Statistica Sinica **19** (2009), 1557-1565

ON PRINCIPAL COMPONENTS AND REGRESSION: A STATISTICAL EXPLANATION OF A NATURAL PHENOMENON

Andreas Artemiou and Bing Li

Pennsylvania State University

Abstract: In this note we give a probabilistic explanation of a phenomenon that is frequently observed but whose reason is not well understood. That is, in a regression setting, the response (Y) is often highly correlated with the leading principal components of the predictor ($\boldsymbol{X}$) even though there seems no logical reason for this connection. This phenomenon has long been noticed and discussed in the literature, and has received renewed interest recently because of the need for regressing Y on $\boldsymbol{X}$ of very high dimension, often with comparatively few sampling units, in which case it seems natural to regress on the first few principal components of $\boldsymbol{X}$. This work stems from a discussion of a recent paper by Cook (2007) which, along with other developments, described a historical debate surrounding, and current interest in, this phenomenon.

Key words and phrases: Dimension reduction, orientationally uniform distribution, principal components, random covariance matrices, regression, stochastic ordering.

1. Introduction

Cook (2007) described an intriguing historical debate surrounding the relation between the regression of a scalar response Y on a random vector $\boldsymbol{X}$ and the principal components of $\boldsymbol{X}$. The debate arose from the practice of regressing Y on the first few principal components of $\boldsymbol{X}$, as suggested and advocated in Kendall (1957, p.75), Hocking (1976), Mosteller and Tukey (1977, p.307), Scott (1992). Some authors, however, question the logic behind this practice, on the basis that in the computation of principal components of $\boldsymbol{X}$, the response Y is never used in any direct or indirect way. For example, Cox (1968, p.272) writes:

> A difficulty seems to be that there is no logical reason why the dependent variable should not be closely tied to the least important principal component.

This view is shared by authors such as Hotelling (1957) and Hawkins and Fatti (1984). See Cook (2007) for a more detailed description.

This question has received renewed interest recently due to the need for handling regression problems with very high-dimensional predictors but relatively few observation units, as one encounters when analyzing microarray data, so that the sample covariance matrix of $\boldsymbol{X}$ is singular and the usual regression techniques cannot be directly applied. Under these circumstances, regressing Y on the first few principal components is a practical solution and often gives reasonable results. See Alter, Brown and Botstein (2000), Chiaromonte and Martinelli (2002), Bura and Pfeiffer (2003), and Li and Li (2004). See also Cook, Li and Chiaromonte (2007) for a different approach to this problem.

In his comments on Cook (2007), Li (2007) made a conjecture in an attempt to explain probabilistically why the response should be related to the leading principal components of the predictors, which was stated roughly as follows:

> If nature arbitrarily selects a covariance matrix $\boldsymbol{\Sigma}$ for $\boldsymbol{X}$ and coefficients $\boldsymbol{\beta}$ for the regression of Y on $\boldsymbol{X}$, then the principal components of $\boldsymbol{X}$ of higher ranks tend to have stronger correlations with Y than do those of lower ranks.

Li (2007) argued intuitively that if $\boldsymbol{X}$ is concentrated on a single direction, then the only way for Y to be correlated with $\boldsymbol{X}$ at all is to be correlated with its first principal component. Likewise if $\boldsymbol{X}$ has an elongated distribution the $\boldsymbol{X}$ components in the longer axes should on average bear stronger correlations with Y. Now if $\boldsymbol{\Sigma}$ is selected arbitrarily then $\boldsymbol{X}$ would have a large probability of having an elongated distribution, and would therefore effect the similar probabilistic ordering of correlations, even if the relation between Y and $\boldsymbol{X}$ is independent of the shape of the distribution of $\boldsymbol{X}$. He supported this conjecture with several simulation studies, that affirmed it.

In this paper we give a precise formulation and a rigorous proof of the conjecture. This provides at least a partial justification for regressing Y on the principal components of $\boldsymbol{X}$, and gives fresh insights into a debate of historical interest and of importance in contemporary data analysis.

In Section 2, we will demonstrate that the conjectured ordering of correlations does occur naturally in practice by analyzing 33 data sets chosen arbitrarily from a database. The conjecture is then formulated and proved in Section 3.

