Statistica Sinica **21** (2011), 741-747

PRINCIPAL COMPONENT REGRESSION REVISITED

Liqiang Ni

University of Central Florida

Abstract: Principal component regression has been perceived as a remedy for multicollinearity. Cook (2007) suggested that principal components and related methodology actually play a broader role than previously thought. Recently, Artemiou and Li (2009) provided a probabilistic explanation of the phenomenon that the response is often highly correlated with the leading principal components of the predictors. This article reinforces the previous results and offers an alternative perspective.

Key words and phrases: Dimension reduction, principal components, regression, spherically symmetric distribution.

1. Introduction

Principal component regression has been used widely for years (Kendall (1957), Massy (1965)) with much emphasis on dealing with the collinearity among predictors. Recently, Cook (2007) argued that the role of principal components and related methodology may be broader than previously seen. Commenting on Cook (2007), Li (2007) conjectured that if nature arbitrarily selects a covariance matrix for the predictor and coefficients for the regression, then the principal components of higher rank tend to have stronger correlations with the response than those of lower rank. This conjecture subsequently was proved by Artemiou and Li (2009), which helps explain the fact that the response is often highly correlated with the leading principal components, even when there is no logical reason for this connection.

Consider a linear model with an additive random error,

$$Y = \beta^T X + \epsilon, \tag{1.1}$$

where $E(X) = 0$, $\mathrm{var}(X) = \Sigma$, $E(\epsilon) = 0$, and $\mathrm{var}(\epsilon) = \sigma^2$, $\mathrm{cov}(X, \epsilon) = 0$. For ease of exposition, we repeat the main result of Artemiou and Li (2009) as Proposition 1:

Proposition 1.(Theorem 3.1 in Artemiou and Li (2009)) *Suppose Σ is a $p \times p$ orientationally uniform random matrix, β is a p-dimensional random vector such* *that* $\beta \perp\!\!\!\perp (X,\Sigma)$ *and* $\epsilon \perp\!\!\!\perp (X,\beta,\Sigma)$ (*here* $\perp\!\!\!\perp$ *indicates independence*), *and* $\Pr(\beta \in G) > 0$ *for any nonempty open set* $G \in \mathbb{R}^p$. *Let* ρ_i *be the squared correlation coefficient between* Y *and the* i*th principal component of* X. *Then, if* $i < j$, $\Pr(\rho_i \geq \rho_j) > 1/2$.

Here a random matrix is said to be orientationally uniform if its eigenvalues are exchangeable random variables, its eigenvectors are exchangeable random vectors, and the eigenvalues are independent of the eigenvectors. The proposition provides a partial explanation of the popularity of the principal component regression. However, as pointed out in the rejoinder of Cook (2007) the leading principal components may not be a sufficient reduction for the regression since they do not take account of the information of Y.

In this article, we discuss two cases: the first focuses on the orientational uniformness of the covariance of X, which follows the path of Artemiou and Li (2009); the second focuses on the randomness of the regression coefficients as an alternative explanation. Both could be reasonable manifestations of the "unbiasedness" of nature. We show that both cases lead to the conclusion that the principal components of higher rank tend to have stronger correlations with the response than those of lower rank.

Working on model (1.1), we introduce some notation. Suppose Σ has a spectral decomposition $\Sigma = U\Lambda U^T$, where U is orthonormal, and Λ is a diagonal matrix. For any vector $\alpha \in \mathbb{R}^p$, let

$$f(\alpha,\beta) = \operatorname{cor}^2(\alpha^T X, Y) = \frac{(\alpha^T U\Lambda U^T \beta)^2}{(\alpha^T U\Lambda U^T\alpha)(\beta^T U\Lambda U^T\beta + \sigma^2)}.$$

2. Orientationally Uniform Σ

Suppose that Σ is an orientationally uniform random matrix. For any set of orthonormal vectors $\{v_1,\ldots,v_p\}$, any random variable among $v_1^T X, \ldots, v_p^T X$ is equally likely to be the first, second, ..., or pth principal component of X. In other words, there is no particular preference among predictors in how they organize themselves. If so, we have Theorem 1 which reinforces Proposition 1.

