NILU OR : 16/84
REFERANSE : E-8414
DATO : MAI 1984

TESTS OF HYPOTHESES IN THE PRINCIPAL COMPONENT ANALYSIS

Alena Moldanova

NILU

NORWEGIAN INSTITUTE FOR AIR RESEARCH

ROYAL NORWEGIAN COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH

NILU OR : 16/84
REFERANSE : E-8414
DATO : MAI 1984

TESTS OF HYPOTHESES IN THE PRINCIPAL COMPONENT ANALYSIS

Alena Moldanova

Present address: Departement of Probability and Mathematical Statistics
Charles University, Prague
Sokolovska 83
186 00 Praha 8
Czechoslovakia

NORSK INSTITUTT FOR LUFT FORSKNING
POSTBOKS 130, 2001 LILLESTRØM
NORGE

ISBN 82-7247-479-4

LIST OF CONTENTS

1 INTRODUCTION

Consider p variables measured on individuals of a given population, for example p chemical components measured in samples of air. The population in this case is the mass of air from which the sample was taken, and our aim is to describe this population in terms of a few physically or chemically meaningful combinations of the original variables, for example characterising sources of components. In other words, we want to transform the original set of variables into a space with fewer dimensions where the axes will describe the behaviour of the whole population more clearly. For this purpose principal component analysis and factor analysis are often used. Whenever we have some insight into the rules that are affecting our population, we may use this knowledge to construct a model for the factor analysis: we assume that a certain percentage of variability in the population is due to some well described causes (for ex. the composition of particles produced by wind erosion or the emission of certain pollutants by vehicles), while the rest is due to random effects influencing the measurements. We usually consider these random or error effects to be independent and identically normally distributed with a zero mean vector. The parameters characterizing such a model can be estimated. Tests of hypotheses for these parameters are well described, at least for normal populations (cf.eg. Lawley & Maxwell [16], Jørgenskog [11], Harman [12]). We shall not, however, consider the factor model in the present paper.

If our knowledge of the population is rather restricted, principal component analysis may be used to investigate the covariance or correlation structure of the underlying space. The present paper is limited to a discussion of the case when a sample correlation matrix is used. Our aim is to give some insight into the power of principal component analysis and to outline several practical directions about how to treat multiple response data in problems concerning air pollution.

The basic concept of principal component analysis, the necessary definitions, and a summary of properties of the principal components is briefly sketched in section 1. Section 2 contains a survey of statistical tests that can be used to control different hypotheses concerning the principal components of a correlation matrix, so that the reader should be able to apply them to his data, even if he is not very familiar with the mathematical statistics. Paragraph 3.2 indicates the problems connected with formulating and testing hypotheses about eigenvectors. To facilitate the reading of this paper, and as it is not possible to avoid some statistical terms, an abstract of terminology used in the statistical discipline of testing hypotheses is included as Appendix A. To illustrate the procedures described in 3.1.1 and 3.1.2 some examples are presented in Appendix B, using some of the precipitation data I had the opportunity to work with at NILU. For completeness, the tables of chisquare distribution are also included in Appendix D.

2 PRINCIPAL COMPONENT ANALYSIS (PCA)- DESCRIPTION AND OBJECTIVES

PCA is one of several methods which can be used to reduce the dimensionality of multiple response data: the main consideration here is the possibility to interpret the lower dimensional representation. It should be stressed here that the PCA does not necessarily yield directly interpretable results.

Circumstances, under which one may be interested in reducing the dimensionality of multiple response data include the following (cf. Gnanadesikan [9]):

1. exploratory situations in data analysis, especially when it is not known what is important in the measurement planning. One may want to screen out redundant coordinates (if any) or to find more insightful ones as a preliminary step to further data analysis or data collection.

2. Preliminary specification of a space that eventually is to be used as a basis for discrimination or classification procedures.

3. Situations in which one is interested in the detection of possible functional dependencies among observations in high-dimensional space.

A problem of particular interest in connection with transformation of coordinates and reduction of dimensionality is that the reduced coordinates should have a meaning or an interpretation that will facilitate an understanding of the problem, although the derived coordinates may not be directly observable.

PCA as a technique was first described by Karl Pearson in 1901. Further developement is due to Harald Hotelling [13], who also was the first to use the term principal components. Since then many statisticians have dealt with problems related to PCA, but the more widely used technique was factor analysis, applied mostly on psychological research problems (cf. Rummel [19] for further discussion and references.)

The basic idea of PCA is to describe the dispersion of an array of N points in a p-dimensional space by introducing a new set of orthogonal linear coordinates, so that the sample variances of the given points with respect to these derived coordinates are in increasing order of magnitude. The first principal component has maximum variance among all possible linear coordinates, the second principal component has maximum variance subject to being orthogonal to the first one and so on.

Principal components are not invariant under the linear transformation of original coordinates, including separate scaling. Therefore the principal components of the covariance matrix are not the same as those of the correlation matrix, or when some other type of scaling is used according to measures of importance. Note, that when the correlation matrix is used, the principal components are invariant to separate scaling of original variables. For this reason we are in favour of preforming PCA on the correlation matrix, especially when the scales of the measured variables are not comparable. However, for reasons of statistical nature (formal statistical inference, distribution theory, asymptotic theory) it is highly preferable to work with covariance matrix, though recently there are some results available for the correlation matrix, too.

Let $X_j' = (X_{1j}, \ldots, X_{pj})$, $j=1,\ldots,N$, be a random sample of size $N=n+1$, $n \geqslant p$, from a p-variate distribution with mean vector μ and a positive semidefinite covariance matrix $\Sigma = (\sigma_{ij})$, let

$$S = (s_{ij}) = \sum_{j=1}^{N} (X_j - \bar{X})(X_j - \bar{X})'$$

$$\bar{X} = (1/N) \sum_{j=1}^{N} X_j \ .$$

$\bar{X}$ is an unbiased estimate of μ, S/n is an unbiased estimate of Σ. Here and further on the prime denotes transposition of a vector resp. matrix. The population correlation coefficient between the i-th and the j-th component of the random vector is defined as

$$\varrho_{ij} = \sigma_{ij} \ / \ (\sigma_{ii}\sigma_{jj})^{\frac{1}{2}} \ .$$

The pxp matrix $P = (\varrho_{ij})$ is called the population correlation matrix. We estimate ϱ_{ij} as

$$r_{ij} = s_{ij}/(s_{ii}s_{jj})^{1/2} \ .$$

The pxp matris $R = (r_{ij})$ is called the sample correlation matrix. We need not suppose that the X_j's are normally distributed to formulate the theory of PCA, however, if the X_j's are drawn from a normal population, the statistical theory is considerably simplified.

Since P is positive semidefinite, there exists an orthogonal matrix H such that

$$H' P H = \Lambda$$

(1) or

$$P H = H \Lambda , \Lambda = \operatorname{diag} (\lambda_1 , \ldots , \lambda_p)$$

where

$$\lambda_1 \geqslant \ldots \geqslant \lambda_p$$

are the latent roots (or eigenvalues) of P, and the columns of

$$H_{pxp} = (h_1, \ldots, h_p)$$

are the corresponding orthonormal latent vectors of P. We can rewrite these relations as

(2)
$$h_j' P h_j = \lambda_j$$
$$h_j' P h_k = 0 \qquad j \neq k.$$

The linear combination $y_j = h_j' X$ is called the j-th principal component of P. It follows from (2) that the principal components are uncorrelated, and that the variance of the j-th principal component is λ_j. It can be shown that the $\operatorname{var}(h_1' X)$ is maximal among all linear combinations of X such that $h_j' h_j = 1$, and that the $\operatorname{var}(h_2' X)$ is maximal among all linear combinations of X such that $h_j' h_j = 1$, $h_j' h_1 = 0$, etc. Consequently, it can be shown that if $B_1, \ldots, B_k$, $k \leqq p$, is a set of orthonormal vectors in p-dimensional space, then

$$\lambda_1 + .. + \lambda_k = \max_{B_1, \ldots, B_k} (\operatorname{var}(B_1' X) + .. + \operatorname{var}(B_k' X))$$
$$= \operatorname{var}(h_1' X) + .. + \operatorname{var}(h_k' X) \quad , \quad k \leqslant p,$$

and that the linear prediction of X based on the first k characteristic vectors is optimal in terms of minimizing the residual variance.

It should be noted, that for the indicated linear transform $Y = HX$ the sum of the variances of all the principal components is p - the same as the sum of the variances of the original normalized variables. Also, the generalized variance (of the normalized variables, i.e. the determinant of P) of the population is preserved. This follows from the fact that H is orthonormal. For proofs of these results cf. Anderson [1], Rao [21].

