A Note on the Use of Principal Components in Regression

Ian T. Jolliffe

Applied Statistics, Vol. 31, No. 3. (1982), pp. 300-303.

For example, if $N = 300$, $p_0 = 5$ per cent, $L = 2$ per cent and $\alpha = 5$ per cent, the required s is given by equation (2) as

$$\left[\frac{4}{3{\cdot}84 \times 475} + \frac{1}{300}\right]^{-1} \approx 180{\cdot}95,$$

whereas the usual normal approximation to the binomial approximation, $B(p, s)$, yields

$$s = \frac{3{\cdot}84 \times 475}{4} = 1824/4 = 456$$

which is much bigger than the batch it is to be selected from.

References

Gnedenko, B. V. (1967). *Theory of Probability*, 4th ed., translated by B. D. Seckler, pp. 94–103. New York: Chelsea.

Appl. Statist. (1982),
31, *No.* 3, *pp.* 300–303

A Note on the Use of Principal Components in Regression

By Ian T. Jolliffe

University of Kent

[Received February 1982. Revised June 1982]

Summary

The use of principal components in regression has received a lot of attention in the literature in the past few years, and the topic is now beginning to appear in textbooks. Along with the use of principal component regression there appears to have been a growth in the misconception that the principal components with small eigenvalues will very rarely be of any use in regression. The purpose of this note is to demonstrate that these components can be as important as those with large variance. This is illustrated with four examples, three of which have already appeared in the literature.

Keywords: PRINCIPAL COMPONENT ANALYSIS; REGRESSION

The idea of using principal components in regression is not new. Kendall (1957) suggested it in his book on Multivariate Analysis, as did Hotelling (1957) in an article in the same year, and a well-known example was given by Jeffers (1967). The use of principal components envisaged by these authors was to replace the original regressor variables by their principal components, thus orthogonalizing the regression problem and making computations easier and more stable.

More recently the subject has received a lot of attention in the literature, including discussion of various alternatives or modifications to the original idea, and of links with other forms of biased regression. The topic has even began to appear in some student textbooks, e.g. Mosteller and Tukey (1977), Mardia *et al.* (1979) and Gunst and Mason (1980).

0035–9254/82/313000 $2.00

Along with the growth of interest in principal component regression, a misconception seems to be becoming established, concerning the rule for deciding which principal components should be kept in the regression. The original idea was to treat the principal components in the same way as ordinary regressor variables, and assess whether they should be included by computing their association with the dependent variable. However, in several recent publications the suggested rule for inclusion is simply based on the variance of the component, i.e. retain those components with large variances and reject those with small variances.

For example, Mansfield *et al.* (1977, p. 38) suggest that if the only components deleted are those with small variance then there is very little loss of predictiveness in the regression. Some examples which follow later will show that this need not be true.

In the book by Gunst and Mason (1980), 12 pages are devoted to principal component regression, and most of the discussion assumes that deletion of principal components is based solely on the sizes of their variances. An alternative strategy is briefly mentioned but is said to be less useful in practice (pp. 327–328). Note, however, that in a later section of the book, dealing with the rather different technique of latent root regression, the authors implicitly recognise that small-variance components may have predictive value.

Mosteller and Tukey (1977, pp. 397–398) argue similarly that the components with small variance are unlikely to be important in regression, apparently on the basis that nature is "tricky, but not downright mean". We shall see in the examples below that without too much effort we can find examples where nature *is* "downright mean".

Hocking (1976, p. 31) is even firmer in defining a rule for retaining principal components in regression based on size of variance. He says that various authors, including Kendall (1957), Jeffers (1967), Massey (1965) and Hawkins (1973), recommend transforming to principal components and deleting components with small variances. Again this is not true; Jeffers (1967, p. 230) specifically states that relations between the dependent variable and *all* of the components should be examined since it is always possible that one of the components with small variance may be related to the dependent variable. Kendall (1957) is more ambiguous, although he does mention using *t*-tests to test the significance of the components (p. 71) without, at that point, mentioning the variance of components. Hotelling (1957) and also Massy (1965) are unequivocal in saying that even the last (i.e. smallest variance) component can be important in the regression.

Part of the misconception over the role of the low-variance components may have been due to the examples given by Kendall (1957) and Jeffers (1967). These were for a long time the best known examples in the literature and in both cases only the large-variance components *were* important–the first 3 of 5 for Kendall, and the first 6 of 13 for Jeffers. Note, however, that Mardia *et al.* (1979) in their analysis of Jeffers' data claim that some of the later components (seventh, eighth and twelfth out of 13 in order of decreasing variance) are important in the regression too. Jeffers (1981) has recently reanalysed his data using one of the several modifications to principal component regression which have been suggested in the past few years. In this technique, due to Hawkins (1973), the principal component analysis is done on all the variables, including the dependent variable, and the principal components of most interest are those with small variances. In introducing this technique, Hawkins (1973) also pointed out that in the usual form of principal component regression there is no guarantee that the low-variance components will be unimportant.

