A Fable of PCA

FRED L. RAMSEY*

1. INTRODUCTION

If an investigator believes one factor is an important factor for the problem at hand, several aspects of that factor may be measured. As the forthcoming fable depicts, this practice—in combination with principal components analysis (PCA)—may give results akin to that of the blind men and the elephant.

2. FABLE

Dr. First studied the Big Forest. Some 100 sites were chosen through skilled use of random number tables. A trained crew of five graduate students visited each site, recording (a) native tree biomass, (b) exotic tree biomass, (c) native shrub biomass, (d) exotic shrub biomass, and (e) total Forbes biomass (all Forbes being native). The skill of the design paid off when the correlation matrix shown in Table 1 appeared. In the published account, Dr. First described the Big Forest with powerful simplicity (see Table 2). I quote:

> The major source of variation is from a treeless state to a fully forested state. The next major source is the variation in shrubs from predominantly native to predominantly exotic. There is some variation in the total amount of shrubs, but this is of less importance than the variation in the total amount of Forbes. There is very little variation in the native/exotic composition of the trees.

As good investigators often do, Dr. Second repeated the study—with a little extra. Having a somewhat larger grant, Dr. Second was able to support a sixth student, who was to share responsibilities for recording Forbes. One student was assigned to grasses and one to herbs. Unfortunately, because of a slight misunderstanding and because there were in fact no grasses at all in the Big Forest, both of these students independently measured exactly the same thing: Forbes. Here, then, was Dr. Second's correlation matrix (see Table 3). This published account was also simple (see Table 4). I quote:

Table 1. Correlations in the Big Forest

	Trees		Shrubs		
	Native	*Exotic*	*Native*	*Exotic*	*Forbes*
Trees					
Native	1.0	+.8	.0	.0	.0
Exotic	+.8	1.0	.0	.0	.0
Shrubs					
Native	.0	.0	1.0	−.6	.0
Exotic	.0	.0	−.6	1.0	.0
Forbes	.0	.0	.0	.0	1.0

> The major source of variation is attributable to changes in the total biomass of Forbes from . . . .

Whereupon both Doctors were expelled from the Academy.

MORAL: You usually see what you set out to see.

3. DISCUSSION

Most multivariate analysis textbooks discuss discarding variables following PCA. Jolliffe (1972, 1973) discussed the problem extensively. These discussions focused on the vector(s) associated with the smallest root(s) of the correlation matrix. The discussions did not, however, include the influence of redundant variables on the largest root(s) and associated vector(s). The example here shows that the largest roots may reflect redundancies only. The lesson of the example does not depend on the choice of using a correlation, rather than a covariance, matrix.

Table 2. Principal Components Analysis

Component	*Root*	*%*	*Description*	*Vector*
1	1.8	36	Total tree biomass	[1, 1, 0, 0, 0]
2	1.6	32	Native vs. exotic shrubs	[0, 0, 1, −1, 0]
3	1.0	20	Forbes	[0, 0, 0, 0, 1]
4	.4	8	Total shrub biomass	[0, 0, 1, 1, 0]
5	.2	4	Native vs. exotic trees	[1, −1, 0, 0, 0]

*Fred L. Ramsey is Professor, Department of Statistics, Oregon State University, Corvallis, OR 97331. This article was written when the author was Visiting Professor, Department of Mathematics, The University of Wollongong, Wollongong, New South Wales 2500, Australia.

Had Dr. Second discarded one of the Forbes variables *and then recomputed the PCA*, both Doctors might still

Table 3. Dr. Second's Correlations

	Trees		Shrubs		Forbes	
	Native	Exotic	Native	Exotic	Grasses	Herbs
Trees						
Native	1.0	+.8	.0	.0	.0	.0
Exotic	+.8	1.0	.0	.0	.0	.0
Shrubs						
Native	.0	.0	1.0	−.6	.0	.0
Exotic	.0	.0	−.6	1.0	.0	.0
Forbes						
Grasses	.0	.0	.0	.0	1.0	+1.0
Herbs	.0	.0	.0	.0	+1.0	1.0

flourish in the Academy. This prescription is seldom offered in the textbooks, however.

[*Received February 1986.*]

REFERENCES

Jolliffe, I. T. (1972), "Discarding Variables in Principal Component Analysis I: Artificial Data," *Applied Statistics*, 21, 160–173.

——— (1973), "Discarding Variables in Principal Component Analysis II: Real Data," *Applied Statistics*, 22, 21–31.

Table 4. Principal Components Analysis

Component	Root	%	Description	Vector
1	2.0	33	Total Forbes biomass	[0, 0, 0, 0, 1, 1]
2	1.8	30	Total tree biomass	[1, 1, 0, 0, 0, 0]
3	1.6	27	Native vs. exotic shrubs	[0 0, 1, −1, 0, 0]
4	.4	7	Total shrub biomass	[0, 0, 1, 1, 0, 0]
5	.2	3	Native vs. exotic trees	[1, −1, 0, 0, 0, 0]
6	.0	0	Grasses vs. herbs	[0, 0, 0, 0, 1, −1]