ORiON, Vol. 18, No. 1/2, pp. 21-36 *ISSN 0259-191-X*

OMITTING CORRELATED VARIABLES

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

ABSTRACT
Data collected on the physical, biological or man-made world are often highly correlated, posing the question of whether fewer variables would contain almost as much information. A crude solution is simply to look at the Pearson correlation matrix and omit one of a pair of highly correlated variables. A more systematic method is to condition on one or more variables, and observe the resulting partial covariance matrix. If the variables have little variance after the conditioning, then the conditioning variables contain most of the information of all the original variables. Paralleling the usual tests applied in judging how many principal components are sufficient to represent all the data, we can use the amount of variance explained by the conditioning variable(s), as a measure of information content. The paper references earlier work in this area, explains the computation and includes examples using published data sets. The approach is found to be highly competitive with using principal components, and has the obvious advantage over principal components of simply omitting some of the original variables from further consideration. The method has been coded in Visual-Basic add-ins to an Excel spreadsheet.

Keywords: Multivariate statistics; Correlation; Data reduction

1. INTRODUCTION

In studying physical and social phenomena, it often happens that two observed variables are highly correlated with one another. This immediately opens up the question of whether there is any need to observe the values of both variables, or is it sufficient to collect data on just one or other. Indeed, often several variables are observed and found to be correlated, and it is useful to know whether collecting data on a smaller number would be sufficient.

When we are interested in a single dependent variable, and all other variables are examined as predictors of the dependent variable, well-known statistical techniques such as analysis of variance or step-wise multiple regression can be applied (Neter, et al., 1996). On the other hand, if we seek a general understanding of the data without an immediate differentiation into dependent and independent variables, the applicable statistical techniques are less well known. A common approach is to use the multivariate method of principal components, or the extension of this into factor analysis. Unfortunately this technique does not address directly the basic question of whether all the original variables yield much more information than just some sub-set of them.

In this paper we present a statistical method that measures the amount of information lost by omitting one or more variables from a set of correlated observations, and thereby identifies which variables are best retained. This is primarily an ex-post analysis when we are simply interested in reducing the total number of variables to allow the underlying phenomena to be understood more easily. However, the method can also be used ex-ante on a preliminary sample of observations to assess the loss in information if data on all variables is not collected for the main analysis.

Although the technique is known to experienced statisticians, it is not well known in the operational research community. (We reinvented it after failing to find to find any references in the literature, including recent texts on multivariate statistics). This article describes the method as we saw it, and refers the reader to our routines, coded in Microsoft Visual Basic, that can be used in an Excel spreadsheet.

The next section summarizes previous literature in this area, and Section 3 introduces notations and definitions. Section 4 explains our method that uses partial covariance to identify which variables are most significant. Section 5 explains the computation, and Section 6 describes tests that can be applied to decide how many variables sufficiently represent the information in all the data. Section 7 summarizes the contrasting approach of principal components analysis. Sample results are given in Section 8.

2. PREVIOUS RESEARCH

The earliest major article on this subject dates back to 1967 [Beale, et al.] and has often been referenced in the context of multivariate regression, where one variable is treated as dependent on the other variables, and the exercise becomes one of choosing just a sub-set of the independent variables as sufficient to explain most, or all, the variance in the dependent variable. In fact, the article by Beale et al., which gives a limpid expression and solution of the problem in multivariate regression, also points out how their method can be modified easily to the situation where none of the variables is considered as dependent on the others. Then the problem, which they describe as ‘Interdependence Analysis’, becomes precisely the problem considered in this paper.

If one pursues the thread of articles on selecting a subset of dependent variables in multivariate regression, one can trace back to at least 1960 [Efroymson] and forward through, for example, Garside [1965] Hocking and Leslie [1967] and a commentary by Beale in 1970, followed by numerous more recent articles. On the other hand, the thread on ‘interdependence analysis’ is very thin. Jolliffe’s interest in principal component analysis led him to explore the idea of forming principal components with a sub-set of variables from a larger set of correlated variables [1972, 1973]. The simplest version of this is simply to use the first variable in each of the first few principal components, and omit all other variables from subsequent consideration. The general problem of forming principal components from only a sub-set of the original variables continues to be explored. A recent web site [Mori et al. 2000, 2002] gives references and a good summary of those methods.