To estimate the characteristic roots and the characteristic vectors of the population correlation matrix P we compute the characteristic roots and vectors of its sample equivalent R.

When the X values are drawn from a normal population and P has p distinct characteristic roots, the estimates of the corresponding population parameters are of maximum likelihood type (it can be shown in the same way as for the covariance matrix, cf. Anderson [1]).

However, let us note that the basis of the PCA is a spectral decomposition of a positive semidefinite matrix. The characteristic roots of such matrix are always real and non-negative, and the characteristic vectors are real. It is only to facilitate the development of the statistical theory, that we require a random sample drawn from a normal population. But it is clear that when one wants to construct a general basis for the theory and its validation, this is a vital consideration, since the statistical inference in non-normal cases is very complicated and of restricted practical value till now.

3 TESTING HYPOTHESES WITHIN THE PCA

By a decomposition of the correlation matrix into the principal components we have obtained a set of new variables - each a linear combination of the original ones. Now two basic problems arise: the first and perhaps the most important one is to find whether we can ascribe some meaning to the new variables. It may not be always true: when the original variables are chosen so that they do not describe the population sufficiently in the sense that the set of original variables is highly intercorrelated, then the largest eigenvalue will be close to p - the number of variables, while the others will be close to zero. In a less boundary case we may obtain a new set of variables such that there are only some sources of variability mixed together in one linear combination. Unless we introduce a new variable that is able to distinguish between these sources, we can not separate them by means of any analysis of linear dependencies. In other words, a failure in interpreting results of the decomposition is likely to lead us to the conclusion that the chosen variables do not provide a suitable description of the population under study.

The second problem concerns statistical inference: we would like to know which principal components represent only random (or error) influences and which can be ascribed to a specific well-described process. Also the true value of components of each principal component vector is of interest, e.g., to deduce that the coefficients of the original variables are either zero or nonzero with some probability.

We shall reformulate these vague questions and describe them in the language of statistical hypotheses.

3.1 Hypotheses about the rank of a population correlation matrix P.

One of the fundamental problems in both PCA and factor analysis is the determination of the number of "common factors" to be used as a basis for a further description of the population. There are several criteria, most of them derived for the factor model, but they are applicable to similar problems in PCA. We wish to determine (on some significance level) the number of substantive influences in our physical population. This will generally be dependent on the number of original variables and on the size of the sample, since the confidence region is a function also of these values, even when using heuristic criteria. Also, the model should fulfill the condition that the number of parameters we wish to estimate is less than the number of observations, otherwise the problem is singular and no statistical inference is possible. On the other hand, having a finite number of variables describing a theoretically infinite population we can always find a finite number of components to fit our observations.

3.1.1 Distribution independent criteria

3.1.1.1 Three lower bounds to the rank of P

Let us consider a population correlation matrix P real positive semidefinite. Let U^2 be an arbitrary real diagonal matrix such that the j-th diagonal element denoted by u_j^2 satisfies

$$0 \leqslant u_j^2 \leqslant 1 \ , \ j = 1,\ldots,p.$$

Then we define a symmetric matrix $G_{p \times p}$ by

$$G = P - U^2 \tag{3}$$

and our objective is to find a matrix U^2 such that G will remain real positive semidefinite with the smallest possible rank. Actually, the relation (3) states a factor model, as U^2 might be regarded as the covariance matrix of unique factors or uncorrelated errors. In (3) we have subtracted from P the variance due to the causes influencing the original variates independently, and we want to estimate the minimum number of causes influencing them as a whole, that is , the rank of G. Again, we are looking only for linear dependencies.

Guttman [10] has found three lower bounds to the minimum rank of G making no additional assumptions about U^2 or about an underlying distribution whatsoever.
He defines the following boundaries s_1, s_2, s_3:

- s_1 equals the number of eigenvalues of P greater than or equal to unity.
- s_2 equals the number of non-negative eigenvalues of the matrix $S_2 = P - D_2$,

where D_2 is a diagonal matrix whose j-th diagonal element is equal to $1-r_j^2$, $j=1,\ldots,p$, r_j denotes the multiple correlation coefficient of the j-th observed variable with the remaining p-1 observed ones.

s_3 equals the number of non-negative eigenvalues of the matrix $S_3 = P-D_3$, where D_3 is the diagonal matrix whose j-th diagonal element is equal to $1-\tilde{r}_j^2$, where $\tilde{r}_j$ is the maximum correlation coefficient between the j-th observed variable and any of the p-1 remaining ones.

Let k be an unknown minimal rank of G given P. Then using only linear algebra it can be shown that $k \geqslant s_2 \geqslant s_3 \geqslant s_1$.

So far we have dealt only with the population characteristics, now we shall use their sample equivalents. The eigenvalues of the sample correlation matrix R give us directly the estimate of s_1. This lower bound to the rank of P has also another practical justification: it determines the number of principal components with a variance larger than or equal to the variance of each primary normalized variable. Let us note in this connection that the mean value of the eigenvalues of P resp. R is 1.

All the three lower boundaries should be used carefully: they certainly underestimate the proper number of components corresponding to the global influences we are looking for.

3.1.1.2 <u>The scree test</u>

This criterion was proposed by Cattel [4] and it is based on his empirical knowledge. To give an impression of the underlying philosophy, let us use Cattel's own words: "In any case the scree-test does not rest for its practical validity upon the correctness of the theory or inferences from it, but on an inductive law, some of the empirical evidence for which is presented here ...". And indeed, the reader of Cattel paper can find a lot of empirical evidence there. A theoretical justification of the suggested criterion is not to be easily found (it can hardly be called a test in the statistical sense), nevertheless, I am quite convinced of its value as a practical guide, when other criteria are employed, too.

As a basis for determining how many of the principal components express non-trivial processes (physical influences), we use a plot of the eigenvalues as shown in Figure 1:

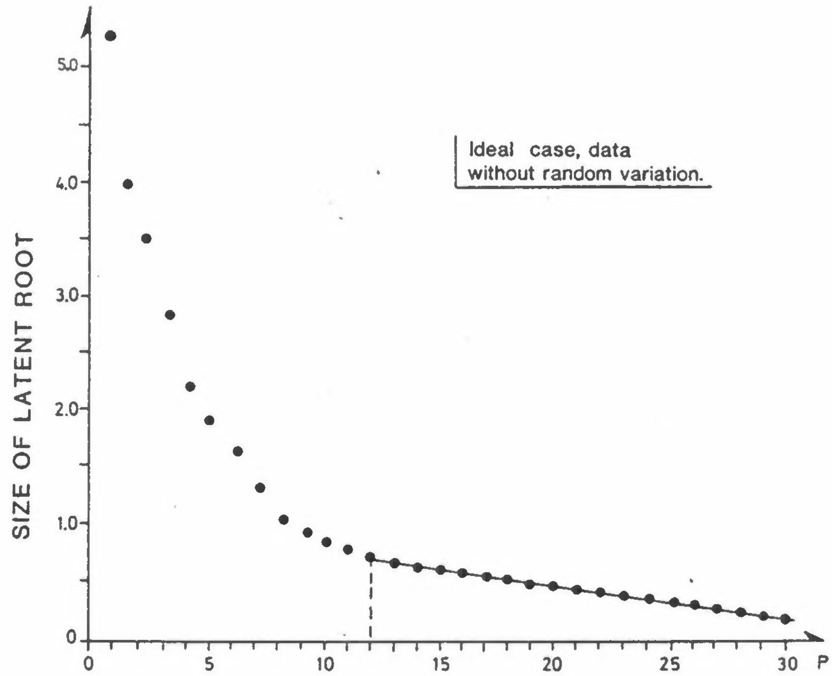

Figure 1: Eigenvalues corresponding to 30 factors - ideal case.

This plot first falls off steep by, and then straightens out in a line which runs only with small irregular deviations from straightness (fig. 1 shows an ideal case). This straight end part we call the scree - "from the straight line of rubble and boulders which forms at the pitch of sliding stability at the foot of mountain", to quote Cattel again. The implication of this is that this scree represents small error factors. The criterion then is to consider all components corresponding to the eigenvalues rising above the scree physical meaningful.

Finally, let us outline some motives that led Cattel to adopt this heuristic theory. He argues that

1. it is not possible to describe the population in terms of a smaller number of linear components; the cut-off point is determined in an objective manner using a concept of non-trivial common variance, which may be adjusted at 95% or 99% of total variability, according to the circumstances,
2. the model for the scree-test is a "complex stratified factor model", different from both the PCA and factor analysis models. It considers contributions from factors of temporarily-specific origine, general error factors and a truly specific primary factors (so-called unique factors), in addition to the variance in each variable accounted for by the substantive common or general physical factor.