Theorem 1. *Suppose* β *is a fixed vector or a measurable random vector independent of* (X,ϵ), *and the covariance matrix* Σ *is an orientationally uniform random matrix. Let* $\lambda_{(i)}$ *and* $U_{(i)}$ *denote its* i*th largest eigenvalue and the corresponding eigenvector. For* $i < j$,

$$\Pr[\operatorname{cor}^2(U_{(i)}^T X, Y|\beta) \geq \operatorname{cor}^2(U_{(j)}^T X, Y|\beta)] = \frac{2}{\pi} E\Big[\arctan\Big(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\Big)\Big] \geq \frac{1}{2}.$$

Proof. Let U_m and U_l denote any generic pair of columns of U. We have

$$\begin{aligned}
&\Pr[f(U_{(i)},\beta) \geq f(U_{(j)},\beta)] \\
&= \Pr\Big[\frac{(U_{(j)}^T\beta)^2}{(U_{(i)}^T\beta)^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}}\Big] \\
&= \Pr\Big[\frac{(U_m^T\beta)^2}{(U_l^T\beta)^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}}\Big] \\
&= E\Big\{\Pr\Big[\frac{(U_m^T\beta)^2}{(U_l^T\beta)^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}}\Big|\beta\Big]\Big\} \\
&= E\Big\{\Pr\Big[\frac{U_m^T\beta}{U_l^T\beta} \leq \sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\Big|\beta, U_m^T\beta > 0, U_l^T\beta > 0\Big]\Big\} \\
&= \frac{2}{\pi}E\Big[\arctan\Big(\frac{\sqrt{\lambda_{(i)}}}{\lambda_{(j)}}\Big)\Big] \\
&\geq \frac{1}{2}.
\end{aligned}$$

The second equality holds because of the independence between U and Λ by the orientationally uniformness. The fourth equality follows the geometric argument in the proof of Theorem 1 in Arnold and Brockett (1992). Conditioning on β, the random vector $U^T\beta$ is uniformly distributed on a p-dimensional hypersphere with a radius $\|\beta\|$. Because of the symmetry, we need only consider the first quadrant. The fifth equality follows the examination of the bivariate joint distribution of $(U_m^T\beta, U_l^T\beta)$ that only depends on the distance to the origin. Therefore, the probability we search for should be proportional to the angle of the area. It is easy to see the proportionality constant should be the inverse of $\pi/2 = \arctan(+\infty)$.

The leading principal component is often of special interest. Theorem 1 states that the first principal component tends to have higher correlation with the response than any other component. In some situations, we can calculate this probability. Consider the simulation example in Li (2007, pp. 33) where $p = 2$, λ_1 and λ_2 are i.i.d. Uniform$(0, c)$, $c > 0$. Let ρ_i denote the squared correlation between the ith principal component and the response. Based on Theorem 1, we have

$$\Pr[\rho_1 > \rho_2] = 2 * \frac{2}{\pi}\int_0^1\int_v^1 \arctan(\sqrt{\frac{u}{v}})dudv = \frac{2}{\pi} = 0.6366.$$

Li (2007) reported a sample estimate of 0.65 with 200 replicates that is quite close to the true probability.

Table 1. Estimated Probabilities. (a) $\Pr[\rho_1 > \rho_2]$; (b) $\Pr[\rho_1 = \max \rho_i]$; (c) $p \Pr[\rho_1 = \max \rho_i]$.

	$\lambda \sim Beta(1,3)$			$\lambda \sim Beta(1,1)$			$\lambda \sim Beta(3,1)$		
p	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
2	0.6647	0.6647	1.3293	0.6383	0.6383	1.2765	0.5532	0.5532	1.1064
3	0.5551	0.5306	1.5918	0.5423	0.4897	1.4691	0.5129	0.3937	1.1810
4	0.5259	0.4538	1.8152	0.5174	0.4055	1.6220	0.5052	0.3115	1.2458
6	0.5116	0.3635	2.1809	0.5033	0.3098	1.8590	0.5030	0.2229	1.3375
8	0.5046	0.3134	2.5075	0.5026	0.2508	2.0064	0.5008	0.1753	1.4022
16	0.5012	0.2154	3.4466	0.5003	0.1528	2.4442	0.5043	0.0981	1.5702

