ORiON, Vol. 19, No. 1/2, pp.75-85 *ISSN 0259-191-X*

AN EXPERIMENT ON SELECTING MOST INFORMATIVE VARIABLES IN SOCIO-ECONOMIC DATA

LARRY JENKINS
Department of Business Administration, Royal Military College of Canada, Kingston, Ontario K7K 7B4

ABSTRACT

In many studies where data are collected on several variables, there is a motivation to find if fewer variables would provide almost as much information. Variance of a variable about its mean is the common statistical measure of information content, and that is used here. We are interested whether the variability in one variable is sufficiently correlated with that in one or more of the other variables that the first variable is redundant. We wish to find one or more 'principal variables' that sufficiently reflect the information content in all the original variables.

The paper explains the method of principal variables and reports experiments using the technique to see if just a few variables are sufficient to reflect the information in 11 socio-economic variables on 130 countries from a World Bank (WB) database. While the method of principal variables is highly successful in a statistical sense, the WB data varies greatly from year to year, demonstrating that fewer variables would be inadequate for this data.

1. INTRODUCTION

In multivariate regression, methods for selecting the best sub-set of independent variables on which to regress the dependent variable are well known as an extension of analysis of variance. Software packages such as SPSS [1999] make the exercise encouragingly simple. On the other hand, when there is no differentiation of the variables into dependent and independent (for example, in multi-criteria decision analysis [Belton and Stewart 2001] or data envelopment analysis [Cooper et al., 2000]) methods for selecting a sub-set of variables that contain most of the 'information' (as measured by variance) are not well known.

An approach proposed in recent texts on multivariate analysis, for example Tabachnik and Fidell [2001] or Johnson and Wichern [2002], is to take the most heavily weighted variable in each of the first few principal components (PCs) as being the 'most informative'. Unfortunately, the amount of variance explained by just the most heavily weighted variable of a PC is obviously not the same as that explained by the whole PC, and PC analysis gives no measure of the variance explained by each variable.

In two recent publication Jenkins and Anderson [2003a, 2003b] described a method for choosing a sub-set of 'most informative' variables from a set of correlated data. They had rediscovered a method developed and promulgated by McCabe [1984, 2002] under the name 'principal variables' (PVs), which in turn owed much to earlier work by Beale et al. [1967]. In spite the usefulness of the method of PVs, even recent textbooks on multivariate analysis do not mention the method, and continue to suggest using the first variable of the first few PCs.

This brief article summarizes the method of PVs and PCs and compares results with some socio-economic data using PVs versus using the first variable in the PCs. The purpose of the experiments was not only to describe and compare results with the two methods, but also to find whether the same variables were consistently 'most informative' over several years of country socio-economic data available from the World Bank. The next section summarizes the method of principal variables, followed by a section on principal components. (These two sections repeat some material from the articles of Jenkins and Anderson [2002, 2003] for completeness). Section 4 defines the data used in the experiments, and Section 5 has the various results of the analyses.

2. PRINCIPAL VARIABLES

McCabe [1984] and Jenkins and Anderson [2003] describe the method of PVs in terms of conditional variance. The approach taken by Beale et al. [1967] was to maximize the minimum multiple correlation between the selected variables and each omitted variable, which is slightly different from the method of PVs described here.

Consider a set of variables i = 1 .. m that are observed on a number of cases j = 1 .. n. The observations may be represented by matrix $\mathbf{X}$ $|x_{ij}$, i = 1 ..m, j = 1 .. n| or column vector variables $\mathbf{X}_i$, i = 1 .. m. Working simply from the value of $\mathbf{X}$, we are interested in inferring whether one or more of the variables i = 1 .. m are so closely correlated to the others that using only a subset and ignoring the rest of the variables would result in little or no loss of information. For convenience of notation, we allow an arbitrary reordering of the variables, and will write about omitting variables i = 1 .. p, and retaining variables i = p+1 .. m.

Since the measurement scale of variable i does not enter into our consideration, each variable can be conveniently normalized to have mean 0 and variance of 1. To simplify notation, we will assume that this transformation has been carried out on all the data, and from now on will use x_{ij} to denote the normalized variables. With all variables initially standardized to a variance of 1, each has the same information content. Then the variance of the i = 1 .. m variables sums to numeric value m, and this total variance can be used as a measure of the information content in any subset of the m variables.