3.1.2 Criteria for samples from a normal population

The tests described in this section are based on the result obtained by Wilks [24], who has shown that under certain conditions imposed on the population distribution, the asymptotic distribution of the logarithm of likelihood ratio is chi-square with degrees of freedom corresponding to the difference between hypothesis and alternative.
Let us assume that we have a random sample drawn from a p-variate normal distribution $N_p(\mu,\Sigma)$
with mean vector μ and covariance matrix Σ. First we shall consider a hypothesis

$$H_0 : \lambda_1 = \ldots\ldots\ldots = \lambda_p = 1 \quad .$$

that the characteristic roots λ_i, $i=1,\ldots,p$, of the population correlation matrix $P_{p\times p}$ are equal and therefore equal to 1. This hypothesis states that

$$P = I \; ,$$

where I denotes the identity matrix, i.e. that the original variables are uncorrelated and therefore - since the underlying distribution is assumed normal - independent. If H_0 holds, there is no point in trying to find a new set of uncorrelated variables as the conditions imposed on the principal components are met by the original variables. A suitable statistics for the test of H_0 against the general alternative that H_0 does not hold is

$$\Lambda_0 = \ln(|R|) = \ln \prod_{j=1}^{p} l_j,$$

where l_j denotes the sample values of λ_j (i.e.,the eigenvalues of R), the $|R|$ denotes the determinant of the sample correlation matrix R, the ln(x) denotes the natural logarith of x. Bartlett [3] has shown that the expression

$$T_0 = -\{n - \frac{1}{6}(2p+5)\}\Lambda_0, \quad n = N-1$$

is asymptotically (for n tending to infinity) distributed after a chi-square distribution with p(p-1)/2 degrees of freedom. To test H_0 we compute the value of T_0. We reject H_0 on the level α when

$$T_0 > \chi^2_{p(p-1)/2}(1-\alpha),$$

and we can not reject H_0 when the opposite inequality is true. $\chi^2_{p(p-1)/2}(1-\alpha)$ denotes the $100(1-\alpha)\%$ quantile of a chi-square distribution with p(p-1)/2 degrees of freedom, α is the chosen significance level; we usualy take $\alpha = .05$ or $\alpha = .01$.

If in the given set of variables some observed correlation coefficients are high (say .95 or higher), then several variables are likely to be lineary dependent and therefore the correlation matrix is near to a singular one. In this case H_0 will be rejected. Also it is quite probable that we shall meet a considerable computational troubles when trying to compute the eigenvalues of a singular matrix R, especially when the number of variables is large, as most of the algorithms do not converge to a suitable solution in such a border case. Both these problems are simplified without much loss of information by removing one or more of the most highly correlated variables from the set under study.

Nagarsenker [18] has derived an exact null distribution of the determinant of the correlation matrix R by using techniques based on series expansion of certain functions. His aim was to test $H_0: P=I$ and in addition to the exact distribution of $|R|$ he computed also its significance points for $\alpha = .05$ and $\alpha = .01$ for a number of variables ranging from 3 to 8 and sample sizes from 4 to 100. The tables are reproduced in Appendix C.

Let us consider a more general hypothesis about the population correlation matrix P in the form

$$H_1: \quad \lambda_j = \lambda_{j+1} = \ldots = \lambda_p \qquad , 1 \leqslant j \leqslant p ,$$

based again on the indicated random sample from $N_p(\mu, \Sigma)$. H_1 includes H_0 as a special case when $j=1$. H_1 states that the $p-j$ smallest eigenvalues of P are equal, that is, we can not distinguish between the variances of the corresponding principal components. If the last eigenvalues are small enough, and if H_1 holds for some j, $j < p$, and does not hold for $j+1 \leqslant p$, then we can consider the corresponding principal components to be the result of some trivial random process, or simply we can focus our further interest on the remaining more important (in the sense of larger variance) ones.

The statistic used to test H_1 against the general hypothesis that H_1 does not hold is

$$\Lambda_1 = (\prod_{k=j}^{p} l_\kappa)/\{[1/(p-j)]\sum_{k=j}^{p} l_\kappa\}^{p-j} \quad .$$

For N large a null distribution of T_1,

$$T_1 = -(n - \frac{1}{6}(2p+5) - \frac{2}{3}(j-1)) \ln \Lambda_1 \ ,$$

can be approximated as a chi-square distribution with $(p-j-1)*(p-j+2)/2$ degrees of freedom. We reject H_1 when $T_1 \geqslant \chi^2_{(p-j-1)(p-j+2)/2}(1-\alpha)$. The number of degrees of freedom connected with the test of H_1 depends even asymptotically on the amount of variance removed from H_1 with $(p-j-1)(p-j+2)/2$ being the maximum value. For details see Bartlett [3], Lawley [15], Rao [20], Konishi [14], Anderson [2].

Of particular interest would be to test a hypothesis

H_1^* : $\lambda_j = \lambda_{j+1} = .. = \lambda_{j+a}$ that is that the eigenvalues from a subset $\lambda_j \geq \lambda_{j+1} \geq ... \geq \lambda_{j+a}$ of $\lambda_1 \geq ... \geq \lambda_p$ have the same value. This would mean that the corresponding principal components are of the same importance, again in the sense of equal variances. The statistic used could be

$$T_1^* = -\left(n - \frac{1}{6}(2p+5)\right) \ln \Lambda_1^*$$

where

$$\Lambda_1^* = \left(\prod_{k=j}^{j+a} l_j \right) / [(a+1)^{-1} \sum_{k=j}^{j+a} l_j]^{a+1}, \quad 1 \leq p-a,\ a<p.$$

But even the asymptotic distribution of T_1^* is not generally solved.

Another special case of interest is to consider $P=P_2$, where

$$P_2 = \begin{pmatrix} 1 & \varrho & \varrho & \cdots & \varrho \\ \varrho & 1 & \cdots & \cdots & \varrho \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \varrho \\ \varrho & \cdots & \cdots & \varrho & 1 \end{pmatrix}$$

that is

$$P_2 = (1-\varrho)I + \varrho ee' ,$$

where I_{pxp} is the identity matrix, ϱ denotes the common value of the correlation coefficients and e is a vector $e_{px1} = (1,..,1)'$. P_2 reflects the situation when the population is affected by only one nontrivial source of variability. The eigenvalues of P_2 are $\lambda_1 = (1+(p-1)\varrho)$ of multiplicity 1 and $\lambda_2 = (1-\varrho)$ of multiplicity p-1.

To test the hypothesis

$$H_2: P = P_2$$

against the alternative that H_2 does not hold we can use the results summarized by Gleser [8]. On the basis of a random sample of size N=n+1 from $N_p(\mu,\Sigma)$ we can chose between two statistics T_{21} and T_{22}.

$$T_{21} = N \left(\log\left(\sum_{i=2}^{p} l_i \right) - \sum_{i=2}^{p} \log l_i \right)$$

where l_i denotes the i-th largest eigenvalue of the sample correlation matrix R. T_{22} can be computed without any knowledge of the eigenvalues of R:

$$T_{22} = \frac{1}{\lambda_2} \left(\sum_{i<j} (y_{ij} - \bar{y})^2 - \gamma \sum_{k=1}^{p} (\bar{y}_k - \bar{y})^2 \right)$$

where

$$\lambda_2 = (1-\varrho),$$

$$\gamma = (p-1)^2 (1-\lambda_2^2)(p-(p-2)\lambda_2^2)^{-1},$$

$$\bar{y}_k = (p-1)^{-1} \sum_{i \neq k} y_{ik},$$

$$\bar{y} = \{p(p-1)\}^{-1} \sum_{i \neq j} y_{ij}$$

$$y_{ij} = n^{1/2} (r_{ij} - \varrho), \ i \neq j,$$

r_{ij} is a sample correlation coefficient between ith and jth variable; λ_2 and γ can be computed from the data by replacing ϱ by $\bar{\varrho}$ (the mean value of the correlation coefficients r_{ij}):

$$\bar{\varrho} = (p(p-1))^{-1} \sum_{i \neq j} r_{ij} .$$

Under H_2 the asymptotic distribution of T_{21} is

$$\pounds_1 = \chi^2_{p(p-3)/2} + (1 - \frac{p-2}{p} \lambda_2^2) \chi^2_{p-1}$$

and of T_{22} is

$$\pounds_2 = \chi^2_{(p+1)(p-2)/2}.$$

In setting up the test of H_2 we want to restrain the maximum probability of type I error, i.e. of rejecting H_2 when it holds. With this consideration in mind we obtain the same rejection region when using T_{21} as for T_{22}, namely

(4) $\quad T_{2i} \geqslant \chi^2_{(p+1)(p-2)/2}(1-\alpha)$, $\quad i=1,2,$

where $\chi^2_{(p+1)(p-2)/2}(1-\alpha)$ is the $100(1-\alpha)\%$ quantile of chi-square distribution with $(p-2)(p+1)/2$ degrees of freedom; α again denotes the chosen significance level. We reject H_2 when (4) is true, otherwise H_2 can not be rejected.

