) = P(\boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{W} \in B|\boldsymbol{V}) &= P(\boldsymbol{W} \in B|\boldsymbol{V}) \\ &= P(\boldsymbol{V}^\top \boldsymbol{W} \in B|\boldsymbol{V}) = P(\boldsymbol{V}^\top \boldsymbol{W} \in B). \end{aligned}$$

This means $\boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{W} \overset{\mathcal{D}}{=} \boldsymbol{V}^\top \boldsymbol{W}$. Hence

$$\boldsymbol{A}\boldsymbol{V}^\top \overset{\mathcal{D}}{=} \boldsymbol{A}\boldsymbol{V}^\top \boldsymbol{W} \overset{\mathcal{D}}{=} \boldsymbol{V}^\top \boldsymbol{W} \overset{\mathcal{D}}{=} \boldsymbol{V}^\top.$$

Because this holds for every $\boldsymbol{A} \in \mathbb{U}^{p\times p}$, we have $\boldsymbol{V}^\top \boldsymbol{A}^\top \overset{\mathcal{D}}{=} \boldsymbol{V}^\top$ for every $\boldsymbol{A} \in \mathbb{U}^{p\times p}$, which is equivalent to statement (17). □

The next theorem extends the result of Artemiou and Li (2009) to model (15), where the conditional mean is the primary interest of dimension reduction. The assertion of the next theorem is more specific than Theorem 3.1 in Artemiou and Li (2009) in that it not only asserts the probability of the desired inequality is greater than a half but also gives the explicit expression of the amount greater than a half. The result is weaker than Theorem 1 in that the probability depends on the distribution of $\boldsymbol{V}^\top \boldsymbol{\beta}$, which cannot be completely determined under permutation invariance.

Theorem 3 *Suppose relation (15) holds with* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$ *and*

$$E(\boldsymbol{Y}^\top \boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) < \infty$$

almost surely. Moreover, suppose:

1. $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$*;*
2. $E(\boldsymbol{X}|\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is a linear function of* $\boldsymbol{\beta}^\top \boldsymbol{X}$*;*
3. $\boldsymbol{\Sigma}$ *satisfies Assumption 4.*

Then, for $i < j$*, and any* $\boldsymbol{\alpha} \in \mathbb{R}^q$*, such that* $\mathrm{cov}(\boldsymbol{\alpha}^\top \boldsymbol{Y}, \boldsymbol{\beta}^\top \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \neq 0$ *almost surely, we have*

$$\begin{aligned} &P\left(\mathrm{corr}^2(\boldsymbol{\alpha}^\top \boldsymbol{Y}, \boldsymbol{v}_i^\top \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(\boldsymbol{\alpha}^\top \boldsymbol{Y}, \boldsymbol{v}_j^\top \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) \\ &\qquad = 1/2 + (1/2)P(\lambda_j/\lambda_i < (\boldsymbol{v}_i^\top \boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top \boldsymbol{\beta})^2 < \lambda_i/\lambda_j). \end{aligned}$$

Proof. Using the same argument in the proof of Theorem 1 we can show that

$$\frac{\mathrm{corr}^2(\boldsymbol{\alpha}^\top \boldsymbol{Y}, \boldsymbol{v}_i^\top \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\mathrm{corr}^2(\boldsymbol{\alpha}^\top \boldsymbol{Y}, \boldsymbol{v}_j^\top \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})} = \frac{\lambda_i (\boldsymbol{v}_i^\top \boldsymbol{\beta})^2}{\lambda_j (\boldsymbol{v}_j^\top \boldsymbol{\beta})^2}.$$

Hence

$$P\left(\mathrm{corr}^2(\boldsymbol{\alpha}^\top \boldsymbol{Y}, \boldsymbol{v}_i^\top \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(\boldsymbol{\alpha}^\top \boldsymbol{Y}, \boldsymbol{v}_j^\top \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) = P((\boldsymbol{v}_i^\top \boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top \boldsymbol{\beta})^2 > \lambda_j/\lambda_i).$$

Because $\boldsymbol{V} \perp\!\!\!\perp \boldsymbol{\beta}$, we have, for any $\boldsymbol{b} \in \mathbb{R}^p$ and $B \in \mathcal{R}^2$,

$$\begin{aligned} P((\boldsymbol{v}_i^\top \boldsymbol{\beta}, \boldsymbol{v}_j^\top \boldsymbol{\beta}) \in B|\boldsymbol{\beta} = \boldsymbol{b}) &= P((\boldsymbol{v}_i^\top \boldsymbol{b}, \boldsymbol{v}_j^\top \boldsymbol{b}) \in B) \\ &= P((\boldsymbol{v}_j^\top \boldsymbol{b}, \boldsymbol{v}_i^\top \boldsymbol{b}) \in B) \\ &= P((\boldsymbol{v}_j^\top \boldsymbol{\beta}, \boldsymbol{v}_i^\top \boldsymbol{\beta}) \in B|\boldsymbol{\beta} = \boldsymbol{b}). \end{aligned}$$