We use the approach of conditioning the observed value of one variable on the observed value of another. That is, a value $x_{i'j}$ observed jointly with a value for x_{ij} is adjusted to a value calculated as if x_{ij} were at the mean value of $\mathbf{X}_i$. If such an adjustment for every $x_{i'j}$, j = 1 .. n, adjusts $x_{i'j}$ to the mean value of $\mathbf{X}_{i'}$, then the partial variance of $\mathbf{X}_{i'}$ conditioned on $\mathbf{X}_i$ is zero. Since by this process all the information (variance) of $\mathbf{X}_{i'}$ is removed, it means that all the information is already contained in $\mathbf{X}_i$, so variable $\mathbf{X}_{i'}$ is redundant. Conditioning on two or more variables is a simple extension of the process. Thus if, in an arbitrarily ordered set of variables $\mathbf{X}_1$ .. $\mathbf{X}_m$, conditioning on $\mathbf{X}_{m-1}$ and $\mathbf{X}_m$ leaves zero partial variance in $\mathbf{X}_1$ to $\mathbf{X}_{m-2}$, then the information contained in all the variables $\mathbf{X}_1$ .. $\mathbf{X}_m$ is contained in $\mathbf{X}_{m-1}$ and $\mathbf{X}_m$. Since perfect correlation is unlikely in any real data, a residual partial variance that is small, rather than 0, is an acceptable goal

Now represent the variance-covariance matrix derived from the m columns of data matrix $\mathbf{X}$ as

$$\text{var}\,(\mathbf{X}) = \mathbf{V} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{pmatrix}$$

Consider partitioning the m variables into two sets, with appropriate relabelling as necessary, so that i = p+1 .. m are the variables retained as representing most of the information of all m variables, and i = 1 .. p variables are to be omitted. The variance-covariance matrix $\mathbf{V}$ can be partitioned as

$$\mathbf{V} = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

where $\mathbf{V_{11}}$ represents the variance-covariance matrix of variables i = p+1 .. m and $\mathbf{V_{22}}$ the variance-covariance matrix of variables i = 1 .. p. Then the partial variance-covariance matrix of (relabelled) $\mathbf{X}_1$, $\mathbf{X}_2$, .. $\mathbf{X}_p$ given $\mathbf{X}_{p+1}$ .. $\mathbf{X}_m$ is $\mathbf{V_{11.2} = V_{11} - V_{12}\, V^{-1}_{22}\, V_{21}}$ (Morrison, 1976 p.92). The trace of $\mathbf{V_{11.2}}$ represents the remaining variance of variables i = 1 .. p after conditioning on the selected variables i = p+1 .. m. If the trace of $\mathbf{V_{11.2}}$ is small, then variables i = p+1 .. m retain sufficient of the information (measured by variance) to represent all the original variables i = 1 .. m.

Once the variance-covariance matrix **V** has been computed initially for **X**, all subsequent calculations are manipulations on **V**. If, as described here, all variables have been normalized to mean 0 and variance 1, there is no difference between the initial variance-covariance matrix **V** and a standard Pearson correlation matrix, which therefore forms a convenient starting point for our procedure.

We have experimented with two approaches, the simpler one we call a 'm*yopic*' or 'greedy' procedure, while the second is comprehensive. In the myopic procedure, we start by taking each of the m variables as the conditioning variable, and find which one has the maximum information content (as computed by the trace of $\mathbf{V_{11.2}}$). With this first variable now selected, we try all the remaining m-1 variables to find which best represents the residual information content. This can be continued until there is only one residual variable, and all the other m-1 are conditioning variables.

While the *Myopic* procedure is computationally simple, unfortunately if conditioning is to be on 2 variables, we cannot be sure that the first variable selected will be one of the best two variables, that the best 2 will be a subset of the best 3, and so on [Jenkins and Anderson, 2000]. The alternative is to try conditioning on all combinations of 2 variables, all combinations of 3 variables etc. When the total number of variables m is small, it will be worthwhile trying all ${}_mC_{m-p}$ combinations for all values of p, to find which m-p variables best represent all the data – a procedure that we code under the name '*PickBest'*.