3.1.3 Samples from a non-normal population

As a first step in any statistical analysis we should evaluate statistical characteristics of the variables under consideration, say the first and second moments and several quantiles. This will give us some ideas about the marginal distribution from which every variable was drawn. Often, when dealing with chemical data, the supposed distribution is lognormal, that is, the logarithm of a theoretical value is distributed according to the normal law. Several tests are available for hypotheses about the shape of the distribution (e.g. a chi-square test of goodness of fitt, cf. e.g. Rao [21]), and when the hypothesis of a lognormal distribution can not be rejected, we may deal with the logarithm of the data as with a random sample from a normal population. A normalizing transformation can be found also for other types of distribution functions.

Sometimes no knowledge about the shape of the distribution is available, but we may usually suppose that several first moments exist. We may also suppose that the unknown distribution function is differentiable with respect to both parameters and a random variable. Then under a general conditions the theory of the logarithm of likelihood ratio partly described in 2.1.1. holds, but the problem lies in finding suitable estimators of the parameters, because the properties of the estimators are dependent on the shape of the unknown distribution of the considered random variable. As we have already seen, the general theory is far from simple, even in the normal case, so we can foresee even more difficulties when less is known. Here we shall contend ourselves by stating that so far only the asymptotic distribution of certain functions of the eigenvalues of a sample correlation matrix as well as the asymptotic distribution of its latent vectors have been developed. The distributions can be derived as shown by Dawis [6], Fang & Krishnaiah [7] and others, but the theory is not yet ready for direct practical use. Recently also two other approaches to PCA have appeared in the literature. The first is a robust PCA by Ruymgaart [23]. The second gives some general results obtained when applying PCA to stochastic processes (cf. Daudiox, Pousse and Romain [5]), but there is still some way to go before we shall be able to use these results for the problems outlined above.

3.2 Hypotheses about eigenvectors of P

This section is based on the asymptotic results obtained by Konishi [4], and its purpose is only to indicate a possible way to achieve a more complete statistical analysis of the results obtained by PCA.

Let the sample correlation matrix R be based on $N=n+1$ observations from a p-variate normal distribution with positive definite covariance matrix Σ. Let $\lambda_1 \geq \ldots \geq \lambda_p > 0$ be the ordered eigenvalues of the population correlation matrix P, and let $h_1, \ldots, h_p$ be the corresponding orthonormal eigenvectors of P, so that

$$H'P\,H = \Lambda \quad , \quad H'H = I \quad ,$$

where $\Lambda_{p\times p} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ is a diagonal matrix and $H_{p\times p} = (h_1, \ldots, h_p)$. We shall now consider two hypotheses:

$$H_3: \quad h_g = h_{g0}$$

that the normalized eigenvector h_g (i.e. $h_g' h_g = 1$) corresponding to the distinct eigenvalue λ_g of multiplicity 1 of P is equal to a specified vector h_{g0} such that $h_{g0}'h_{g0} = 1$, and

$$H_4: \quad h_j = h_{j0} \ , \ j = 1, \ldots, a \quad , \ a \leq p$$

that a specified set of orthonormal vectors are eigenvectors of P. Let us denote by f_g an eigenvector corresponding to the g-th largest eigenvalue of the sample correlation matrix R. First, let us focus on H_3. Konishi has shown that

$$T_3 = n(f_g - h_g)' H_g \; Q_g^{-1} \; H_g' \; (f_g - h_g)$$

has a limiting chi-square distribution with p-1 degrees of freedom. Here $H_g = (h_1, \ldots, h_{g-1}, h_{g+1}, \ldots, h_p)$ and $Q_g = (q_{ij.g})$ $i,j \neq g$ are a (p-1)x(p-1) matrices,

$$q_{ij.g} = (\lambda_i - \lambda_g)^{-1} (\lambda_j - \lambda_g)^{-1} \{\delta_{ij}\lambda_i\lambda_j - (2\lambda_i\lambda_j\lambda_g + \lambda_g^2(\lambda_i+\lambda_j))* \sum_{k=1}^{p} h_{ki}h_{kj}h_{kg}^2 + \frac{1}{2}(\lambda_i+\lambda_g)(\lambda_j+\lambda_g) \sum_{k=1}^{p} \sum_{l=1}^{p} \varrho_{kl}^2 \; h_{ki}h_{lj}h_{kg}h_{lg}\}.$$

δ_{ij} denotes the Kronecker's delta function δ_{ij}: $\delta_{ij} = 1$ $i=j$, $\delta_{ij} = 0$ $i \neq j$. Testing H_3 we shall replace h_g by a specified h_{g0} and we shall estimate the unknown parameters λ_g, h_{ij}, $j \neq g$, ϱ_{ij} by their sample values. After evaluating the T_3 we shall reject H_3 on the significance level α if

$$(5) \qquad T_3 \geqslant \chi^2_{p-1} \; (1-\alpha),$$

otherwise H_3 can not be rejected. The symbols in (5) are used with the same meaning as in section 3.1.2.

H_4 is even more complicated. Also it may not be easy to formulate H_4 so that the designed set of h_{j0}, $j=1,\ldots,a$ is orthonormal. Therefore we shall only sketch the idea of deriving the test with no details; an interested reader is referred to Konishi [14]. Let us denote

$$H_{10} = (h_{10}, \ldots, h_{a0})$$

$$\Lambda_a = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_a) \quad .$$

Let $H_2 = (h_{a+1}, \ldots, h_p)$ be any px(p-a) matrix such that $H = (H_{10}, H_2)$ is an orthogonal pxp matrix. Then using the presumption that H_4 holds, Konishi [14] suggests the statistic T_4 for the test H_4:

$$T_4 = N \ln \left(\prod_{j=1}^{a} (H_{10}' R \; H_{10})_{jj} \; |H_2' R \; H_2| \, |R|^{-1} \right) \quad .$$

The problem of finding its distribution he labels as "intractable" (as in general T_4 will not have a chi-square distribution); however, a chi-square approximation could be obtained using the expectation of T_4 as the degrees of freedom (cf. Konishi [14], p.681).

4 CONCLUDING REMARKS

The presented paper outlines possible ways to handle and analyze multiple response data. Several rather simple criteria for reducing the dimensionality of these have been described, so that a reader interested in using the results in practice would find all the necessary information here. If the problem is more complicated, and no previous use of theory in applications is referred in literature as in case of sections 3.1.3 and 3.2, our aim was to provide a brief description and to indicate possible further literature on the topic.

Finally, I would like to point out a few more papers and textbooks dealing with PCA in a way accessible to non-mathematicians. Morrison [17] has dedicated two chapters of his book to PCA and factor analysis, pointing out for ex. their interpretation and sampling properties, as well as the most frequent special cases. The books of Anderson [1] and Rao [21] are written for statisticians,but especially in the latter others may also find a lot of inspiration. The book of Gnanadesikan [9], is even more readable. A valuable and detailed examination of the role of the PCA in applied research is provided by Rao [22].

Acknowledgement

I am strongly indebted to Dr B. Ottar for his helpful comments and suggestions as well as for correcting my English in the main part of this paper.

5 REFERENCES

[1] Anderson, T.W. An introduction to multivariate statistical analysis. New York, Wiley, 1958.

[2] Anderson, T.W. Asymptotic theory for principal component analysis. Ann. Math. Statist., 34,122-148 (1963).

[3] Bartlett, M.S. A note on the multiplying factors for various Chi-square approximations. J. Roy. Statist. Soc. Ser. B, 16, 296-298 (1954).

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

[5] Daudoix, J., Pousse, J., Romain, Y. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. J. Multivar. Anal., 12, 136-154 (1982).

[6] Dawis, A.W. Symptotic theory for principal component analysis: Non-normal case. Austral. J. Statist., 19, (3), 206-212 (1977).

[7] Fang, C., Krishnaiah, P.R. Asymptotic distributions of functions of the eigenvalues of some random matrices for non-normal populations. J. Multivar. Anal., 12, 39-63 (1982).

[8] Gleser, L.J. On testing a set of correlation coefficients for equality: Some asymptotic results. Biometrika, 55, 513-517 (1968).

[9] Gnanadesikan, R. Methods for statistical data analysis of multivariate observations. New York, Wiley, 1977.