It is not until McCabe’s article [1984] that the idea of selecting a subset of variables from a larger correlated set, without any consideration of a dependent variable, is addressed as the central goal. McCabe uses the name “principal variables”. The Fortran program that he wrote to test the method benefits from ideas used in the 1967 program of Beale et al., McCabe having come to it via his earlier work in discriminant analysis [McCabe 1975]. Recently he has made available on his web-site downloadable Fortran routines callable in the statistical software packages IMSL and SAS [McCabe, 1998] to select principal variables from a set of correlated variables.

Jolliffe’s text on principal components [1986, pp.108-113] compares the first variable of the principal components with the variables selected by McCabe’s method of principal variables, though the emphasis is still toward choosing a sub-set of variables from which to form the principal components. The results on two real data-sets using a number of different statistical criteria for “best” are inconclusive. The first few variables selected by either approach are the same, but there is divergence if more variables are to be included in the chosen sub-set.

A few years ago a real-world multi-criteria analysis by Jenkins [2001] elicited a number of criteria (variables) with which participants in the workshop were very comfortable, but were obviously highly correlated. While not a problem in that particular analysis, it posed the general academic question of how to deal with a number of highly correlated criteria (variables). A literature search (since found to be inadequate!) found references only to choosing a subset of the independent variables in multiple regression, so we invented a method [Jenkins and Anderson, 2000] that we later found to be a simple recasting of McCabe’s approach of principal variables. The method has proven interesting and useful [Jenkins and Anderson, 2002, 2003] and, since we can offer Visual Basic routines that can be used in Microsoft’s Excel spreadsheet software, we describe here our development and coding of this very useful technique.

3. NOTATION AND PREPARATION

Consider a number of variables i = 1 .. m that are observed on a number of cases j = 1 .. n, with datum x_{ij} assumed to have an unique scalar value. The observations may be represented by matrix $\mathbf{X}$ $|x_{ij}$, i = 1 ..m, j = 1 .. n$|$ or row 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, simply by subtracting the mean of the observations from each observed value. Similarly we can normalize the variance and standard deviation to 1 by dividing each x_{ij} by the standard deviation of $\mathbf{X}_i$. That is, $x'_{ij} = (x_{ij} - \mu_i) / \sigma_i$ where μ_i and σ_i represent the mean and standard deviation respectively of the observed $\mathbf{X}_i$. 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 now standardized, there is essentially no distinction between them. The only special case that could arise would be if all elements of some original $\mathbf{X}_i$ had the same value. In this case the variance would be 0, and our normalization would be mathematically undefined. But if some vector $\mathbf{X}_i$ has all elements identical, then it contains no information. Then, after omitting any such uninformative variables, the variance can be used as a measure of information. 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.

This total variance as a measure of information content is exploited in the procedure below. 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.

4. SELECTING VARIABLES BASED ON PARTIAL COVARIANCE

We now outline the procedure used to select variables i = p+1 .. m to retain as representing most of the information in all m variables, and identify those variables i = 1 .. p that contain little additional information. It should be emphasized that if two variables $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ are perfectly correlated, then whichever we choose first makes the second redundant. In other words, it is not intended to say that some variables are inherently more informative than others. We are simply trying to identify some sub-set i = p+1 .. m of a set of inter-related variables that contains as much information as possible. Some completely different sub-set of m-p variables may contain almost as much information.

The partial variance of a variable, denoted $\sigma_{ii.i''}$ is the variance remaining in variable i when the effect of variable i" is removed. This is equivalent to calculating the value that each x_{ij} would have if $x_{i''j}$ were at the mean value of $\mathbf{X}_{i''}$. If variable i is perfectly correlated with i", then conditioning variable i on i" will leave $\sigma_{ii.i''}$ with a value of 0. (Similarly, $\sigma_{i''i''.i}$ will be zero.) 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 $\mathbf{X}_m$.