3. PRINCIPAL COMPONENTS

In contrast with the approach of omitting variables that do not bring significantly more information than is already contained in our selected p+1 .. m variables, PC analysis retains all the original variables and forms artificial variables that are linear combinations of them. The first PC is a weighted sum of all the input variables, calculated so that as much as possible of the variance of all the raw variables is contained in that component. With variables $\mathbf{X}_i$ i = 1 .. m, then

$$\mathbf{PC}_{(1)} = w_{(1)1}\mathbf{X}_1 + w_{(1)2}\mathbf{X}_2 + \dots w_{(1)i}\mathbf{X}_i + \dots w_{(1)m}\mathbf{X}_m$$

where the weights $w_{(1)1}$, $w_{(1)2}$, .. $w_{(1)i}$, … $w_{(1)m}$ have been chosen to maximize the variance of $\mathbf{PC}_{(1)}$ - the first principal component - subject to the constraint that $\Sigma_{i=1}^{m} w^2_{(1)i} = 1$ (Dillon and Goldstein, 1984).

After this first component is extracted, the raw variables have some residual variance. Then a second principal component, $\mathbf{PC}_{(2)}$, is extracted to include as much as possible of this residual variance. To extract the maximum remaining variance, this second PC will be orthogonal to the first. Technically, the PCs of matrix **X** are the eigenvectors of the sample covariance matrix, while the eigenvalue corresponding to each eigenvector is the amount of variance explained by that eigenvector. Thus many common matrix manipulation programs can be used to calculate the principal components.

4. DATA USED IN THE EXPERIMENTS

Country annual data were downloaded from the World Bank (WB) website [2002] for the years 1997-2001. From the WB variables: *Surface area (sq km); Population, total; and GDP (current US$)* were derived:

Density – Population per sq. km.
LogPop – Log_{10} (total population)
LogGDP/Cap – Log_{10}(GDP/population)

Other variables downloaded and used directly in the analysis were:

Popgrow - *Population growth (annual %)*
Urbanpop - *Urban population (% of total)*
Agric - *Agriculture, value added (% of GDP)*
Gdpgrow - *GDP growth (annual %)*

Fertility	- *Fertility rate, total (births per woman)*
Lifexp	- *Life expectancy at birth, total (years)*
Mortinf	- *Mortality rate, infant (per 1,000 live births)*
IllitF	- *Illiteracy rate, adult female (% of females ages 15 and above)*

From the original WB list of 225, countries with populations of less than 1 million people (commonly islands dependent on tourist trade and offshore banking) were deleted, as well as a few other countries that had no data for GDP, but there was no filtering on factors such as geographic region. This left a sample of 130 countries for which data was available in most of the years, though the website had no data for Fertility, Lifexp and Mortinf in 2001.

5. EXPERIMENTAL RESULTS

The goal of the experiments was two-fold. The first purpose was a straightforward evaluation of the relative success of different statistical methods of selecting ‘most informative variables’. The secondary purpose was based on the hope that just a few variables in this particular socio-economic data would be strong indicators for the rest of the data. An indication of this would be if just two or three variables explained more than 50% of the variance of all eleven variables, and furthermore if the same variables succeeded in doing this for the five years for which data were examined.

A further experiment was to select a few variables that might be the easiest to measure in practice, and find how much of the variance of all eleven variables was explained by the selected variables. Again one would hope that the selected variables explained at least 50% of the variance of all eleven variables, and that the percentage explained would be approximately the same in all five years.

All computation was performed in Microsoft Excel 2000 spreadsheet software, with routines coded in Visual Basic for *Myopic* and *PickBest* [Anderson and Jenkins, 2002]. Murray Anderson also modified these routines to create a routine *PickSubset,* used to measure the variance explained by one or more variables specified by the user. The Eigenanalysis routine in the Excel add-in PopTools [Hood, 2002] was used to compute the PCs. Excel was used to calculate the Pearson correlation matrix from the data for each year, then this became all that was required as input to the subsequent analyses.

Table 1 lists the 1997 PVs selected by the *Myopic* routine and the proportion of variance explained by the sequentially selected variables. The third column gives the proportion of variance explained by the very best combination of variables. As highlighted in the first four rows, *Myopic* succeeds in choosing the very best one, two, three and four variables, and even for more than four variables, the *Myopic* selection is almost as good as the very best combination.