[10] Guttman, L. Some necessary conditions for common-factor analysis. Psychometrika, 19, 149-161 (1954).

[11] Jörenskog, K.G. Statistical estimation in factor analysis. Stockholm, Almqvist and Wiksell, 1963.

[12] Harman, H.H. Modern factor analysis, 2nd. ed. University of Chicago Press 1967.

[13] Hotelling, H. Analysis of a complex of statistical variables into principal components. J. Educ. Physchol., 24, 417-441, 498-520 (1933).

[14] Konishi, S. Asymptotic expansions for the distributions of statistics based on the sample correlation matrix in principal component analysis. Hiroshima math. J., 9, 647-700 (1979).

[15] Lawley, D.N. Asymptotic expansions for the distributions of statistics based on covariance and correlation matrices. Biometrika, 43, 128-136 (1956).

[16] Lawley, D.N. Maxwell, A.E. Factor analysis as a statistical method. 2nd. ed. London, Butterworth, 1971.

[17] Morrison, D.F. Multivariate statistical methods. New York, McGraw-Hill, 1967.

[18] Nagarsenker, B.N. The distribution of the determinant of correlation matrix useful in principal component analysis.
Commun. Statist. - Simula. Computa. B5, (1), 1-13 (1976).

[19] Rummel, R.J. Applied factor analysis.
Evanston, Ill., Northwestern Univ. Press, 1970.

[20] Rao, C.R. Estimation and tests of significance in factor analysis.
Psychometrika, 20 (2), 93-111 (1955).

[21] Rao, C.R. Linear statistical inference and its applications.
New York, Wiley, 1965.

[22] Rao, C.R. The use and interpretation of principal component analysis in applied research.
Sankhyã, 26A, 329-357 (1965).

[23] Ruymgaart, F.H. A robust principal component analysis.
J. Multivar. Anal., 11, 485-497 (1981).

[24] Wilks, S.S. The large-sample distribution of the likelihood ratio for testing composite hypotheses.
Ann. Math. Statist., 9, 60-63 (1938).

APPENDIX A

AN INTRODUCTION TO TESTING OF STATISTICAL HYPOTHESES

AN INTRODUCTION TO TESTING OF STATISTICAL HYPOTHESES

Let us have a theoretical probability space and a random variable X defined on it. Let S denote the sample space of outcomes of an experiment (through which we observe the probability space) and x denote an arbitrary element of S, say, S being a (p-dimensional) real Euclidean space with (p-dimensional) vectors as its elements and (p-dimensional) intervals as sets in S. Let H_0 be a hypothesis (to be called a <u>null hypothesis</u>) which specifies partly or completely the distribution function over the sets in S. Clearly, x is an observed value of X, and the distribution function describes the properties of the population of X's. The problem of testing of hypotheses is then to decide on the basis of an observed x, whether H_0 is true or not.

Whatever procedure may be employed for testing a null hypothesis, that is, deciding whether to reject H_0 or not, there are two types of error involved, viz., that of rejecting H_0 when it is true (<u>the type I error</u>; its probability is called the level of significance), and not rejecting H_0 when an <u>alternative hypothesis</u> is true (<u>the type II error</u>).

A test procedure consists in dividing the sample space into two regions, w and S-w, and deciding to reject H_0 if the observed x falls into w and not to reject H_0 otherwise. We call w <u>the critical region</u>. To test H_0 we may write a function T defined over S as a function T(X) of X with the value T(x) when X=x, where X denotes a theoretical random variable, x its sample equivalent. T is called <u>a test function</u> or, sometimes, <u>a test statistic</u>. The term statistic in general denotes a random variable, and consequently, a function of a random variable.

There exists a class of tests T with an optimal property of minimizing the type II error when the probability of type I error is prescribed - the <u>likelihood ratio tests</u>. When we consider the distribution function as a function of its parameters instead of the random variable, we call it a likelihood function. As a likelihood ratio we denote the ratio of the likelihood function with the parameters (now regarded as a variables) specified by alternative hypothesis divided by the likelihood function with the parameters specified by the null hypothesis.

A likelihood ratio or some other principle provides us with a statistic T whose value is determined on the basis of the random sample. The function T(X) itself is a random variable - a transformation of the random variable X - so that we are usually able to ascertain its distribution. The value T(x) enables us to decide whether x is an element of w, that is, whether or not to reject H_0. The boundaries of critical regions (on level α) for several most common distributions of T(X) are tabulated in statistical tables - mostly as a $100(1-\alpha)\%$ quantiles of the distribution of T (X).

Another problem is to estimate the parameters of a theoretical distribution function. For this purpose we first derive a suitable estimator, that is, a function of the original random variable. Its distribution again can usually be found. Here, too, we can use the likelihood function as a basis. To estimate the parameter on the basis of a given random sample we calculate a value corresponding to this sample for the estimator - we obtain the point estimate. Using the distribution of the estimator, we can determine the confidence region for this point estimate on a level α, that is, the random set which with probability $(1-\alpha)$ contains the true value of the parameter.

The reader should not be confused by alternative use of $1-\alpha$ and α values. Their use differs from author to author, the meaning is usually clear from the context.

APPENDIX B

Several examples

SEVERAL EXAMPLES

We shall illustrate with several examples some of the tests described in sections 3.1.1 and 3.1.2. We shall use some of the precipitation data obtained from background stations in Norway in 1980, viz.,N15 Tustervatn and the adjoining meteorological station Bolna, and meteorological stations at Røros and Tynset. Station N15 was selected because of the small number of days with precipitation (95), so that the tables computed by Nagarsenker can be used. The stations at Røros and Tynset are separate from Bolna. They are supposed to be sufficiently close to represent the same population.

From the analysis of separate variables (which is not presented here) we could conclude that the chemical variables are distributed according to the log-normal distribution, while the meteorological variables except precipitation amounts are mixtures of normal distributions. The observed distribution of the precipitation amounts is usually very close to the exponential law. To obtain a sample from a population as close to a normal one as possible we logarithmically transform the chemical variables; in the course of the present analysis no transformation was used on the precipitation amounts. As the assumption of normality may not be fulfilled, the results based on the normal theory should be regarded as proximate. Also, we have neglected possible autocorrelation within the data, which violates the assumption of random sample. No regard was taken to possible time dependence either.

The characteristics of variables, correlation matrices and corresponding eigenvalues and eigenvectors were computed on a NORD 100 computer using the double precision arithmetics for the eigenvalue analysis. The original programme was written by R.C. Henry. The values of remaining statistics were obtained with the aid of the scientific calculator Sharp EL-512.

The variables are listed in Table B1 and their characteristics are given in Table B2 for the stations Tustervatn and Bolna. The correlation matrix R_0 of all these variables is presented in Table B3 together with the matrix of coefficients of the principal components of R_0. Tables B4 and B5 contain the results of statistical analysis of correlation matrix R_0. Figure B1 shows the plot of eigenvalues with the scree marked.

The character of presented tests and estimators can be seen from the analysis of R_0. Table B2 can be used also for the tests of normality: critical values for skewness and kurtosis of normal distribution can be found in tables (e.g. in Snedekor,G.W. & Cochran,W.G.: Statistical methods, The Iowa State University Press, Ames, Iowa, U.S.A. 1937 and many later editions). We should be aware of errors in our data: outlying observations considerably increase the values of skewness and kurtosis. Tables B4 and B5 show that we usually underestimate the number of principal components which are needed to describe the population. When we are looking for some well-described influences we should bear this fact in mind. After the ana- lysis of R_0 I have selected three of its submatrices:

R_1 (Tables B6 and B7) with at least 2 non-trivial components,
R_2 (Tables B8 and B9) with at least 3 non-trival components,
and R_3 (Tables B10 and B11) with also at least 3 nontrivial principal components.

The analysis of meteorological data from Røros and Tynset (Tables B12-B14 and B15-B17) show that the stations are similar in many ways. For a valid inference about this similarity we need some other statistical technique, e.g., the canonical correlations. The departure from normality of TEMP and all the windspeeds is due to their being very likely a mixture of at least two normal distributions.

Whenever there is an asterisk (*) attached to the test value, (except in the tests for normality - skewness and kurtosis), it means that on the basis of this test value we reject the hypothesis on the level α= .05. Two asterisks (**) mark the rejection on the level α= .01. In the tables of skewness and kurtosis asterisk (*) marks the significance level α=. 10, two asterisks denote the level $\alpha = .02$. For the tests about kurtosis the levels are only proximate (we have used one-tailed tables and the distribution is not symmetrical).

