An inequality for Principal Components and Regression

Andreas Artemiou and *Bing Li*
Department of Statistics, Pennsylvania State University

1. Introduction

Cook (2007) presented an interesting historical debate regarding the relationship between the regression of a scalar response Y on a random vector X and the principal components of X.
The debate arose from the practice to regress Y on the first few principal components of X. While some researchers support this practice like Mosteller and Tukey (1977, page 307) and Kendall (1957, page 75) there are also researchers who question it like Cox (1968, page 272) and Hotelling (1957). The main reason is that "A difficulty seems to be as there is no logical reason why the dependent variable should not be closely tied to the least principal component", as Cox (1968) phrases it. It is obvious that since the calculation of principal components is based on the eigenvectors of the covariance matrix of X that the concern mentioned above is a legitimate one.
As regression problems where p > n arise even more frequently in research, people use even more the method of principal components to reduce the dimension of their regression problems (see for example Chiaromonte and Martinelli (2002)), so there is an even bigger necessity to justify the principal component analysis procedure

2. Previous work

In the discussion of Cook (2007), Li (2007) has stated a conjecture that higher order principal components will be more likely to be more correlated with the response than the lower order ones. Artemiou and Li (to appear) formulated the conjecture into a theorem for the simple case of linear regression, under mild assumptions. That theorem appears here:

Theorem

Suppose $\boldsymbol{\Sigma}$ is a $\boldsymbol{p} \times \boldsymbol{p}$ orientationally uniform random matrix, and

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

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

3. Extensions

In this work we present extensions od the above theorem in three different directions:

- For nonlinear regression settings where the link function is unknown.
- For multivariate response variables (in both linear and non-linear settings).
- For fixed covariance matrices (in both linear and non-linear regression settings) under the assumption of spherically distributed regression coefficients.

3.1 Definition of multivariate correlation

First we give the definition of multivariate correlation:

Definition

The square multiple correlation between $\boldsymbol{U}$, a $\boldsymbol{p}$-dimensional random vector, and $\boldsymbol{V}$, a $\boldsymbol{q}$-dimensional random vector, is defined as follows

$$\mathbf{mcorr}^2(\boldsymbol{U}, \boldsymbol{V}) = \mathbf{tr}\left(\boldsymbol{\Sigma}_U^{-\frac{1}{2}}\boldsymbol{\Sigma}_{UV}\boldsymbol{\Sigma}_V^{-1}\boldsymbol{\Sigma}_{VU}\boldsymbol{\Sigma}_U^{-\frac{1}{2}}\right)$$

where $\boldsymbol{\Sigma}_U$, $\boldsymbol{\Sigma}_V$ are the covariance matrices or $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively, and $\boldsymbol{\Sigma}_{UV}$, is the covariance matrix between $\boldsymbol{U}$ and $\boldsymbol{V}$.

3.2 Fixed covariance matrix and multivariate response

Here, I present two useful results that are needed in order to prove the main theorem, that follows. In later theorems similar results are also needed but are omitted due to space limitations. Both results are basically formulating the conditions under which a ratio between two variables of interest will have unique median equal to 1.

Lemma

Let $\boldsymbol{\Sigma}$ $\boldsymbol{p} \times \boldsymbol{p}$ and $\boldsymbol{\Gamma}$ $\boldsymbol{q} \times \boldsymbol{q}$, fixed matrices, β be a $\boldsymbol{p} \times \boldsymbol{q}$ random matrix with spherical distribution, $\boldsymbol{a}, \boldsymbol{b}$ be p-dimensional known vectors, eigenvectors of $\boldsymbol{\Sigma}$ and $\mathbf{D} = \left(\beta^T\boldsymbol{\Sigma}\beta + \boldsymbol{\Gamma}\right)^{-1/2}$. Then the random variables $\boldsymbol{a}^T\beta\mathbf{D}\mathbf{D}^T\beta^T\boldsymbol{a}$ and $\boldsymbol{b}^T\beta\mathbf{D}\mathbf{D}^T\beta^T\boldsymbol{b}$ are exchangeable.