If all m variables are normalized to have unit variance, then the sum of their variances is simply m. Thus if i = p+1 .. m are the variables retained as representing most of the information of all m variables, and i = 1 .. p are omitted, ideally the variance explained by variables p+1 .. m will be m, and the partial variance of variables 1 .. p will be zero. Since perfect correlation is unlikely in any real data, a residual partial variance that is small, rather than 0, is an acceptable goal. We use the proportion of variance explained by the selected p+1 .. m variables, or alternately the residual partial variance in the remaining p variables, to decide how many variables reasonably represent all the information.

5. COMPUTATION

The covariance between two random variables $\mathbf{X}_i$, $\mathbf{X}_{i'}$ is given by the joint moment

$$\mathrm{Cov}\,(\mathbf{X}_i , \mathbf{X}_{i'}) = \mathrm{E}\{[\mathbf{X}_i - \mathrm{E}(\mathbf{X}_i)]\ [\mathbf{X}_{i'} - \mathrm{E}(\mathbf{X}_{i'})]\} = \mathrm{E}\,(\mathbf{X}_i\ \mathbf{X}_{i'}) - \mathrm{E}\,(\mathbf{X}_i)\ \mathrm{E}\,(\mathbf{X}_{i'}) = \sigma_{ii'}$$

Now represent the variance-covariance matrix derived from the m rows of data matrix **X** as

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

Consider partitioning the m variables into two sets, with appropriate relabelling a snecessary, 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 **V** can be partitioned as

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

where $\boldsymbol{V}_{11}$ represents the variance-covariance matrix of variables i = 1 .. p, $\boldsymbol{V}_{22}$ the variance-covariance matrix of variables i = p+1 .. m and $\boldsymbol{V}_{12}$ matrix of covariances between these two sets of variables. Then the partial variance-covariance matrix of $\mathbf{X}_1$, $\mathbf{X}_2$, .. $\mathbf{X}_p$ given $\mathbf{X}_{p+1}$ .. $\mathbf{X}_m$ is $\boldsymbol{V}_{\mathbf{11.2}} = \boldsymbol{V}_{\mathbf{11}} - \boldsymbol{V}_{\mathbf{12}}\ \boldsymbol{V}^{\mathbf{-1}}_{\mathbf{22}}\ \boldsymbol{V}_{\mathbf{21}}$ (Morrison, 1976 p.92). The trace of $\boldsymbol{V}_{\mathbf{11.2}}$ represents the remaining variance of variables i = 1 .. p after conditioning on the selected variables i = p+1 .. m. If the trace of $\boldsymbol{V}_{\mathbf{11.2}}$ 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 $\boldsymbol{V}$ has been computed initially for **X**, all subsequent calculations are manipulations on $\boldsymbol{V}$. Since, in fact, we have an initial step of normalizing all variables to mean 0 and variance 1, there is no difference between the initial variance-covariance matrix $\boldsymbol{V}$ and a standard Pearson correlation matrix, which therefore forms a convenient starting point for our procedure.

Having discussed how to compute the partial covariance matrix and how to monitor the amount of variance explained by the selected conditioning variables, we still need to consider how the conditioning variables will be selected. (It is worth pointing out that the procedure can be controlled by the analyst's knowledge of what the variables represent, and can serve as a diagnostic of how much information is retained or lost by including or omitting specific variables. However, here we are concerned only with a mechanistic procedure having the goal of retaining most information with the least number of variables).

We have experimented with two approaches, the simpler one referred to as a "myopic" 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 $\boldsymbol{V}_{\mathbf{11.2}}$). 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 this 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.