2. The Phenomenon as Seen Through 33 Data Sets

In this section we analyze a collection of data sets in *Arc* software database, which can be found at `http://www.stat.umn.edu/arc/software.html`. From this collection we select 33 suitable data sets. The excluded data sets either have

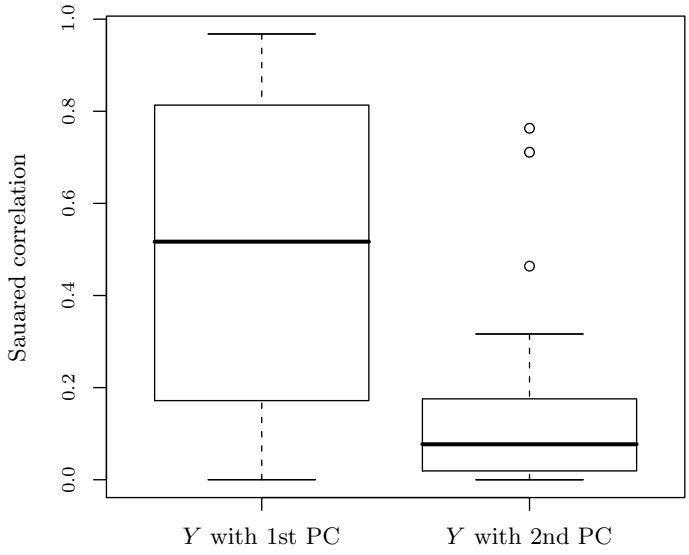

Figure 1. Boxplots of the squared correlations between Y and the first principal component of $\mathbf{X}$ (left) and those between Y and the second principal component of $\mathbf{X}$ (right) for 33 data sets.

only one predictor, or have categorical predictors or responses, or are artificially constructed.

The selected data sets contain from 2 to 12 predictors. For each of them we calculated all principal components of the predictor and their squared correlation with the response. Among these 33, in 24 cases the first principal components have the highest correlation with the responses, in 6 cases the second principal components have the highest squared correlation, in 2 cases the third principal components have the highest squared correlation, and in 1 case the fifth principal component has the highest square correlation. In Figure 1 we present the boxplots of the squared correlation coefficients between the responses and the first principal components of the predictors (left) and between the responses and the second principal components of the predictors (right) for the 33 data sets. This does indicate the tendency for the response to have higher squared correlation with the first principal component of the predictor.

3. Formulation and Proof of the Conjecture

Recall that p random elements, say $W_1, \ldots, W_p$, are exchangeable if, for any permutation $(i_1, \ldots, i_p)$ of $(1, \ldots, p)$, we have $(W_{i_1}, \ldots, W_{i_p}) \stackrel{\mathcal{D}}{=} (W_1, \ldots, W_p)$ where $\stackrel{\mathcal{D}}{=}$ indicates two random elements having the same distribution. To give the conjecture a rigorous formulation we need a precise definition of a random

covariance matrix that has equal probability of any orientation.

Definition 3.1. We say that a $p \times p$ positive semidefinite random matrix $\boldsymbol{\Sigma}$ has an orientationally uniform distribution if $\boldsymbol{\Sigma} = \sigma_1^2 \boldsymbol{v}_1 \boldsymbol{v}_1^T + \cdots + \sigma_p^2 \boldsymbol{v}_p \boldsymbol{v}_p^T$, where each $(\sigma_i^2, \boldsymbol{v}_i)$ is a pair of random elements in which σ_i^2 is a positive random variable and $\boldsymbol{v}_i$ is a p-dimensional random vector, such that $(\sigma_1^2, \ldots, \sigma_p^2)$ are exchangeable with distribution dominated by Lebesgue measure, $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ are exchangeable and orthonormal, $(\sigma_1^2, \ldots, \sigma_p^2)$ and $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ are independent.

Thus if $\boldsymbol{\Sigma}$ is orientationally uniform, the random ellipsoid $\{\boldsymbol{x} : \boldsymbol{x}^T \boldsymbol{\Sigma} \boldsymbol{x} \le c\}$ has equal probability to have any orientation. Or, from a different perspective, suppose that $\boldsymbol{X}$ is a p-dimensional random vector satisfying $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{0}$ and $\operatorname{Var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$, then any random variable among $\boldsymbol{v}_1^T \boldsymbol{X}, \ldots, \boldsymbol{v}_p^T \boldsymbol{X}$ is equally likely to be the 1st, 2nd, ..., or pth principal component of $\boldsymbol{X}$.