The right-hand three columns give results with a PC analysis. Obviously the first variable of each PC explains (column 5) less variance than the whole PC (column 6). Surprising here is how successful the first variable of the PCs is compared with PVs. The highlighting shows that the first variable of the first PC is the most informative of all the variables when taken on its own. Oddly the best combination of two and three variables is not found by the PC method, but coincides again on four selected variables. (The footnote to Table 1 notes that some variables repeat as the most heavily weighted variable in a PC. The heuristic for identifying "most informative" variables via PC analysis then selects the most heavily weighted variable that has not been selected already).

Table 1. PV and PC results for 1997 data

Myopically chosen variable	**Cumulative variance by Myopic**	**Cumulative variance by PickBest**	**First variable in principal component**	**Cumulative variance by 1st variable of PC**	**Cumulative variance by PCs**
Mortinf	48.65%	48.65%	Mortinf	48.65%	53.52%
Gdpgrow	58.08%	58.08%	Density	58.05%	64.76%
LogPop	67.34%	67.34%	LogPop	67.32%	73.75%
Density	76.56%	76.56%	Gdpgrow	76.56%	82.25%
Agric	83.42%	83.79%	Popgrow	82.57%	88.65%
Popgrow	89.43%	89.43%	IllitF[1]	86.46%	92.78%
IllitF	92.90%	93.32%	Urban	92.40%	95.66%
Urban	96.35%	96.52%	LogGDP/Cap	95.61%	97.65%
LogGDP/Cap	98.16%	98.17%	Fertility	96.86%	98.88%
Fertility	99.25%	99.45%	Lifexp	97.67%	99.64%
Lifexp	100.00%	100.00%	Agric[2]	100.00%	100.00%

[1] Popgrow was more heavily weighted than IllitF in the 6th PC.

[2] Mortinf, Lifexp, Fertility and LogGDP/Cap were all weighted more heavily than Agric in the 11th PC.

Table 2. PV and PC results for 1997-2001 data

1998

Myopically chosen variable	Cumulative variance by Myopic	Cumulative variance by PickBest	First variable in principal component	Cumulative variance by 1st variable of PC	Cumulative variance by PCs
LogGDP/Cap	31.39%	31.39%	LogGDP/Cap	31.39%	38.17%
Mortinf	42.10%	42.10%	Mortinf	42.10%	50.12%
Gdpgrow	51.52%	51.52%	Gdpgrow	51.52%	60.78%

1999

Myopically chosen variable	Cumulative variance by Myopic	Cumulative variance by PickBest	First variable in principal component	Cumulative variance by 1st variable of PC	Cumulative variance by PCs
LogGDP/Cap	28.13%	28.13%	LogGDP/Cap	28.13%	34.37%
Mortinf	39.75%	39.75%	Mortinf	39.75%	48.88%
Gdpgrow-	49.44%	49.44%	LogPop	48.61%	59.50%

2000

Myopically chosen variable	Cumulative variance by Myopic	Cumulative variance by PickBest	First variable in principal component	Cumulative variance by 1st variable of PC	Cumulative variance by PCs
Mortinf	49.52%	49.52%	Mortinf	49.52%	54.77%
Gdpgrow	58.98%	58.98%	Gdpgrow	58.98%	65.55%
LogPop	68.17%	68.20%	Density	68.06%	75.42%

2001

Myopically chosen variable	Cumulative variance by Myopic	Cumulative variance by PickBest	First variable in principal component	Cumulative variance by 1st variable of PC	Cumulative variance by PCs
IllitF	30.14%	30.14%	IllitF	30.14%	38.96%
Density	44.32%	44.32%	Density	44.32%	55.65%
Gdpgrow	57.72%	57.72%	LogGDP/Cap	54.80%	68.50%

Table 2 follows the same style as Table 1, but just the first few variables selected by *Myopic* are shown. For these five years' worth of data, a number of conclusions are obvious. Firstly, the myopic way of selecting PVs performs almost as well as a comprehensive (*PickBest)* method. Obviously this is simply an experimental result, though our trials with other data [Jenkins and Anderson, 2000, 2003a, 2003b] gave similar results. Secondly with the selection of WB data used, the heuristic of using the first variable of each PC performs reasonably well for much of this data. However, selecting the third variable in 2001 using the PC is a poor result for just eight variables and a sample as large as 130 cases.