Section 8 includes illustrative results using both the myopic and the comprehensive approach to selecting the subset of conditioning variables. Murray Anderson has coded each procedure as a Visual Basic add-in to an Excel spreadsheet (Roman, 1999) that uses a correlation matrix as initial input. The two procedures, both of which include in the output measures of the variance explained by the selected variables, can be downloaded from an appendix to this paper on Larry Jenkins' website (Anderson and Jenkins, 2002). The first add-in is named *Myopic*, and selects all m variables sequentially in a single run. The second add-in, named *PickBest*, asks the user how many m-p variables to use to explain the variance. Output consists of results with m-p explanatory variables on a new Excel worksheet. The output lists all possible m-p combinations of the original variables, in decreasing order of total variance explained.

6. HOW MANY VARIABLES SUFFICIENTLY REPRESENT ALL THE DATA?

We discuss in this section some heuristic and statistical rules to help decide how many variables might satisfactorily represent all the information contained in the original variables. Without any special knowledge of what each variable measures, our guidelines depend strictly on statistical tests. The simplest approach focuses on the proportion of total variance m contained in the subset of variables. The other approach considers the residual (partial) variance in the variables omitted. The heuristics and statistical rules are all drawn from studies in principal components and factor analysis (see Section 7), but are equally applicable to our analysis.

Rules based on the variance explained by each variable:

1. *Base the stopping rule on the percentage of total variance explained by m- p variables*
   Our procedure identifies, for each value of p, the sub-set of variables that explains as much as possible of the total variance m. Then we examine the percentage of total variance explained by m-p variables, and decide when the proportion is large enough that it is not worth retaining more than m-p variables. A graph plotting number of variables omitted against proportion of variance explained can easily suggest an appropriate

stopping point. This was named the *scree* test (Cattell, 1966, cited in Bernstein, 1988 p.174)

2. *Base the stopping rule on the percentage of remaining variance explained by a variable*
   In this case we include an additional variable as long as it explains a large proportion of the variance remaining. This is simply a variation on rule 1.

3. *A variable can be omitted if it explains less than 1 unit of the remaining variance*
   The logic of this rule is simply that if adding another variable cannot bring to the analysis at least as much variance as it explained as a stand-alone variable (1 unit), then its contribution to total explanation is inadequate. The heuristic is usually attributed to Kaiser (1958, cited in Dillon and Goldstein, 1984 p.48).

Rule based on the partial correlation matrix:
There are statistical tests of significance that can be applied to the partial correlation matrix remaining after conditioning on the variables p+1 .. m. They all test whether the matrix is significantly different from the identity matrix, for if there was 0 correlation remaining between the variables, the partial correlation matrix would have 1s on the principal diagonal, and 0s everywhere else. In other words, all common factors would have already been extracted by the conditioning variables p+1 … m. The tests of significance depend both on the total number of variables m, and the total sample size n. While m is implicit in the correlation matrix, n needs to be known independently as the size of the sample from which the original correlation matrix was calculated. Bartlett (1950, cited in Bernstein, 1988 p.175) developed a chi-square test of whether the partial correlation matrix is significantly different from the identity matrix as follows:
Calculate $\chi^2 = -[n - 1 - (2m + 5)/6] \ln|V|$, where n is the number of cases, m is the number of variables and $|V|$ is the determinant of the partial correlation matrix for the 1 .. p variables, conditioned on variables p+1 .. m. The degrees of freedom for this computed χ^2 is m(m-1)/2.

7. PRINCIPAL COMPONENTS ANALYSIS

In contrast with our approach of omitting variables that do not bring significantly more information than is already contained in our selected p+1 .. m variables, principal components (PC) analysis retains all the original variables and forms artificial variables that are linear combinations of them. We mention it here briefly since the method is well known and its results provide a valuable benchmark against which to compare our approach. PC analysis was conceived as a method of trying to identify a few underlying factors that reasonably explained most of the variability in a set of related observations.

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

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

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. The process can continue until m principal components are extracted, and these will always suffice to explain all the variance in the original data. However, since PC analysis maximizes the reduction in the variance at each step, often just the first few components contain most of the information in the original data. 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.