It may also be of interest to obtain the probability that the leading principal component has the highest correlation among all components, i.e. $\Pr[\rho_1 = \max \rho_i]$. Under the assumptions in Theorem 1, the probability depends on both the dimension p and the distributions of λ_i's. While a generic closed-form solution seems elusive, we can always estimate it by Monte Carlo simulation. Table 1 provides the result of a simulation study based on 100,000 replicates, where the eigenvalues are i.i.d. from $Beta(1,3)$, $Beta(1,1)$ (uniform), and $Beta(3,1)$, respectively. For any setting, Column (a) shows the advantage of the first component over the second component; Column (b) shows the probability that the first component has the highest correlation among all components. As expected, these probabilities are decreasing as the number of predictors increases. However, if one considers the ratio of Column (b) to $1/p$, the fair share for any component if no partiality exists, Column (c) suggests the advantage of the leading component actually is strengthening rather than diminishing. Meanwhile, the comparison of the sections of the Table clearly demonstrates the impact of the distribution of λ's on the probabilities. The Beta distributions were used only for illustration here. However, the stipulation of orientationally uniform covariance provides a framework where the exchangeability and the unspecified distribution of eigenvalues offer a great deal of flexibility.

3. Spherically Symmetric β

Assume that β is a random vector that is spherically symmetric, i.e., $\Gamma\beta$ has the same distribution as β for any orthonormal matrix Γ. In words, regardless of the structure of predictors, the response has no preference among the predictors in terms of the magnitude of the linear coefficients. As a first step, we consider a fixed Σ. For simplicity of the notation, let λ_i indicate the ith largest eigenvalue.

For $i < j$,

$$\begin{aligned}\Pr[f(U_i,\beta) \geq f(U_j,\beta)] &= \Pr[\lambda_i (U_i^T\beta)^2 \geq \lambda_j (U_j^T\beta)^2] \\ &= \Pr\Big[\frac{(U_j^T\beta)^2}{(U_i^T\beta)^2} \leq \frac{\lambda_i}{\lambda_j}\Big] \\ &= \frac{2}{\pi}\arctan\Big(\sqrt{\frac{\lambda_i}{\lambda_j}}\Big),\end{aligned}$$

which is no less than 1/2 since $\lambda_i \geq \lambda_j$. The third equality follows the logic of the proof of Theorem 1. In other words, the leading eigenvectors are more likely to have closer correlations with the response than any other eigenvector.

We can take a further step to show that the leading eigenvector has highest expectation of the squared correlation with the response among all possible directions without knowledge of β. Let $a = U^T\alpha$ and $b = U^T\beta$. Searching for a decision rule α is equivalent to searching for $a = U^T\alpha$. Note that b also has the spherically symmetric distribution. Let

$$g(\alpha) = E_\beta[f(\alpha,\beta)] = E_\beta\Big[\frac{\sum \lambda_i^2 a_i^2 b_i^2}{(\sum \lambda_i a_i^2)(\sum \lambda_i b_i^2 + \sigma^2)}\Big] = \sum_i w_i E(v_i), \qquad (3.1)$$

where $w_i = \lambda_i a_i^2 / \sum \lambda_j a_j^2$, $v_i = \lambda_i b_i^2 / (\sum \lambda_j b_j^2 + \sigma^2)$. The second equality holds because $E[\lambda_i\lambda_j a_i a_j b_i b_j] = 0$ when $i \neq j$. Obviously $\sum w_i = 1$, so $\min_i E(v_i) \leq g(\alpha) \leq \max_i E(v_i)$. Based on the Lemma 1 in the Appendix, we have $E(v_1) = \max E(v_i)$ and $E(v_p) = \min E(v_i)$. It is easy to see that $g(\cdot)$ reaches a maximum when $\alpha = U_1$,

$$g(U_1) = E\Bigg[\frac{\lambda_1 (U_1^T\beta)^2}{\Big(\sum \lambda_i (U_i^T\beta)^2 + \sigma^2\Big)}\Bigg].$$

Similarly, we can show that the $g(\cdot)$ reaches a minimum at $\alpha = U_p$. These results with a fixed Σ can be generalized to a random Σ, as summarized in Theorem 2.