In the following, the symbol $\perp\!\!\!\perp$ indicates independence between random elements. We adopt the convention that if U is a random variable, then any constant α that satisfies $P(U < \alpha) \le 1/2 \le P(U \le \alpha)$ is a median of U.

Lemma 3.1. *Suppose $\boldsymbol{\beta}$ and $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ are p-dimensional random vectors such that* 1. $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{v}_1, \boldsymbol{v}_2)$; 2. $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set G;* 3. $\boldsymbol{v}_1$ *and* $\boldsymbol{v}_2$ *are linearly independent and exchangeable. Then $(\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2$ has a unique median of* 1.

Proof. First, we show that 1 is a median of $(\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2$; that is,

$$P\left(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_2)^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2} < 1\right) \le \frac{1}{2} \le P\left(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_2)^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2} \le 1\right). \tag{3.1}$$

Because $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ are exchangeable and $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{v}_1, \boldsymbol{v}_2)$, the random variables $(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2$ and $(\boldsymbol{\beta}^T \boldsymbol{v}_2)^2$ are exchangeable. Hence $P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \le 1) = P((\boldsymbol{\beta}^T \boldsymbol{v}_1)^2/\ (\boldsymbol{\beta}^T \boldsymbol{v}_2)^2 \le 1) = 1 - P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 < 1)$. It follows that $P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 < 1) \le 1 - P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 < 1)$ and $P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \le 1) \ge 1 - P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \le 1)$, which implies (3.1).

Now we show that 1 is the only number that satisfies (3.1). In other words, for any $0 < c_1 < 1$ and $c_2 > 1$ we have $P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \le c_1) < 1/2$ and $P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 < c_2) > 1/2$. We only show the first inequality; the second can be shown similarly. Since $P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \le c_1) = E[P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \le c_1 | \boldsymbol{v}_1, \boldsymbol{v}_2)]$, it suffices to show that for any nonrandom, linearly independent $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, we have $P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \le c_1 | (\boldsymbol{v}_1, \boldsymbol{v}_2) = (\boldsymbol{a}, \boldsymbol{b})) < 1/2$. However, because $(\boldsymbol{v}_1, \boldsymbol{v}_2) \perp\!\!\!\perp \boldsymbol{\beta}$, the above inequality is equivalent to

$$P\left(\frac{(\boldsymbol{\beta}^T \boldsymbol{b})^2}{(\boldsymbol{\beta}^T \boldsymbol{a})^2} \le c_1\right) < \frac{1}{2}. \tag{3.2}$$

Let $c_3 \in (c_1, 1)$. Since $(\boldsymbol{a}, \boldsymbol{b})$ has full column rank, the system of equations

$$\begin{cases} \boldsymbol{\beta}^T \boldsymbol{b} = \sqrt{c}_3 \\ \boldsymbol{\beta}^T \boldsymbol{a} = 1 \end{cases}$$

has a solution, say $\boldsymbol{\beta}_0$. Note that $(\boldsymbol{\beta}_0^T \boldsymbol{b})^2/(\boldsymbol{\beta}_0^T \boldsymbol{a})^2 = c_3 \in (c_1, 1)$. Because $\boldsymbol{\beta} \mapsto (\boldsymbol{\beta}^T \boldsymbol{b})^2/(\boldsymbol{\beta}^T \boldsymbol{a})^2$ is continuous there is a neighborhood of $\boldsymbol{\beta}_0$, say G, such that $\boldsymbol{\beta} \in G \Rightarrow (\boldsymbol{\beta}^T \boldsymbol{b})^2/(\boldsymbol{\beta}^T \boldsymbol{a})^2 \in (c_1, 1)$. By assumption, $P(\boldsymbol{\beta} \in G) > 0$. Therefore $P((\boldsymbol{\beta}^T \boldsymbol{b})^2/(\boldsymbol{\beta}^T \boldsymbol{a})^2 \in (c_1, 1)) > 0$ which, combined with (3.1), implies (3.2).

We are now ready to establish the main result of the paper.

Theorem 3.1. *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *orientationally uniform random matrix,*
2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\operatorname{Var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$,
3. $Y = \boldsymbol{\beta}^T \boldsymbol{X} + \delta$, *where* $\boldsymbol{\beta}$ *is a* p*-dimensional random vector and* δ *is a random variable such that* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$, $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, $E(\delta) = 0$ *and* $\operatorname{Var}(\delta) < \infty$.
4. $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set* $G \in \mathbb{R}^p$.