8. ILLUSTRATIVE RESULTS

To illustrate our approach, we start with a small artificial example. The data are listed in Table 1, with the correlation matrix in Table 2.

Table 1. Artificial data

Case j	X1	X2	X3
1	2.30	3.30	2.07
2	1.50	0.50	3.55
3	2.20	3.20	2.08
4	1.80	2.30	2.47
5	0.50	2.60	1.48
6	1.30	2.80	1.82
7	1.40	2.10	2.37
8	0.30	1.80	1.92

Table 2. Correlation matrix for artificial data

	X1	X2	X3
X1	1.000		
X2	0.366	1.000	
X3	0.326	-0.760	1.000

Though it is hardly obvious from the correlation matrix for this data, any two of the three variables contains all the information of the three variables. ($\mathbf{X}_3$ was calculated by $\mathbf{X}_3 = 0.6\mathbf{X}_1 - 0.7\ \mathbf{X}_2 + 3.0$). $\mathbf{X}_2$ alone can account for 57.07% of the variance of all three variables, $\mathbf{X}_3$ alone for 56.14%, and $\mathbf{X}_1$ alone for 41.35%. Any two will account for all the variance. By comparison, the first principal component can account for 58.72% of the variance of all three variables, (which is barely more than 57.07% with the most informative variable) and of course two principal components account for all the variance.

For our second example, we start simply with a Pearson correlation matrix. The matrix is taken from an example in the SPSS manual related to smoking (SPSS, 1999, p.318) but we simply label the variables 1 to 12.

Table 3. Correlation matrix of variables characterizing different smokers

	Var1	*Var2*	*Var3*	*Var4*	*Var5*	*Var6*	*Var7*	*Var8*	*Var9*	*Var10*	*Var11*	*Var12*
Var1	1.00											
Var2	0.85	1.00										
Var3	0.81	0.78	1.00									
Var4	0.82	0.81	0.79	1.00								
Var5	0.06	0.12	0.14	0.12	1.00							
Var6	0.11	0.16	0.19	0.22	0.80	1.00						
Var7	0.10	0.17	0.24	0.24	0.74	0.70	1.00					
Var8	0.12	0.21	0.22	0.30	0.71	0.73	0.71	1.00				
Var9	0.04	0.23	0.09	0.20	0.56	0.60	0.49	0.58	1.00			
Var10	0.13	0.28	0.14	0.21	0.36	0.34	0.24	0.27	0.46	1.00		
Var11	0.14	0.27	0.20	0.27	0.41	0.43	0.34	0.36	0.51	0.80	1.00	
Var12	0.04	0.20	0.10	0.22	0.58	0.61	0.61	0.59	0.80	0.61	0.70	1.00

Table 4. Comparison of most informative fewer variables with PC analysis for smokers

Variables selected	% total variance explained by best m-p variables	% total variance explained by same number of PCs
V12	33.59%	45.19%
V1, V12	59.14%	70.22%
V1, V5, V12	70.82%	81.75%
V1, V5, V10, V12	78.24%	86.33%
V1, V6, V7, V9, V10	83.09%	89.27%
V1, V5, V7, V8, V9, V11	86.50%	91.80%
V1, V3, V5, V7, V8, V9, V11	89.51%	93.82%
V1, V3, V5, V7, V8, V9, V10, V11	92.47%	95.48%
V1, V3, V5, V6, V7, V8, V9, V10, V11	94.86%	96.88%
V2, V3, V4, V5, V6, V7, V8, V9, V10, V11	97.01%	98.09%
V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12	98.56%	99.16%

From the results in Table 4, we see that the “myopic” approach would be valid for up to 4 variables, but would not select the best 5 variables, and that V1, which forms part of the most informative sub-set of 2 through 9 variables, is omitted if we use 10 or 11 variables.

As a third example we present results for a fairly large data set used to illustrate factor analysis in the recent release of the SPSS software for statistical analysis (SPSS, 1999, p.324), and consists of 11 population-related variables for 74 countries. Table 5 shows the initial correlation matrix.