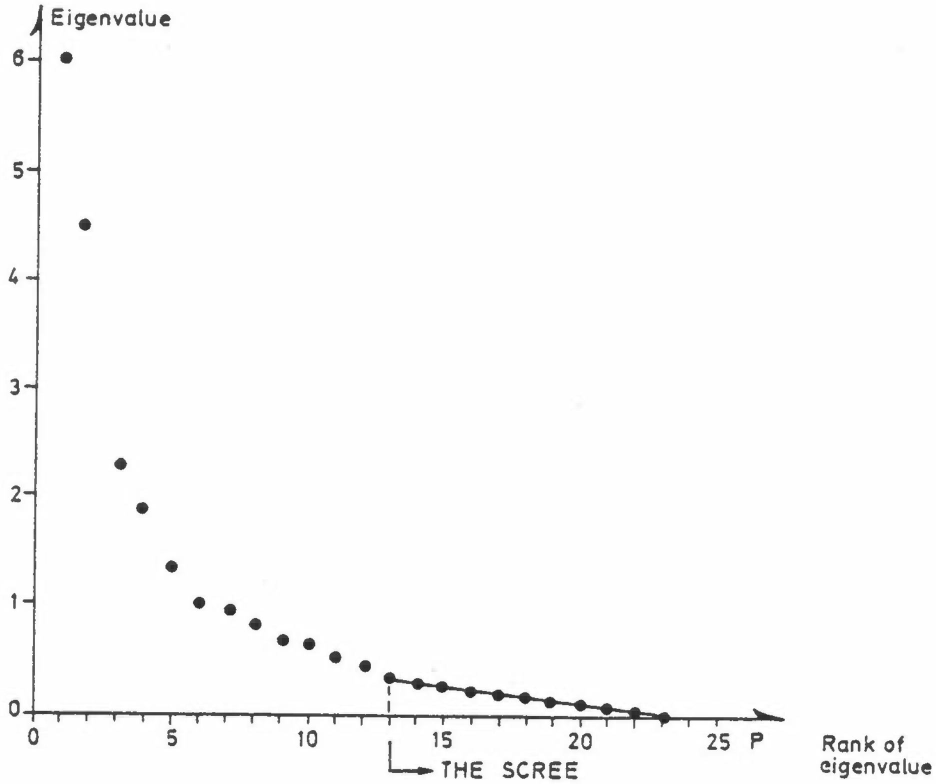

Figure B1: The scree for correlation matrix R_0 (all variables from Tustervatn and Bolna included).

Table B1

LIST OF VARIABLES

MM1	Precipitation amount
SO4C	Sulphate in precipitation (corrected for sea salts)
NH4	Ammonium in precipitation
NO3	Nitrate in precipitation
NA	Sodium in precipitation
MG	Magnesium in precipitation
CA	Calcium in precipitation
H	Strong acid in precipitation
K	Potassium in precipitation
COND	Conductinity
PSUM	Sulphate in aerosols (measured)
SO2M	Sulphur dioxide in aerosols (measured)
U850	Wind speed, east-west component at 850mB
V850	Wind speed, north-south component at 850mB
SO2C	Sulphur dioxide in aerosols (computed)
PSUC	Sulphate in aerosols (computed)
MM4	Precipitation amount at neighbour meteorological station
TEMP	Air temperature at neighbour meteorological station
HUMI	Relative humidity at neighbour meteorological station
USUR	Windspeed, east-west component at surface at neighbour meteorological station
VSUR	Windspeed, north-south component at surface at neighbour meteorological station
WSPE	Windspeed, total at surface at neighbour meteorological station

Table B2

VARIABLE	MEAN	STD. DEV.	VARIANCE	SKEWNESS	KURTOSIS	MAXIMUM	MINIMUM
MM1	1.6727	.6753	.4560	.1032	-.4024	3.2465	.1823
SO4C	-1.8979	1.3447	1.8082	-.0556	-.2143	1.1606	-5.2983
NH4	-2.3617	1.2679	1.6077	.4407	-.7919	.4447	-3.9120
NO3	-2.9218	1.2506	1.5640	.2011	-.4205	.1310	-5.2983
NA	-1.0545	1.3756	1.8922	.3924	-.6239	2.7726	-2.9957
MG	-2.9365	1.2789	1.6357	.3380	-.4520	.6729	-5.2983
CA	-1.7869	1.0182	1.0367	1.7674**	3.0894**	2.1041	-2.9957
H	1.5110	1.7491	3.0595	-.9135*	2.0057**	5.0657	-5.5215
K	-2.0603	.8885	.7895	-.2997	1.7585*	.2231	-5.2983
COND	2.4224	.7210	.5198	.5227*	-.1589	4.2556	1.0986
PSUM	-1.3473	1.1117	1.2359	.4339	-.5593	1.3083	-3.5066
SO2M	-1.6565	1.0171	1.0345	1.4552**	2.5984**	2.2925	-2.9957
U850	1.1218	1.1366	1.2919	-.5315*	-.2088	2.8303	-2.8134
V850	1.1630	1.1150	1.2432	-.3156	-1.0092*	2.8009	-1.3863
SO2C	-2.5817	1.2567	1.5793	.9783**	1.4516*	2.2680	-4.6052
PSUC	-4.8280	3.8869	15.1081	-.0528	-1.6695**	1.7156	-9.2103
SUDE	.5282	1.7142	2.9384	-3.7376**	19.2290**	3.3586	-9.2103
MM4	1.0601	1.1506	1.3239	-.6231*	.1705	3.3673	-2.3026
TEMP	-.4006	7.9689	63.5034	-.0743	-.4129	16.7653	-22.8514
HUMI	86.2303	8.1840	66.9776	-1.1641**	1.8119*	98.0972	56.0139
USUR	.6041	3.4940	12.2078	-.2526	-1.1007**	7.5154	-6.3579
VSUR	1.7741	3.5598	12.6719	-.7918**	-.5956	7.0175	-6.3991
WSPE	5.1915	1.0912	1.1907	.4904*	-.0856	8.3304	2.9435

* ... α= .10 significance level
** ... α= .02 significance level

Table B3

Correlation matrix R_0

	MM1	SO4C	NH4	NO3	NA	MG	CA	H	K	COND	PSUM	SO2M	U850	V850	SO2C	PSUC	SUDE	MM4	TEMP	HUMI	USUR	VSUR	WSPE
MM1	1.000	-.362	-.509	-.473	-.044	-.085	-.057	-.064	-.341	-.228	-.143	-.077	.092	.325	-.320	-.183	.345	.575	.111	.285	.006	.130	.063
SO4C	-.362	1.000	.758	.832	.068	.139	.019	.536	.411	.567	.601	.238	-.275	-.028	.558	.552	-.110	-.264	.259	-.412	-.208	.139	.101
NH4	-.509	.758	1.000	.709	.211	.283	.149	.270	.599	.606	.406	.189	-.047	-.071	.477	.382	-.220	-.274	.104	-.259	-.161	.070	.108
NO3	-.473	.832	.709	1.000	-.087	.010	-.014	.534	.325	.473	.651	.360	-.409	-.036	.619	.616	-.118	-.308	.205	-.409	-.241	.022	.033
NA	-.044	.068	.211	-.087	1.000	.927	.462	-.285	.566	.590	-.120	-.069	.629	.197	-.225	-.345	-.079	.163	-.374	.184	.228	.172	.291
MG	-.085	.139	.283	.010	.927	1.000	.554	-.309	.631	.673	-.049	-.058	.573	.116	-.170	-.308	-.116	.122	-.284	.136	.112	.177	.263
CA	-.057	.019	.149	-.014	.462	.554	1.000	-.631	.415	.561	-.100	.027	.185	.210	-.142	-.313	-.114	-.005	-.319	.068	.000	-.036	.053
H	-.064	.536	.270	.534	-.285	-.309	-.631	1.000	-.105	.026	.457	.198	-.358	-.042	.473	.562	.122	-.128	.454	-.333	-.197	.143	-.008
K	-.341	.411	.599	.325	.566	.631	.415	-.105	1.000	.564	.040	-.045	.182	.112	.175	.040	-.315	-.132	-.017	-.084	-.070	.148	.119
COND	-.228	.567	.606	.473	.590	.673	.561	.026	.564	1.000	.365	.259	.225	.150	.203	.026	-.093	-.099	-.195	-.145	-.084	.068	.295
PSUM	-.143	.601	.406	.651	-.120	-.049	-.100	.457	.040	.365	1.000	.660	-.281	.097	.504	.522	.040	-.159	.075	-.313	-.158	.000	.067
SO2M	-.077	.238	.189	.360	-.069	-.058	.027	.198	-.045	.259	.660	1.000	-.215	.104	.409	.352	.168	-.155	-.276	-.113	-.237	-.171	.032
U850	.092	-.275	-.047	-.409	.629	.573	.185	-.358	.182	.225	-.281	-.215	1.000	.003	-.433	-.528	.039	.200	-.406	.369	.213	.029	.222
V850	.325	-.028	-.071	-.036	.197	.116	.210	-.042	.112	.150	.097	.104	.003	1.000	-.073	-.176	.199	.304	.024	.068	.127	.388	.079
SO2C	-.320	.558	.477	.619	-.225	-.170	-.142	.473	.175	.203	.504	.409	-.433	-.073	1.000	.840	.049	-.303	.126	-.459	-.275	-.124	.047
PSUC	-.183	.552	.382	.616	-.345	-.308	-.313	.562	.040	.026	.522	.352	-.528	-.176	.840	1.000	.056	-.183	.297	-.407	-.306	-.106	-.016
SUDE	.345	-.110	-.220	-.118	-.079	-.116	-.114	.122	-.315	-.093	.040	.168	.039	.199	.049	.056	1.000	.258	-.115	.331	.084	.075	.066
MM4	.575	-.264	-.274	-.308	.163	.122	-.005	-.128	-.132	-.099	-.159	-.155	.200	.304	-.303	-.183	.258	1.000	.025	.572	.314	.137	.092
TEMP	.111	.259	.104	.205	-.374	-.284	-.319	.454	-.017	-.195	.075	-.276	-.406	.024	.126	.297	-.115	.025	1.000	-.119	-.109	.392	-.205
HUMI	.285	-.412	-.259	-.409	.184	.136	.068	-.333	-.084	-.145	-.313	-.113	.369	.068	-.459	-.407	.331	.572	-.119	1.000	.336	.199	-.139
USUR	.006	-.208	-.161	-.241	.228	.112	.000	-.197	-.070	-.084	-.158	-.237	.213	.127	-.275	-.306	.084	.314	-.109	.336	1.000	.002	.294
VSUR	.130	.139	.070	.022	.172	.177	-.036	.143	.148	.068	.000	-.171	.029	.388	-.124	-.106	.075	.137	.392	.199	.002	1.000	.023
WSPE	.063	.101	.108	.033	.291	.263	.053	-.008	.119	.295	.067	.032	.222	.079	.047	-.016	.066	.092	-.205	-.139	.294	.023	1.000