For these particular data, the hope that the same few variables each year would be strong indicators for the rest of the data was obviously unsatisfied. In fact, the large variation between years was a surprise. (Just looking at the values in the different correlation matrices indicated this, though in the interests of rigour we also performed statistical tests for significant differences in correlation across the years).

The last experiment was to choose a few variables of the eleven (eight in 2001) that might be the easiest to observe, and find if they explained a large portion of the total variance. These could then be considered as indicators for the more comprehensive socio-economic data. Table 3 shows the proportion of variance explained as the three chosen variables are introduced incrementally. The results make it obvious that the gross change from year to year makes this approach unsatisfactory for these particular data sets.

Table 3. Proportion of total variance explained with selected variables

	1997	**1998**	**1999**	**2000**	**2001**
LogGDP/Cap	42.78%	31.39%	28.13%	43.06%	25.52%
Popgrow	54.04%	41.98%	38.49%	55.89%	38.07%
Gdpgrow	63.09%	51.46%	48.52%	65.08%	51.91%

6. CONCLUSION

Selecting a sub-set of PVs from a set of correlated data so that the selected variables explain most of the variance (information content) of all of them is straightforward. We have coded the procedure as an add-in to Excel spreadsheet software, and it is available as a Fortran routine from G. P. McCabe, the originator of PVs. Even our simplistic *Myopic* routine gives good results, better than the common heuristic of taking the first variable of the first few PCs as ‘most informative’ variables.

We had hoped to find that just a few variables would fairly represent all 11 common socio-economic variables taken from World Bank data on 130 countries. This proved unsuccessful, owing primarily to the surprisingly variability in correlations of these 11 variables over the five years of data with which we experimented.

REFERENCES

Anderson, M. and Jenkins, L., Excel visual-basic add-ins *Myopic* and *PickBest,* http://www.rmc.ca/academic/busadm/staff/jenkins-e.html (January 2002).

Beale, E. M. L., Kendall, M. G. and Mann, D. W., The Discarding of Variables in Multivariate Analysis, *Biometrika,* **54** (1967) 357-366.

Belton, V. and Stewart, T. J., *Multiple Criteria Decision Analysis: An Integrated Approach,* Kluwer Academic, Dordrecht, The Netherlands, (October 2001).

Cooper, W. W., Seiford, L. M. and Tone, K., *Data Envelopment Analysis*. Kluwer Academic Publishers, Norwell, MA. (2000).

Dillon W. R., and Goldstein, M, *Multivariate Analysis,* Wiley, New York (1984).

Hood, G., *PopTools 2.5,* available at: http://www.cse.csiro.au/poptools/ (October 2002).

Jenkins, L., and Anderson, M., A statistical approach to eliminating some of the criteria in a multi-criteria decision analysis with little loss of information (Presented at INFORMS Annual Conference, San Antonio, TX, November 2000).

Jenkins, L. and Anderson, M., Omitting Correlated Variables, *ORiON*, 18 (2002) 21-36.

Jenkins, L. and Anderson, M., A multivariate statistical approach to reducing the number of variables in data envelopment analysis, *European Journal of Operational Research*, **147** (2003) 51-61.

Johnson, R. A. and Wichern, D.W., *Applied Multivariate Statistical Analysis,* 5th ed., Prentice Hall, Upper Saddle River, NJ (2002).

McCabe, G. P., Principal variables, *Technometrics*, **26** (1984) 137-144.

McCabe, G. P., http://www.stat.purdue.edu/people/mccabe (accessed May 2002).

Morrison, D. F., *Multivariate Statistical Methods*, 2nd ed. (McGraw-Hill, New York, 1976).

SPSS Inc. *SPSS Base 10.0 Applications Guide* (SPSS Inc. Chicago, IL, 1999).

Tabachnik, B.G. and Fidell, L. S., *Using Multivariate Statistics,* 4th ed., Allyn & Bacon, Boston, MA (2001).

World Bank website http://devdata.worldbank.org/data-query (accessed 9 October 2002).