Let $w_1, \ldots, w_p$ *be the* 1*st, ..., pth principal components of* $\boldsymbol{X}$, *and let* $\rho_i = \rho_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \operatorname{Corr}^2(Y, w_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})$. *Then, whenever* $i < j$, $P(\rho_i \geq \rho_j) > 1/2$.

Proof. Let τ^2 denote $\operatorname{Var}(\delta)$. Let $(\sigma^2_{(1)}, \boldsymbol{v}_{(1)}), \ldots, (\sigma^2_{(p)}, \boldsymbol{v}_{(p)})$ be the reordered $(\sigma_1^2, \boldsymbol{v}_1), \ldots, (\sigma_p^2, \boldsymbol{v}_p)$ such that $\sigma^2_{(1)} \geq \cdots \geq \sigma^2_{(p)}$. We derive an explicit expression for ρ_i. Note that

$$\begin{aligned} \operatorname{Cov}(Y, \boldsymbol{v}_{(i)}^T \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \operatorname{Cov}(\boldsymbol{\beta}^T \boldsymbol{X} + \delta, \boldsymbol{v}_{(i)}^T \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &= \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{v}_{(i)} + \operatorname{Cov}(\delta, \boldsymbol{v}_{(i)}^T \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}). \end{aligned} \tag{3.3}$$

Because $\delta \perp\!\!\!\perp (\boldsymbol{\Sigma}, \boldsymbol{X}, \boldsymbol{\beta})$, we have $\delta \perp\!\!\!\perp (\boldsymbol{v}_{(i)}^T \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$. This implies $\delta \perp\!\!\!\perp \boldsymbol{v}_{(i)}^T \boldsymbol{X}|(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, and hence that the second term in (3.3) is zero. Because $(\sigma^2_{(i)}, \boldsymbol{v}_{(i)})$ is an eigen pair of $\boldsymbol{\Sigma}$, we have $\boldsymbol{\Sigma} \boldsymbol{v}_{(i)} = \sigma^2_{(i)} \boldsymbol{v}_{(i)}$. Hence

$$\operatorname{Cov}^2(Y, \boldsymbol{v}_{(i)}^T \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sigma^4_{(i)} (\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2. \tag{3.4}$$

In the meantime, $\operatorname{Var}(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \operatorname{Var}(\boldsymbol{\beta}^T \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) + 2\operatorname{Cov}(\boldsymbol{\beta}^T \boldsymbol{X}, \delta|\boldsymbol{\beta}, \boldsymbol{\Sigma}) + \operatorname{Var}(\delta|\boldsymbol{\beta}, \boldsymbol{\Sigma})$. Because $\delta \perp\!\!\!\perp (\boldsymbol{\beta}, \boldsymbol{\Sigma})$, the last term on the right is simply τ^2. Because $\delta \perp\!\!\!\perp (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{X})$, we have $\delta \perp\!\!\!\perp \boldsymbol{\beta}^T \boldsymbol{X}|(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. So the second term on the right is 0. Hence

$$\operatorname{Var}(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2. \tag{3.5}$$

Moreover,

$$\operatorname{Var}(\boldsymbol{v}_{(i)}^T \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \boldsymbol{v}_{(i)}^T \boldsymbol{\Sigma} \boldsymbol{v}_{(i)} = \sigma^2_{(i)}. \tag{3.6}$$

Now combine (3.4), (3.5), and (3.6) to obtain

$$\rho_i = \overset{2}{\mathrm{Corr}}(Y, \boldsymbol{v}_{(i)}^T \boldsymbol{X}) = \frac{\sigma_{(i)}^2 (\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2}. \tag{3.7}$$

Cook (2007), in his rejoinder to Li (2007), gave a special case of (3.7).

Let $i < j$. Then, using (3.7), we deduce

$$P(\rho_i \geq \rho_j) = P\bigg(\frac{\sigma_{(i)}^2 (\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2} \geq \frac{\sigma_{(j)}^2 (\boldsymbol{\beta}^T \boldsymbol{v}_{(j)})^2}{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2} \bigg) = P\bigg(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_{(j)})^2} \geq \frac{\sigma_{(j)}^2}{\sigma_{(i)}^2} \bigg).$$

The right hand side can be written as

$$\sum_{k \neq l} P\bigg(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_{(j)})^2} \geq \frac{\sigma_{(j)}^2}{\sigma_{(i)}^2} \bigg| \sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2 \bigg) P\Big(\sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2 \Big).$$