Lemma

Suppose that β is a $\boldsymbol{p} \times \boldsymbol{q}$ spherically distributed random matrix, $\boldsymbol{\Gamma}$ is a $\boldsymbol{q} \times \boldsymbol{q}$ fixed matrix, $\boldsymbol{\Sigma}$ $\boldsymbol{p} \times \boldsymbol{p}$ fixed matrix and $\boldsymbol{v_1}, \ldots, \boldsymbol{v_p}$ are $\boldsymbol{p}$-dimensional fixed vectors (eigenvectors of $\boldsymbol{\Sigma}$). Define $\boldsymbol{A} = \beta\left(\beta^T\boldsymbol{\Sigma}\beta + \boldsymbol{\Gamma}\right)^{-\frac{1}{2}}$. If $\boldsymbol{P}(\beta \in \boldsymbol{G}) > \boldsymbol{0}$ for any nonempty open set $\boldsymbol{G}$, then for any chosen pair $(\boldsymbol{v_i}, \boldsymbol{v_j})$,$\boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{p}, \boldsymbol{i} \neq \boldsymbol{j}$ the random variable $\boldsymbol{v_i}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v_i}/\boldsymbol{v_j}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v_j}$ has unique median, which equals 1.

Theorem

Suppose $\boldsymbol{\Sigma}$ fixed $\boldsymbol{p} \times \boldsymbol{p}$ and $\boldsymbol{\Gamma}$ fixed $\boldsymbol{q} \times \boldsymbol{q}$ matrices. Let $\boldsymbol{X}$ is a $\boldsymbol{p}$-dimensional random vector with $\boldsymbol{E}(\boldsymbol{X}) = \boldsymbol{0}$ and $\mathbf{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$; β is a $\boldsymbol{p} \times \boldsymbol{q}$ spherically distributed random matrix independent of $\boldsymbol{X}$ and $\boldsymbol{P}(\beta \in \boldsymbol{G}) > \boldsymbol{0}$ for any nonempty open set $\boldsymbol{G}$; ϵ is a $\boldsymbol{q}$ dimensional random vector independent of $(\boldsymbol{X}, \beta)$ with $\boldsymbol{E}(\epsilon) = \boldsymbol{0}, \mathbf{var}(\epsilon) = \boldsymbol{\Gamma}$. Let $\boldsymbol{Y} = \beta^T\boldsymbol{X} + \epsilon$, $\boldsymbol{w_1}, \ldots, \boldsymbol{w_p}$ be the $\boldsymbol{1}^{\mathbf{st}}, \ldots, \boldsymbol{p}^{\mathbf{th}}$ principal components of $\boldsymbol{\Sigma}$, and let $\rho_i = \rho_i(\beta) = \mathbf{mcorr}^2(\boldsymbol{Y}, \boldsymbol{w_i}|\beta)$.Then, whenever $\boldsymbol{i} < \boldsymbol{j}$, $\boldsymbol{P}(\rho_i \geq \rho_j) > \boldsymbol{1/2}$.

3.3 Nonlinear regression assuming conditional independence

Theorem

Suppose

1. *$\boldsymbol{X}$ is a $\boldsymbol{p}$-dimensional random vector with $\boldsymbol{E}(\boldsymbol{X}) = \boldsymbol{0}$ and $\mathbf{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$; and $\boldsymbol{v_1}, \ldots, \boldsymbol{v_p}$ are the ordered eigenvectors of $\boldsymbol{\Sigma}$ in the sense that $\boldsymbol{v_1}$ corresponds to the largest eigenvalue λ_1, $\boldsymbol{v_2}$ corresponds to λ_2 and so on, were $\lambda_1 \geq \ldots \geq \lambda_p$;*
2. *β is a $\boldsymbol{p} \times \boldsymbol{q}$ spherically distributed random matrix and $\beta \perp\!\!\!\perp \boldsymbol{X}$;*
3. *$\boldsymbol{E}(\boldsymbol{X}|\beta^T\boldsymbol{X}, \beta)$ is linear in $\beta^T\boldsymbol{X}$.*

If $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|(\beta^T\boldsymbol{X}, \beta)$, then, for any square-integrable function $\boldsymbol{f}(\boldsymbol{Y})$ such that $\boldsymbol{E}[\boldsymbol{X}\boldsymbol{f}(\boldsymbol{Y})|\beta] \neq \boldsymbol{0}$, $\mathbf{var}[\boldsymbol{f}(\boldsymbol{Y})|\beta]$ positive definite matrix and $\boldsymbol{i} < \boldsymbol{j}$ we have $\boldsymbol{P}(\rho_i > \rho_j) > \boldsymbol{1/2}$, where $\rho_i = \mathbf{mcorr}^2(\boldsymbol{f}(\boldsymbol{Y}), \boldsymbol{v_i}^T\boldsymbol{X}|\beta)$.