Matrix of the principal component coefficients (the principal components are in columns in descending order of magnitude of the respective eigenvalues)

MM1	.471	-.262	.552	.091	-.186	-.173	.272	-.335	-.193	.082	-.017	-.032	.113	-.091	.103	.230	.062	-.048	-.073	-.064	.030	.012	.000
SO4C	-.830	.297	.182	-.162	.073	.048	-.025	-.014	-.055	.199	-.097	.030	-.027	.056	.135	.118	.029	-.031	.209	.105	.100	.006	.016
NH4	-.605	.499	-.012	-.171	.092	.187	-.104	-.027	.017	.053	.015	.290	.145	-.010	-.004	.148	-.229	-.010	-.104	-.008	-.03[illegible]	-.051	.009
NO3	-.883	.179	.098	-.036	.013	.079	-.159	.044	-.057	.125	-.063	.088	-.137	.045	.047	-.113	.095	-.168	-.085	-.165	.057	.044	-.025
NA	.243	.868	.125	-.017	.123	.037	.161	.005	-.004	-.173	-.089	-.199	-.117	-.011	.012	.041	-.072	-.107	-.014	.062	-.010	-.043	-.114
MG	.154	.906	.083	-.056	.030	.069	.168	-.067	-.033	-.052	-.010	-.239	-.075	-.018	.014	-.086	-.075	-.023	-.074	.003	.081	-.007	.118
CA	.165	.672	-.141	.141	-.489	-.114	-.246	-.188	.080	.264	.029	-.101	.044	.065	-.072	-.001	.041	-.010	.073	-.078	-.053	-.138	.001
H	-.643	-.335	.354	-.167	.286	.056	.265	.127	-.075	-.024	-.228	-.012	-.151	-.107	-.149	.044	.078	.054	.017	-.053	-.067	-.128	.026
K	-.256	.747	-.066	-.301	-.001	.063	-.052	-.170	.107	-.265	.029	.091	.042	-.297	.161	-.048	.112	.111	.044	-.052	-.037	.021	-.009
COND	-.355	.815	.146	.111	-.072	-.021	.058	-.021	-.083	.246	-.081	-.012	-.025	.015	-.199	.046	.075	.105	-.052	.059	-.075	.146	-.010
PSUM	-.701	.046	.320	.330	-.045	-.003	-.038	.226	-.328	.024	.056	-.118	.183	.071	.181	-.128	.043	.099	-.083	.063	-.024	-.065	-.018
SO2M	-.445	.030	.185	.689	-.223	.086	.014	.233	-.158	-.148	.216	-.038	-.011	-.195	-.111	.044	-.108	-.038	.113	-.070	.020	.036	.003
U850	.526	.525	-.005	.065	.293	.193	.355	.107	-.046	-.074	-.065	.114	.314	.170	-.056	-.061	.066	-.045	.093	-.088	-.006	.011	.002
V850	.133	.172	.608	-.008	-.398	-.391	-.169	.187	.119	-.272	-.246	.195	.053	.028	-.059	-.049	.018	-.080	-.007	.072	.003	-.001	.025
SO2C	-.805	-.070	.053	.185	.050	.010	-.050	-.256	.277	-.249	.047	-.088	.093	.133	-.147	.061	.062	.103	-.045	-.013	.146	-.023	-.025
PSUC	-.791	-.268	.125	.094	.091	.082	-.015	-.343	.110	-.187	.104	-.133	.036	.114	.034	.000	.018	-.140	.029	.040	-.169	.035	.036
SUDE	.149	-.204	.543	.399	.061	.240	.077	.000	.534	.255	-.164	-.045	.039	-.093	.091	-.093	-.079	.033	.004	-.013	-.009	.016	-.011
MM4	.478	-.012	.607	.005	.075	.149	-.180	-.387	-.280	-.087	-.006	.128	-.165	.148	-.009	-.109	-.108	.109	.064	-.043	-.002	.008	-.013
TEMP	-.281	-.372	.262	-.711	-.069	-.029	-.031	-.116	-.101	.127	.047	-.160	.224	-.158	-.156	-.153	-.077	-.050	.030	.017	.033	.016	-.026
HUMI	.609	.050	.324	.015	.023	.592	-.236	.029	-.008	.024	.207	.089	-.007	-.083	-.084	.027	.164	-.074	-.046	.105	.023	-.043	.010
USUR	.392	.118	.169	-.001	.544	-.127	-.612	.156	-.029	-.035	-.118	-.214	.099	-.034	.008	.121	.004	.011	.010	-.067	-.017	.029	.013
VSUR	.041	.119	.533	-.564	-.163	-.021	.101	.354	.207	-.008	.345	-.075	-.073	.170	.046	.084	-.005	.055	.008	-.068	-.028	.019	-.003
WSPE	.003	.331	.213	.203	.555	-.551	.070	-.107	.090	.139	.325	.157	-.034	-.078	-.030	-.080	.026	-.031	.002	.035	.010	-.036	-.001

Table B4: Analysis of R_o

The eigenvalues of R_o	of S_3
5.99	2.35
4.50	1.72
2.26	.09
1.88	.07
1.35	.05
1.06	.04
.94	.02
.86	.02
.74	.02
.61	.01
.52	.002
.43	.001
.34	0.00
.32	0.00
.24	0.00
.22	0.00
.18	0.00
.14	0.00
.12	0.00
.11	0.00
.09	0.00
.08	0.00
.03	0.00

lower bounds to the rank of R_o:
$s_1 = 6$
$s_2 = 12$

Number of observations: 95

The determinant of R_o: $|R_o| = 2.57E-10$

Test of the hypothesis H_o: $R_o = I$ $T_o = 1972.64^{**}$ D.F. = 253
χ^2_{253} (.95) = 289.89
(normal approximation of χ^2)

Test of the hypothesis H_2: $R=(1-\varrho)I\ \varrho ee'$
ϱ... the common value of correlation coefficient
H_2: only one non-trivial principal component

$T_{21} = 1948.10^{**}$ D.F. = 504
χ^2_{504} (.95) = 556.07
(normal approximation of χ^2)

**.. α = .01 significane level.

Table B5: Analysis of R_o:

Test of the hypothesis H_1: $\lambda_j^o = \lambda_{j+1}^o = .. = \lambda_p^o$ for various j.