Because $\sigma_1^2, \ldots, \sigma_p^2$ are exchangeable, $(\sigma_{(i)}^2,\ \sigma_{(j)}^2)$ has equal probability to be $(\sigma_k^2, \sigma_\ell^2)$ for any $k \neq \ell$, and that probability is $\binom{p}{2}^{-1}$. Hence the above reduces to

$$\begin{aligned} &\binom{p}{2}^{-1} \sum_{k \neq l} P\bigg(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_{(j)})^2} \geq \frac{\sigma_{(j)}^2}{\sigma_{(i)}^2} \bigg| \sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2 \bigg) \\ &\quad = \binom{p}{2}^{-1} \sum_{k \neq l} P\bigg(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_\ell)^2} \geq \frac{\sigma_\ell^2}{\sigma_k^2} \bigg| \sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2 \bigg), \end{aligned} \tag{3.8}$$

where the equality follows from the fact that, conditioning on the event $(\sigma_{(i)}^2, \sigma_{(j)}^2)$ $= (\sigma_k^2, \sigma_\ell^2)$, one has $(\boldsymbol{v}_{(i)}^2, \boldsymbol{v}_{(j)}^2) = (\boldsymbol{v}_k^2, \boldsymbol{v}_\ell^2)$.

Reexpress each term in the summation in (3.8) as

$$E\bigg[P\bigg(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_\ell)^2} \geq \frac{\sigma_\ell^2}{\sigma_k^2} \bigg| \sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2, \sigma_k^2, \sigma_\ell^2 \bigg) \bigg| \sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2 \bigg]. \tag{3.9}$$

From Definition 3.1 we have

$$\begin{aligned} (\boldsymbol{v}_k, \boldsymbol{v}_\ell) \perp\!\!\!\perp (\sigma_1^2, \ldots, \sigma_p^2) &\Rightarrow (\boldsymbol{v}_k, \boldsymbol{v}_\ell) \perp\!\!\!\perp (\sigma_1^2, \ldots, \sigma_p^2; \sigma_{(1)}^2, \ldots, \sigma_{(p)}^2) \\ &\Rightarrow (\boldsymbol{v}_k, \boldsymbol{v}_\ell) \perp\!\!\!\perp (\sigma_k^2, \sigma_\ell^2, \sigma_{(i)}^2, \sigma_{(j)}^2) \\ &\Rightarrow (\boldsymbol{v}_k, \boldsymbol{v}_\ell) \perp\!\!\!\perp (\sigma_{(i)}^2, \sigma_{(j)}^2) | (\sigma_k^2, \sigma_\ell^2). \end{aligned}$$

Thus the event $\{\sigma_k^2 = \sigma_{(i)}^2, \sigma_\ell^2 = \sigma_{(j)}^2\}$ can be removed from the conditional probability inside the conditional expectation (3.9), which then reduces to

$$E\bigg[P\bigg(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_\ell)^2} \geq \frac{\sigma_\ell^2}{\sigma_k^2} \bigg| \sigma_k^2, \sigma_\ell^2 \bigg) \bigg| \sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2 \bigg]. \tag{3.10}$$

Because $(\boldsymbol{\beta}, \boldsymbol{v}_k, \boldsymbol{v}_\ell) \perp\!\!\!\perp (\sigma_k^2, \sigma_\ell^2)$, for each fixed $0 < s < t$,

$$P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{s}{t}\middle|\sigma_k^2 = t, \sigma_\ell^2 = s\right) = P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{s}{t}\right) > \frac{1}{2},$$

where the inequality follows from Lemma 3.1. By Definition 3.1, the event $\{\sigma_k^2 = \sigma_\ell^2\}$ has probability 0. It follows that

$$P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{\sigma_\ell^2}{\sigma_k^2}\middle|\sigma_k^2, \sigma_\ell^2\right) > \frac{1}{2}$$

almost surely on the event $\{\sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2\}$. Therefore (3.10), and hence (3.8), are strictly greater than $1/2$.

The inequality $P(\rho_i \geq \rho_j) > 1/2$ in Theorem 3.1 is equivalent to

$$P(\rho_i \geq \rho_j) > P(\rho_i < \rho_j). \tag{3.11}$$

In other words ρ_i has a larger probability to be larger than ρ_j than to be smaller than ρ_j. We recall the definition of stochastic ordering.