This means $\boldsymbol{v}_i^\top \boldsymbol{\beta}$ and $\boldsymbol{v}_j^\top \boldsymbol{\beta}$ are exchangeable conditioning on $\boldsymbol{\beta}$. That is

$$P((\boldsymbol{v}_i^\top \boldsymbol{\beta}, \boldsymbol{v}_j^\top \boldsymbol{\beta}) \in B|\boldsymbol{\beta}) = P((\boldsymbol{v}_j^\top \boldsymbol{\beta}, \boldsymbol{v}_i^\top \boldsymbol{\beta}) \in B|\boldsymbol{\beta}).$$

Consequently,

$$\begin{aligned} P((\boldsymbol{v}_i^\top \boldsymbol{\beta}, \boldsymbol{v}_j^\top \boldsymbol{\beta}) \in B) &= E[P((\boldsymbol{v}_i^\top \boldsymbol{\beta}, \boldsymbol{v}_j^\top \boldsymbol{\beta}) \in B|\boldsymbol{\beta})] \\ &= E[P((\boldsymbol{v}_j^\top \boldsymbol{\beta}, \boldsymbol{v}_i^\top \boldsymbol{\beta}) \in B|\boldsymbol{\beta})] \\ &= P((\boldsymbol{v}_j^\top \boldsymbol{\beta}, \boldsymbol{v}_i^\top \boldsymbol{\beta}) \in B). \end{aligned}$$

This shows that $(\boldsymbol{v}_i^\top\boldsymbol{\beta}, \boldsymbol{v}_j^\top\boldsymbol{\beta})$ is exchangeable unconditionally. Hence

$$\begin{aligned}
&P((\boldsymbol{v}_i^\top\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top\boldsymbol{\beta})^2 > \lambda_j/\lambda_i) \\
=&P((\boldsymbol{v}_j^\top\boldsymbol{\beta})^2/(\boldsymbol{v}_i^\top\boldsymbol{\beta})^2 > \lambda_j/\lambda_i) \\
=&P((\boldsymbol{v}_i^\top\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top\boldsymbol{\beta})^2 < \lambda_i/\lambda_j) \\
=&1 - P((\boldsymbol{v}_i^\top\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top\boldsymbol{\beta})^2 \geq \lambda_i/\lambda_j) \\
=&1 - P((\boldsymbol{v}_i^\top\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top\boldsymbol{\beta})^2 > \lambda_i/\lambda_j) + P(\lambda_j/\lambda_i < (\boldsymbol{v}_i^\top\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top\boldsymbol{\beta})^2 < \lambda_i/\lambda_j)
\end{aligned}$$

which implies the desired relation. □

We now generalize this result to the sufficient dimension reduction model (16). The proof is similar to that of Theorem 3 and is omitted.

Theorem 4 *Suppose relation (16) holds with* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma})$*. Then, under the conditions 1, 2, and 3 in Theorem 3 we have, for* $i < j$*, and any* $f \in \mathbb{R}^q \to \mathbb{R}$*, such that*

$$\mathrm{var}(f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\Sigma}) < \infty, \quad \mathrm{cov}(f(\boldsymbol{Y}), \boldsymbol{\beta}^\top\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \neq 0 \ \textit{almost surely},$$

we have

$$\begin{aligned}
&P\left(\mathrm{corr}^2(f(\boldsymbol{Y}), \boldsymbol{v}_i^\top\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(f(\boldsymbol{Y}), \boldsymbol{v}_j^\top\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) \\
&\qquad = 1/2 + (1/2)P(\lambda_j/\lambda_i < (\boldsymbol{v}_i^\top\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\top\boldsymbol{\beta})^2 < \lambda_i/\lambda_j).
\end{aligned}$$

Interestingly, unlike Theorem 1, Corollary 2, and Theorem 2, where Assumption 3 can be replaced by Assumption 2, there is no weaker version of the inequality with Assumption 4 replaced by the assumption that $\boldsymbol{\beta}$ is an exchangeable random vector. This is because the exchangeability of $\boldsymbol{\beta}$ does not guarantee the exchangeability of $\boldsymbol{V}^\top\boldsymbol{\beta}$ for any orthogonal matrix $\boldsymbol{V}$.