It is not too difficult to find examples in the literature where some of the last few (low variance) components *are* important. As a first example, Smith and Campbell (1980) give an example from chemical engineering. There are nine regressor variables and when principal component regression is used the principal components which are important are (in order of decreasing variance) the first, third, fourth, sixth, seventh and eighth. The eighth accounts for only 0·06 per cent of the variance and would be rejected on any "low variance" criterion.

A second example is provided by Kung and Sharif (1980). In a study of the prediction of monsoon onset date from ten meteorological variables it is found that the most important principal components in a regression equation are the eighth, second and tenth, in that order, i.e. the principal component with the smallest variance, accounting for less that 1 per cent of the variability in the original regressor variables, is the third most important variable in the regression equation.

A third example, this time from economics, is given by Hill *et al.* (1977). In this six-variable example the fourth and fifth components should both be included in the regression despite contributing only 0·25 per cent, 0·04 per cent of the variation in the original variables.

The fourth and final example is a simple one, drawn from meteorology. Suppose that it is required to predict the height of the cloud-base, H, an important problem at airports. Various climatic variables are measured including surface temperature T_s, and surface dewpoint, T_d. Here, T_d is the temperature at which the surface air would be saturated with water vapour, and the difference $T_s - T_d$ is a measure of surface humidity. Now T_s, T_d are generally positively correlated, so a principal component analysis of the climatic variables will have a high-variance component which is highly correlated with $T_s + T_d$, and a low-variance component which is similarly correlated with $T_s - T_d$. But H is related to humidity and hence to $T_s - T_d$, i.e. to a low-variance rather than a high-variance component, so a strategy which rejects low-variance components will give poor predictions for H.

The discussion of this example is necessarily vague because of the unknown effects of any other climatic variables which are also measured and included in the analysis. However, it shows a physically plausible case where a dependent variable will be related to a low-variance component, confirming the three empirical examples from the literature.

Furthermore, the cloud-base example has been tested on data from Cardiff (Wales) Airport for the period 1966–73 with one extra climatic variable, sea-surface temperature, also included. Results were essentially as predicted above. The last principal component was approximately $T_s - T_d$, and it accounted for only 0·4 per cent of the total variation. However, in a principal component regression it was easily the most important predictor for H.

The above examples have shown that it is not necessary to find obscure or bizarre data in order for the last few principal components to be important in principal component regression. Rather it seems that such examples may be rather common in practice. Hill *et al.* (1977) give a thorough and useful discussion of strategies for selecting principal components which should have buried forever the idea of selection based solely on size of variance. Unfortunately this does not seem to have happened, and the idea is perhaps more widespread now than 20 years ago.

Acknowledgements

I am very grateful to Roy Maryon for the data on cloud base heights. Part of this research was done while the author was a Visiting Associate Professor at the University of Guelph, Ontario, Canada.

References

Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and its Application: A Data-oriented Approach.* New York: Marcel Dekker.

Hawkins, D. M. (1973). On the investigation of alternative regressions by principal component analysis. *Appl. Statist.*, **22**, 275–286.

Hill, R. C., Fomby, T. B. and Johnson, S. R. (1977). Component selection norms for principal component regression. *Commun. Statist.–Theor. Method*, **A6**, 309–334.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *Brit. J. Stat. Psychol.*, **10**, 69–79.

Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Appl. Statist.*, **16**, 225–236.

—— (1981). Investigation of alternative regressions: some practical examples. *The Statistician*, **30**, 79–88.

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

KUNG, E. C. and SHARIF, T. A. (1980). Multi-regression forecasting of the Indian summer monsoon with antecedent pattern of the large scale circulation. In *WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, pp. 295–302.

MANSFIELD, E. R., WEBSTER, J. T. and GUNST, R. F. (1977). An analytic variable selection technique for principal component regression. *Appl. Statist.*, **26**, 34–40.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis.* London: Academic Press.

MASSY, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Ass.*, **60**, 234–256.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics.* Reading, Mass.: Addison-Wesley.

SMITH, G. and CAMPBELL, F. (1980). A critique of some ridge regression methods. *J. Amer. Statist. Ass.*, **75**, 74–103 (including discussion).