Table 5. Correlation matrix of population-related data for 74 countries

	urban	**lifeexpf**	**literacy**	**pop_incr**	**babymort**	**birth_rt**	**death_rt**	**log_gdp**	**b_to_d**	**fertilty**	**log_pop**
urban	1.000										
lifeexpf	0.685	1.000									
literacy	0.526	0.867	1.000								
pop_incr	-0.204	-0.507	-0.642	1.000							
babymort	-0.667	-0.975	-0.855	0.509	1.000						
birth_rt	-0.473	-0.801	-0.824	0.837	0.810	1.000					
death_rt	-0.319	-0.470	-0.298	-0.303	0.463	0.076	1.000				
log_gdp	0.734	0.829	0.673	-0.499	-0.817	-0.725	-0.147	1.000			
b_to_d	-0.022	-0.186	-0.361	0.879	0.179	0.608	-0.598	-0.295	1.000		
fertilty	-0.387	-0.751	-0.823	0.831	0.754	0.967	0.103	-0.599	0.595	1.000	
log_pop	-0.315	-0.265	-0.185	-0.075	0.297	0.020	0.168	-0.288	-0.218	0.002	1.000

Table 6. Amount of variance in population-related data explained by conditioning versus PCA

Number of variables/components extracted	1	2	3	4	5	6	7	8	9	10	11
Cumulative % additional variance extracted by conditioning - myopic approach	52.1%	72.2%	79.9%	86.6%	92.2%	94.8%	96.9%	98.8%	99.4%	99.7%	100.0%
Cumulative % additional variance extracted by conditioning - most informative variables	52.1%	75.7%	84.0%	89.6%	93.1%	95.5%	97.8%	98.9%	99.5%	99.8%	100.0%
Cumulative % of total variance extracted by PCA	57.0%	79.5%	88.3%	93.9%	96.1%	97.6%	98.7%	99.3%	99.7%	99.9%	100.0%

This last analysis is offered only as an example of applying our technique, but in terms of just a few of the original variables representing most of the information the results are impressive. Selecting myopically, two out of the original 11 variables contain 72.2% of the information in all the data, and the best combination of two variables (log_gdp and literacy) explains 75.7% of the total variance, whereas the first two PCs contain 79.5%. With five variables chosen myopically the proportion of explained variance is 92.2%, and with the best five variables it is 93.1%, versus 96.1% with five PCs.

Since, with this population-related data, we know the sample size n = 74 countries, we can apply the Bartlett test for significance of the residual partial correlation matrix. We find that with conditioning on the best 9 variables (but not less) the residual partial correlation is not significant at the 5% level. This contradicts simpler heuristic rules such as the scree test, or ignoring variables (or PCs) that account for less than 1 unit of variance.

9. EXTENSIONS

The procedures described above have been coded to operate in a purely mechanical fashion to select the most informative variables. By treating the correlation matrix of the data as the starting point, "most informative" is implicitly evaluated on the basis of each variable having a variance normalized to 1. There are two very simple modifications of the mechanical process that may be useful when working with specific data:

1. The user selects specific variables of interest and uses the program to calculate the proportion of total variance that would be explained by just these selected variables;

2. The user decides that, for the specific study, some variables should be weighted as more important than others. This could be achieved by inputting the relative weighting at the beginning of the routine *Myopic* or *PickBest*, as data additional to the original correlation matrix. Thus if a study involves three variables, and the user decides that their relative importance is in the ratio 3:2:1, the information content selection step will weight the variables accordingly.

10. SUMMARY

We have described a method of multivariate statistics that selects a "most informative" subset of variables from a total set of observations where the variables are correlated with one another. The method was discovered some time ago, but is not at all well known in the operational research community. It is computationally simple to apply with modern software, and is much superior to attempting to guess which variables to retain simply by looking at the correlation matrix. Sample analyses with two moderately large sets of data (11 and 12 variables) illustrate that the approach is competitive with PC analysis, and has the obvious advantage of selecting fewer than all the original variables, while PC analysis uses all the original variables. A statistical test of significance on a data set with 74 cases suggests a caveat for both this method and PC analysis on how many variables (or PCs) can be omitted with little loss of information if the sample size is reasonably large.

