Statistical Science
2007, Vol. 22, No. 1, 32–35
DOI: 10.1214/088342307000000069
Main article DOI: 10.1214/088342306000000682

Comment: Fisher Lecture: Dimension Reduction in Regression

Bing Li

This paper puts dimension reduction into the historical context of sufficiency, efficiency and principal component analysis, and opens up an avenue toward efficient dimension reduction via maximum likelihood estimation of inverse regression. I congratulate Professor Cook for this insightful and groundbreaking work. My discussion will focus on two points that explore and extend Cook's ideas. The first is about the relationship between the principal component analysis of the predictor and the regression of the response on the predictor; the second explores various ways of extending Cook's inverse regression to characterize and estimate variance components.

1. PCA OF X AND REGRESSION OF Y

In his paper Professor Cook has told an intriguing and fascinating history of the opposing views regarding the relationship between the principal component analysis of X and the regression of Y on X. On the one hand, it is often the case in practice that the first few principal components of X tend to have higher correlations with Y than the other principal components of X, but on the other hand there seems no logical reason to believe that the direction along which X varies the most should somehow have a relation with Y. In this section I ask, and attempt to answer, the following question: is it possible for the first principal component of X to have higher correlation with Y (than the other principal components of X) even if nature is "neutral" in assigning a relation between X and Y and "arbitrary" in assigning a covariance matrix to X?

To pursue this curiosity let us consider the following situation. Let $\mathbb{R}_+^{p\times p}$ be the collection of all p by p positive definite matrices, and let F be a distribution over $\mathbb{R}_+^{p\times p}$ that is in some sense uniform. Suppose nature randomly selects a covariance matrix Σ according to F, and generates X from $N(0,\Sigma)$. Furthermore, suppose that nature selects a linear relation between X and Y completely independently of the way it selected Σ; that is, $Y=\beta^T X+\varepsilon$, where β is a random vector in $\mathbb{R}^p$, $\beta \perp\!\!\!\perp (\Sigma, X)$, and $\varepsilon \perp\!\!\!\perp (X,\beta,\Sigma)$ (here $\perp\!\!\!\perp$ indicates independence). Let $v_1,\dots,v_p$ be the eigenvectors of the random matrix Σ, arranged so that their eigenvalues satisfy $\lambda(v_1)\ge\cdots\ge\lambda(v_p)$. Let $\rho_i(\beta,\Sigma)$ be the correlation coefficient between $v_i^T X$ and Y, conditioning on β and Σ. Thus $\rho_1(\beta,\Sigma),\dots,\rho_p(\beta,\Sigma)$ are random variables depending on β and Σ. The question is: does $|\rho_1(\beta,\Sigma)|$ in any sense tend to be larger than $|\rho_2(\beta,\Sigma)|,\dots,|\rho_p(\beta,\Sigma)|$?

To make the situation as simple as possible we take $p=2$. We consider two ways of generating Σ "uniformly" over $\mathbb{R}_+^{2\times 2}$. Let λ_1,λ_2 be i.i.d. $U(0,c)$, where c is a large number, say $c=1000$. Let A be a random rotation matrix, say

$$A=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix},$$

where $\theta\sim U(0,2\pi)$ and $\theta \perp\!\!\!\perp (\lambda_1,\lambda_2)$. Let

$$\Sigma = A[\operatorname{diag}(\lambda_1,\lambda_2)]A^T.$$

Intuitively, we first create a horizontal (or vertical) ellipse with arbitrary lengths of axes and then rotate it to an arbitrary angle θ. Since c is large this provides a reasonable approximation to a uniformly distributed Σ over $\mathbb{R}_+^{2\times 2}$. Let X, β and Y be generated according to the procedure described in the last paragraph, with $\beta\sim N(0,I_p)$. For simplicity, we take $\varepsilon=0$ because it has no bearing on the problem. We compute the probability

$$P\{\rho_1(\beta,\Sigma)>\rho_2(\beta,\Sigma)\} \tag{1}$$

by simulation, as follows. First, generate an i.i.d. sample $(\Sigma_1,\beta_1),\dots,(\Sigma_n,\beta_n)$. For each (β_i,Σ_i), generate an i.i.d. sample $(X_{i1},Y_{i1}),\dots,(X_{im},Y_{im})$. Using this sample we estimate $\rho_1(\beta_i,\Sigma_i)$ and $\rho_2(\beta_i,\Sigma_i)$ by the method of moments. Denote these estimates by

Bing Li is Professor, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, USA (e-mail: bing@stat.psu.edu).