3.4 Nonlinear regression assuming $\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X}, \beta) = \boldsymbol{E}(\boldsymbol{Y}|\beta^T\boldsymbol{X}, \beta)$

Theorem

Suppose

1. *$\boldsymbol{X}$ is a $\boldsymbol{p}$-dimensional random vector with $\boldsymbol{E}(\boldsymbol{X}) = \boldsymbol{0}$ and $\mathbf{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$; and $\boldsymbol{v_1}, \ldots, \boldsymbol{v_p}$ are the ordered eigenvectors of $\boldsymbol{\Sigma}$ in the sense that $\boldsymbol{v_1}$ corresponds eigenvalue λ_1, $\boldsymbol{v_2}$ corresponds to λ_2 and so on, were $\lambda_1 \geq \ldots \geq \lambda_p$*
2. *β is a $\boldsymbol{p} \times \boldsymbol{q}$ is a spherically distributed random matrix and $\beta \perp\!\!\!\perp \boldsymbol{X}$;*
3. *$\boldsymbol{E}(\boldsymbol{X}|\beta^T\boldsymbol{X}, \beta)$ is linear in $\beta^T\boldsymbol{X}$.*
4. *$\boldsymbol{Y}$ is a random vector such that $\boldsymbol{E}(\boldsymbol{Y}|\beta) = \boldsymbol{0}$, $\mathbf{var}(\boldsymbol{Y}|\beta)$ positive definite matrix.*

If $\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X}, \beta) = \boldsymbol{E}(\boldsymbol{Y}|\beta^T\boldsymbol{X}, \beta)$, then for $\boldsymbol{i} < \boldsymbol{j}$, $\boldsymbol{P}(\rho_i > \rho_j) > \boldsymbol{1/2}$, where $\rho_i = \mathbf{mcorr}^2(\boldsymbol{Y}, \boldsymbol{v_i}^T\boldsymbol{X}|\beta)$.

4. Conclusions

In this work we have extended a previous work by Artemiou and Li (to appear) in different ways, which make the result stronger. Our main contribution is the extension of the previous results in the non-linear regression functions, in two different ways; one assuming conditional independence and the other assuming $\boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X}, \beta) = \boldsymbol{E}(\boldsymbol{Y}|\beta^T\boldsymbol{X}, \beta)$. Other important extensions are the ones that assumes multivariate response $\boldsymbol{Y}$ and the ones that involve fixed covariance matrix $\boldsymbol{\Sigma}$. All the results shown are true for both multivariate and univariate response variables.
As Artemiou and Li (to appear) pointed out, this inequality shows a natural tendency and is not definite nor particularly strong. It can tell us that in the majority of cases the leading principal components will be viable predictors for the response $\boldsymbol{Y}$, but there is an unmeasurable risk that this will not be the case. This fact though, raises other questions that researchers need to attack. One direction towards overcoming this problem has been extensively worked and there are already proposed methods that use the information of $\boldsymbol{Y}$ to reduce the dimension of $\boldsymbol{X}$ as the research in sufficient dimension reduction (Li (1991), Li (1992), Cook (1994), Cook (1996), Li and Wang (2007)).

5. Main references

- Artemiou A. A. and Li B. (to appear). On principal components and regression: A probabilistic explanation of a natural phenomenon. *Statistica Sinica*. Accepted.
- 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. *Statistical Science*, **22**, 1–40.
- Cook, R.D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, 18–25.
- Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.
- Cox, D. R. (1968). Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society, Ser. A.*, **131**, 265–279.
- Hotelling, H. (1957). The relationship of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, **10**, 69–79.
- Kendall, M. G. (1957). *A course in Multivariate Analysis*. London: Griffin.
- Li, B. (2007). Comment: Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 32–35.
- 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, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma. *Journal of the American Statistical Association*, **87**, 1025–1039.
- Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Reading, Massachusetts: Addison-Wesley.

Special thanks to the Department of Statistics and the Harkness Statistics Enhancement Fund for funding my poster presentation in this Conference.