Journal of Multivariate Analysis 119 (2013) 176–184

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

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

Andreas Artemiou [a,*], Bing Li [b]

[a] Department of Mathematical Sciences, Michigan Technological University, United States

[b] Department of Statistics, Pennsylvania State University, United States

ARTICLE INFO

Article history:
Received 3 December 2012
Available online 3 May 2013

AMS subject classifications:
62A01
62B10
62H99
62G08

Keywords:
Permutation invariance
Principal component analysis
Rotation invariance
Single-index model
Sufficient dimension reduction

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 of why it is often beneficial to perform 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) [19], Artemiou and Li (2009) [2], and Ni (2011) [24], where the same phenomenon was conjectured and rigorously demonstrated for linear regression.

1. Introduction

Principal component analysis has been used for dimension reduction for regression problems ever since its introduction by Pearson [25] and Hotelling [14]. 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 ill-conditioned matrix inversion. However, the validity of this tactic, often referred to as the Principal Component Regression [17], 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 $\boldsymbol{\Sigma}$ 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 [4] in the context of Fitted Principal Components. See also [2,12].

Artemiou and Li [2], inspired by a conjecture by Li [19], 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 $\boldsymbol{\Sigma}$, then, under the linear regression model

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

* Correspondence to: 306 Fisher Hall, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, 906-487-2884, United States.
E-mail addresses: aartemio@mtu.edu, artemiou@gmail.com (A. Artemiou).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.04.015

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 $\boldsymbol{\Sigma}$ corresponding to its ith and jth largest eigenvalues, with $i < j$. Then the probability of

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

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 $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$.

Using an invariant argument by Arnold and Brockett [1] and a stronger invariant assumption, Ni [24] 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 $\boldsymbol{\Sigma}$. 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 [24] also generalized this result in several other directions, but confined his analysis to linear regression.

From a different angle, Hall and Yang [12] 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 [26,13,16]. 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

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

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

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, [20,5–8,28,23].

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, [3,22]. 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 [24] are somewhat stronger than those used in [2]: the former involves permutation invariance and the latter involves rotation invariance. In sections and 3 and 4 we generalize the stronger result of [24] 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 [2] under permutation invariance. These are followed by a short discussion in Section 6.