Definition 3.2. Let U_1 and U_2 be two random variables whose distributions are dominated by a common measure μ. We say that U_1 is stochastically no greater than U_2 if, for each $c \in \mathbb{R}$, $P(U_1 \leq c) \geq P(U_2 \leq c)$. In this case we write $U_1 \overset{\mathcal{D}}{\leq} U_2$. If in addition, $\mu(\{c : P(U_1 \leq c) > P(U_2 \leq c)\}) > 0$ then we say that U_1 stochastically (strictly) less than U_2, and write $U_1 \overset{\mathcal{D}}{<} U_2$.

This version of the definition of stochastic ordering is used, for example, in Li, Zha and Chiaromonte (2005). In general, inequality (3.11) is neither stronger nor weaker than stochastic ordering. However, in a special case it is weaker than stochastic ordering, as we show below.

For $i = 1, 2$, let F_i be the distribution of U_i and f_i be the density of U_i with respect to μ. We say that U_1 and U_2 have a common support if $\{f_1 > 0\} = \{f_2 > 0\}$. It is easy to see that if $U_1 \overset{D}{<} U_2$ and U_1 and U_2 have a common support, then

$$F_1(\{c : F_1(c) > F_2(c)\}) > 0, \quad F_2(\{c : F_1(c) > F_2(c)\}) > 0. \tag{3.12}$$

The following proposition gives a sufficient condition for $U_1 \overset{\mathcal{D}}{<} U_2$ to imply $P(U_1 \leq U_2) > 1/2$.

Proposition 3.1. *Suppose U_1 and U_2 are random variables whose distributions are dominated by a common measure μ; $U_1 \overset{\mathcal{D}}{<} U_2$; $U_1 \perp\!\!\!\perp U_2$; and U_1 and U_2 have a common support. Then $P(U_1 \leq U_2) > 1/2$.*

Proof. By independence of U_1 and U_2 and by Fubini's Theorem,

$$P(U_1 \leq U_2) = \int_{\mathbb{R}} \left[\int_{u_1 \leq u_2} f_1(u_1)\mu(du_1) \right] f_2(u_2)\mu(du_2)$$
$$= \int_{\mathbb{R}} F_1(u_2) f_2(u_2)\mu(du_2) = \int_{\mathbb{R}} F_1(u_2) dF_2(u_2).$$

By the second inequality in (3.12) the right hand side above is (strictly) greater than

$$\int_{\mathbb{R}} F_2(u_2) dF_2(u_2) = \left[\frac{F_2^2(u_2)}{2} \right]_{-\infty}^{\infty} = \frac{1}{2},$$

which completes the proof.

We must point out that the natural tendency described in this paper is neither definite nor particularly strong, and there is much room for improvement by sufficient dimension reduction, which reduces the dimension of $\boldsymbol{X}$ in reference to Y. See Cook (2007).

Acknowledgements

We are grateful to a referee, an associate editor, and the Editor for their very helpful and prompt reviews. The research is supported in part by a National Science Foundation grant (DMS-0704621) to Bing Li.

References

Alter, O., Brown, P. and Botstein, D. (2000). Singular value decomposition for gene-wide expression data processing and modelling. *Proc. Nat. Acad. Sci.* **97**, 10101-10106.

Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics* **19**, 1252-1258.

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. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22**, 1-40.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction without matrix inversion. *Biometrika* **94**, 569–584.

Cox, D. R. (1968). Notes on some aspects of regression analysis. *J. Roy. Statist. Soc. Ser. A* **131**, 265-279.

Hawkins, D. M. and Fatti, L. P. (1984). Exploring multivariate data using the minor principal components. *The Statistician* **33**, 325–338.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1-49.

Hotelling, H. (1957). The relationship of the newer multivariate statistical methods to factor analysis. *British J. Math. Statist. Psych.* **10**, 69-79.

Kendall, M. G. (1957). *A course in Multivariate Analysis*. Griffin, London.

Li, B. (2007). Comment: Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22**, 32-35.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

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

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, Massachusetts.

Scott, D. (1992). *Multivariate Density Estimation*. Wiley, New York.

Department of Statistics, Pennsylvania State University, 325 Thomas Building, Universtiy Park, PA 16802, U.S.A.

E-mail: aaa195@stat.psu.edu

Department of Statistics, Pennsylvania State University, 410 Thomas Building, University Park, PA 16802, U.S.A.

E-mail: bing@stat.psu.edu

(Received February 2008; accepted April 2008)