REFERENCES

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

[2] M.S. BARTLETT, Tests of significance in factor analysis, *British Journal of Psychology,* **3**, 77-85 (1950).

[3] E.M.L. BEALE, Note on Procedures for Variable Selection in Multiple Regression, *Technometrics,* **12** 909-914 (1970).

[4] E.M.L. BEALE, M.G. KENDALL, and D.W. MANN, The Discarding of Variables in Multivariate Analysis, *Biometrika* **54**, 3/4 357-366 (1967).

[5] I.H. BERNSTEIN, *Applied Multivariate Analysis*, Springer-Verlag, New York, (1988).

[6] R.B. CATTELL, The scree test for the number of factors, *Multivariate Behavioral Research*, **1** 245-276 (1966).

[7] W.R. DILLON and M. GOLDSTEIN, *Multivariate Analysis,* Wiley, New York, (1984).

[8] M.A. EFROYMSON, Multiple Regression Analysis. In A. Ralston, H. S. Wilf (Eds.) *Mathematical Methods for Digital Computers*, Wiley, New York, (1960).

[9] M.J. GARSIDE, The Best Sub-set in Multiple Regression Analysis, *Applied Statistics*, **14,** 196-200 (1965).

[10] R.L. GORSUCH, *Factor Analysis*, 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ (1983).

[11] R.R. HOCKING and R.N. LESLIE, Selection of the Best Subset in Regression Analysis, *Technometrics*, **9,** 531-540 (1967).

[12] L. JENKINS, Prioritizing Clean-up of Contaminated Lightstation Sites. In: M. Koksalan, S. Zionts (Eds.), *Multiple Criteria Decision Making in the New Millenium,* Springer-Verlag, Berlin, 361-369 (2001).

[13] L. JENKINS and M. ANDERSON, 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).

[14] L. JENKINS and M. ANDERSON, A comparison of data envelopment analysis using fewer variables versus principal components, Working paper, Dept. of Business Administration, Royal Military College of Canada (October 2002).

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

[16] I.T. JOLLIFFE, Discarding variables in a principal component analysis. I. Artificial data, *Applied Statistics,* **21,** 160-173 (1972) .

[17] I.T. JOLLIFFE, Discarding variables in a principal component analysis. II. Real data. *Applied Statisics*, **22,** 21-31 (1973).

[18] I.T. JOLLIFFE, *Principal Components Analysis*, Springer-Verlag, New York, (1986).

[19] H.F. KAISER, The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23,** 187-200 (1958).

[20] G.P. McCABE, Principal variables, *Technometrics*, 26, 137-144 (1984)

[21] G.P. McCABE , http://www.stat.purdue.edu/people/mccabe (accessed March 2003).

[22] Y. MORI, M. IIZUKA, T. TARUMI, and Y. TANAKA, Statistical Software "VASPCA" for Variable Selection in Principal Component Analysis, In: *COMPSTAT2000 Proceedings in Computational Statistics* (*Short Communications*) (Edited by Jansen, W. and Bethlehem, J.G.) 73-74 (2000).

[23] Y. MORI, M. IIZUKA, T. TARUMI, and Y**.** TANAKA, *VASPCA/Web* –
Variable Selection in Principal Component Analysis,
http://mo161.soci.ous.ac.jp/vaspca/indexE.html (2002).

[24] D.F. MORRISON, *Multivariate Statistical Methods*, 2nd ed., McGraw-Hill, New York, (1976).

[25] J. NETER, M.H. KUTNER, C.J. NACHTSHEIM, and W. WASSERMAN, *Applied Linear Statistical Models*, 4th ed., Irwin, Chicago, (1996).

[26] S. ROMAN, *Writing Excel Macros*, O'Reilly & Associates, Sebastopol, CA, (1999).

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