Theorem 2. *Suppose β is a random spherically symmetric vector. Let $\lambda_{(i)}$ and $U_{(i)}$ denote the ith largest eigenvalue and its eigenvector of either a random or a fixed covariance matrix $\Sigma = \mathrm{var}(X)$. For $i < j$,*

$$\Pr[\mathrm{cor}^2(U_{(i)}^T X, Y|\Sigma) \geq \mathrm{cor}^2(U_{(j)}^T X, Y|\Sigma)] = \frac{2}{\pi}E\Bigg[\arctan\Bigg(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}}\Bigg)\Bigg] \geq \frac{1}{2}.$$

Moreover, $E[cor^2(U_{(1)}^T X, Y|\Sigma)] = \max_\alpha E[cor^2(\alpha^T X, Y|\Sigma)]$, where α is restricted to be a function of Σ.

Proof. Let U_m and U_l denote any generic pair of columns of U. We have

$$\begin{aligned}
&\Pr[f(U_{(i)},\beta) \geq f(U_{(j)},\beta)] \\
&= E\Big\{ \Pr\Big[\frac{(U_{(j)}^T\beta)^2}{(U_{(i)}^T\beta)^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}} \Big| \Lambda, U\Big]\Big\} \\
&= E\Big\{ \Pr\Big[\frac{(U_m^T\beta)^2}{(U_l^T\beta)^2} \leq \frac{\lambda_{(i)}}{\lambda_{(j)}} \Big| \Lambda\Big]\Big\} \\
&= \frac{2}{\pi} E\Big[\arctan(\sqrt{\frac{\lambda_{(i)}}{\lambda_{(j)}}})\Big] \\
&\geq \frac{1}{2}.
\end{aligned}$$

The second equality holds because $U^T\beta$ has the same spherically symmetric distribution as β, regardless of U. Based on (3.1), we know that conditioning on $\text{var}(X)$, $\alpha = U_{(1)}$ maximizes $g(\alpha)$, i.e., $U_{(1)}$ has the highest expectation of the squared correlation.

Acknowledgement

The author is grateful to Andreas Artemiou and Bing Li for providing their manuscript. He also wishes to thank the Editor, an associate editor, and two referees for their helpful comments, which have substantially improved the paper. This work was supported in part by National Science Foundation grant DMS 0805409.

Appendix

Lemma 1. *Suppose that $\{W_1, \ldots, W_n\}$ are nonnegative exchangeable random variables. For any constants $a_1 \geq \cdots \geq a_n \geq 0$ and $c > 0$, with*

$$E\Big[\frac{a_1W_1}{(\sum a_iW_i + c)}\Big] \geq E\Big[\frac{a_kW_k}{\Big(\sum a_iW_i + c\Big)}\Big], k = 2, \ldots, n.$$

Proof. Let $T = \sum a_iW_i + c$, and

$$T_k = \sum_{i \notin \{1,k\}} a_iW_i + (a_1 + a_k)\frac{(W_1 + W_k)}{2} + c.$$

Since the W_i are exchangeable, $E[(a_1 + a_k)(W_1 - W_k)/T_k] = 0$. We only need to show that

$$E\Big[\frac{(a_1W_1 - a_kW_k)}{T} - \frac{(a_1 + a_k)}{2}\frac{(W_1 - W_k)}{T_k}] \geq 0,$$

which is equivalent to

$$\begin{aligned}&E\{(a_1W_1-a_kW_k)T_k-\frac{(a_1+a_k)}{2}(W_1-W_k)T\}\\&\quad=E\Big\{\frac{1}{2}(a_1-a_k)(W_1+W_k)\Big(\sum_{i\notin\{1,k\}}a_iW_i+c\Big)+(a_1+a_k)W_1W_k\Big\}\geq 0.\end{aligned}$$

The above inequality holds since all W_i are nonnegative.

References

Arnold, B. C. and Brockett, P. L. (1992). On distributions whose component ratios are Cauchy. *Amer. Statist.* **46**, 25-26.

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

Cook, R. D. (2007). Fisher Lecture: dimension reduction in regression (with discussions). *Statist. Sci.* **22**, 1-43.

Kendall, M. G. (1957). *A Course in Multivariate Analysis.* Charles Griffin & Company, London.

Li, B. (2007). Comment on Cook (2007). *Statist. Sci.* **22**, 32-35.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60**, 234-256.

Department of Statistics, University of Central Florida, Orlando, FL 32816-2370, USA.

E-mail: lni@mail.ucf.edu

(Received May 2009; accepted October 2009)