6 Discussion

The results of this paper, as well as the earlier papers by Li (2007), Artemiou and Li (2009), Hall and Yang (2010), and Ni (2011) reveal a natural tendency that had not been rigorously characterized previously — that is, the strongest traits or features in high-dimensional data tend to have some predictive power for a given response variable, even if that variable has no pre-designed linkage

with these traits. The easiest way to understand this phenomenon is to imagine a random vector with a single, dominant principal component, and with all the other principal components negligible. In this case, if a response has any relation with the random vector at all, then it has to be related to the dominant principal component, because all the other principal components are essentially 0 and cannot be correlated with the response. This tendency justifies some popular but heuristic statistical practices that seem to have worked well for no obvious reasons, such as principal component regression and performing unsupervised dimension reduction before supervised dimension reduction.

The understanding of this phenomenon is particularly relevant given the heightened interaction and synthesis between supervised (or sufficient) and unsupervised dimension reduction in the recent literature. See, for example, Wu (2008), Fukumizu, Jordan, and Bach (2009), Yeh, Huang, and Lee (2009), Li, Artemiou, and Li (2011), and Lee, Li, and Chiaromonte (2012). Our results provide fresh insights into the interrelation between these two fields.

It is possible that even more general version of this inequality exists. For example, though we have only considered the case where $\boldsymbol{\beta}$ is a vector, it is reasonable to speculate that similar inequalities might hold when $\boldsymbol{\beta}$ is a matrix. Also, we believe that the present results can be extended to the cases where $\boldsymbol{X}$ is a random function, where sufficient dimension reduction takes the form

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} | \langle \boldsymbol{X}, g \rangle.$$

See Ferré and Yao (2003) and Hsing and Ren (2009). This corresponds to the nonlinear version of the setting considered by Hall and Yang (2010).

Finally, we would like to point out that the relation between the unsupervised PCA and supervised dimension reduction is probabilistic and is not particularly strong when λ_i and λ_j are close. As noted, the probability has a lower bound of 1/2 which shows that there is a risk of getting the wrong results if we use PCA alone. Indeed, as noted by Joliffe (1982), Hadi and Ling (1998), there are where the last few principal components are more related to the response, than the first few ones. Nevertheless, when there are a large number of predictors, it is justifiable to use PCA as a pre-screening device, followed by more powerful and computationally intensive methods such as sufficient dimension reduction.

References

Arnold, B. C. and Brockett, P. L. (1992). On distributions whose component ratios are Cauchy. *American Statistician*, **46**, 25–26

Artemiou, A. and Li, B. (2009). On principal components and regression: A statistical explanation of a natural phenomenon, *Statistica Sinica*, **19**, 1557–1565

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.*, **176**, 123–144.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**,983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 1–40.

Cook, R. D. and Li, B. (2002). Dimension Reduction for the conditional mean. *The Annals of Statistics*, **30**, 455–474.

Cook, R. D. and Li, B. (2004). Determining the dimension in Iterative Hessian Transformation. *The Annals of Statistics*, **32**, 2501–2531.

Ferré, L. and Yao, A. F. (2003). Functional slice inverse regression analysis. *Statistics*, **37**, 475–488.

Fukumizu, K., Bach, F. R. and Jordan M. I. (2009). Kernel dimension reduction in regression. *Annals of Statistics*, **4** 1871–1905.

Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principal components in regression. *The American Statistician*, **52**, 15–19.

Hall, P. and Yang, Y.-J. (2010). Ordering and selecting components in multivariate or functional data linear prediction. *Journal of the Royal Statistical Society, Series B*, **72**, 93 – 110.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, **1993**, 157–178.

Hotelling, H. (1933). Analysis of a complex statistical variable into its principal components *Journal of Educational Psychology*, **24**, 417–441.

Hsing, T. and Ren, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *The Annals of Statistics*, **37** 726–755.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **1993**, 71–120.

Joliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **31**, 300–303.

Lee, K.-Y., Li, B., and Chiaromonte, F. (2012). A general theory of nonlinear sufficient dimension reduction: formulation and estimation. Submitted to *The Annals of Statistics*.

Li, B. (2007). Comment: Fisher Lecture: Dimension Reduction in Regression. *StatisticalScience*, **22**, 32–35.

Li, B., Kim, M., and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics*, **38**, 1094–1121.

Li, B., Artemiou, A. and Li, L. (2011). Principal Support Vector Machine for linear and nonlinear sufficient dimension reduction *The Annals of Statistics*, **39**, 3182–3210

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of American Statistical Association*, **102**, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–342.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406–3412.

Ni, L. (2011). Principal Regression Revisited. *Statistica Sinica*, **21**, 741–747.

Pearson, K (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine (6)*, **2**, 559–572.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403–1430.

Wu, H. M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, **17**, 590–610.

Yeh, Y.-R., Huang, S.-Y., and Lee, Y.-Y. (2009). Nonlinear Dimension Reduction with Kernel Sliced Inverse Regression. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1590–1603.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. *J. Roy. Statist. Soc. Ser. B* **64**, 363–410.