j	p-j-1	$\sum_{i=j}^{p} l_j$	log Λ_1	m = -{n- (2p+5)/6 -2(j-1)/3}	T_1 = m log Λ_1	D.F.	critical value $\chi^2_{D.F.}$ (.95)	critical value $\chi^2_{D.F.}$ (.99)
22	2	.109	-.140	71.500	10.010 **	2	5.991	6.635
21	3	.199	-.218	72.167	15.732 **	5	11.071	15.086
20	4	.308	-.375	72.833	27.313 **	9	16.919	21.666
19	5	.429	-.478	73.500	35.133 **	14	23.685	29.141
18	6	.574	-.630	74.167	46.735 **	20	31.410	37.566
17	7	.751	-.797	74.833	59.642 **	27	40.113	46.963
16	8	.969	-1.099	75.500	82.975 **	35	49.802	57.342
15	9	1.210	-1.387	76.167	105.643 **	44	60.481	68.710
14	10	1.530	-2.900	76.833	222.816 **	56	74.468	83.513

** .. α = .01 significance level

Table B6: ANALYSIS OF R_1

Correlation Matrix R_1

	SO4	NH4	NO3	K	COND	PSUM	U850
SO4	1.00						
NH4	.76	1.00					
NO3	.83	.71	1.00				
K	.41	.60	.33	1.00			
COND	.57	.61	.47	.56	1.00		
PSUM	.60	.41	.65	.04	.36	1.00	
U850	-.26	-.47	-.41	.18	.23	-.28	1.00
HUMI	-.41	-.26	-.41	-.08	-.15	-.31	-.37

Eigenvalues of R_1 : 3.92 1.72 .77 .63 .34 .30 .18 .15
Eigenvalues of S_3 : 3.40 1.46 .51 .37 0.00 0.00 0.00 0.00

Number of observations: 95

Value of the determinant of R_1 : $|R|$ = 8.818E-3**

Critical values for test of $H_0: R_1 = I$ α= .05 crit.v. = .63

(Nagarsenker) α= .01 crit.v. = .59

Test of the hypothesis H_0 : $R_1 = I$ T_0 = 422.63** D.F. = 28

χ^2_{28} (.95) = 41.34

Test for the hypothesis H_2 : $R_1 = (1-\varrho)I + \varrho ee'$
ϱ... the common value of corr. coef.
H_2 : only one non-trivial principal component. T_{21} = 2739.67** D.F. = 27

χ^2_{27} (.95) = 40.11

Lower bounds to the rank of R_1 : $s_1 = 2$ $s_3 = 4$

Matrix of the principal component coefficients (in columns; the columns are arranged in descending order of magnitude of the respective eigenvalues).

	1.	2.	3.	4.	5.	6.	7.	8.
SO_4	.92	.06	.05	-.06	.15	-.10	-.25	.22
NH_4	.85	-.26	-.07	-.17	.30	.02	.28	.05
NO_3	.90	.22	.09	-.15	.09	-.05	-.11	-.31
K	.54	-.62	-.38	-.20	-.23	.28	-.07	.01
COND	.69	-.51	.12	.23	-.28	-.33	.07	-.01
PSUM	.68	.33	.52	.21	-.14	.30	.06	-.04
U850	-.26	-.81	.18	.39	.26	.11	-.10	-.06
HUMI	.49	-.45	.53	-.53	-.04	-.03	-.02	-.02

** ... α = .01 significance level.

Table B7: Analysis of R_1

Test of the hypothesis H_1: $\lambda_j^1 = \lambda_{j+1}^1 = .. = \lambda_p^1$ for various j.

j	p-j-1	$\sum_{i=j}^{p} l_j$	log Λ_1	m = - {m- (2p+5)/6 - 2(j-1)/3}	T_1 = m log Λ_1	D.F.	critical values $\chi^2_{D.F.}$ (.95)	$\chi^2_{D.F.}$ (.99)
7	2	.328	-.007	86.500	.640	2	5.991	6.635
6	3	.629	-.136	87.167	11.820 *	5	11.071	15.086
5	4	.969	-.231	87.833	20.298 *	9	16.919	21.666
4	5	1.594	-.656	88.500	58.038 **	14	23.685	29.141
3	6	2.360	-1.039	89.167	92.617 **	20	31.410	37.566

* ... α = .05 significance level
** ... α = .01 significance level

TABLE B8: Analysis of R_2

Correlation matrix R_2

	SO_4	NA	MG	CA	H	U850	V850
SO4	1.00						
NA	.07	1.00					
MG	.14	.93	1.00				
CA	.02	.46	.55	1.00			
H	.54	-.28	-.31	-.63	1.00		
U850	-.27	.63	.57	.18	-.36	1.00	
V850	-.03	.20	.12	.21	-.04	.00	1.00
HUMI	-.41	.18	.14	.07	-.33	.37	.07

Eigenvalues of R_2: 3.19 1.70 1.11 .95 .57 .28 .15 .06
Eigenvalues of S_3: 2.67 1.33 .93 .73 .34 0.00 0.00 0.00

Number of observations: 95

Value of the determinant of R_2: $|R_2|$ = 7.771E-3**

Critical values for test of H_0: R_2 = I: α =.05 crit.v. = .63
(Nagarsenker) α =.01 crit.v. = .59

Test of the hypothesis H_0: R_2=I T_0 = 439.58** D.F.= 28

$\chi^2_{28}(.95)$= 41.34

Test of the hypothesis H_2: $R_2=(1-\varrho)I+\varrho ee'$ T_{21} = 398.91** D.F.= 27
ϱ...the common value of correl. coef.

H_2: only one non-trivial principal component. $\chi_{27}(.95)$= 40.11

Lower bounds to the rank of R_2: s_1 = 3 , s_3 = 5.

Matrix of the principal component coefficients (in columns; the columns are arranged in descending order of magnitude of the respective eigenvelues).

	1.	2.	3.	4.	5.	6.	7.	8.
SO4	-.26	-.86	.04	-.04	-.33	-.23	-.15	-.01
NA	.85	-.39	.20	.10	.06	.20	-.05	-.16
MG	.84	-.45	.15	-.04	-.01	.19	.00	.17
CA	.68	-.17	-.56	-.31	-.18	-.11	.22	-.03
H	-.68	-.46	.37	.34	-.04	.06	.26	-.01
U850	.74	.11	.49	.09	.25	-.36	.05	.01
V850	.22	-.12	-.55	.79	.13	-.06	-.04	.02
HUMI	.43	.57	.23	.31	-.58	.01	.01	.00

* .... α= .05 significance level
** ... α= .01 significance level

Table B9: Analysis of R_2

Test of the hypothesis H_1: $\lambda^2_{j+1} = \ldots = \lambda^2_p$ for various j

j	p-j-1	$\sum_{i=j}^{p} l_j$	log Λ_1	m=-(n-(2n+5)/6 - 2(j-1)/2)	T_1 = m log Λ_1	D.F.	criticle value $\chi^2_{D.F.}$ (.95)	$\chi^2_{D.F.}$ (.99)
7	2	.205	-.208	86.500	17.992 **	2	5.991	9.210
6	3	.485	-.567	87.167	49.406 **	5	11.071	15.086
5	4	1.055	-1.265	87.833	111.117 **	9	16.919	21.666
4	5	2.004	-2.078	88.500	183.885 **	14	23.685	29.141
3	6	3.114	-2.610	89.167	232.726 **	20	31.410	37.566

* ... α = .05 significance level
** ... α = .01 significance level

TABLE B10: ANALYSIS OF R_3

Correlation matrix R_3

	SO4C	NA	CA	K	U850	V850	SUDE
SO4C	1.00						
NA	.07	1.00					
CA	.02	.46	1.00				
K	.41	.57	.41	1.00			
U850	-.27	.63	.18	.18	1.00		
V850	-.03	.20	.21	.11	.00	1.00	
SUDE	-.11	-.08	-.11	-.32	.04	.20	1.00
HUMI	-.41	.18	.07	-.08	.37	.07	.33

Eigenvalues of R_3 : 2.38 1.95 1.16 .84 .65 .53 .30 .21
Eigenvalues of S_3 : 2.06 1.76 1.05 .68 .44 .34 .05 0.00

Number of observations: 95
Value of the determinent of R_3: $|R3|$ = 9.690E-2**

Critical values for the test of $H_0: R_3 = I$ α= .05 crit.v. = .63
(Nagasenker) α= .01 crit.v. = .59

Test of hypothesis H_0 : $R_3 = I$ $T_0 = 211.2^{**}$ D.F. = 28

χ^2_{28} (.95) = 40.01

Test of hypothesis H_2: $R_3 = (1-\varrho)I+\varrho ee'$ $T_{21} = 298.28^{**}$ D.F. = 27
ϱ.. common value of the correl.coef.
H_2: only 1 nontrivial principal component χ^2_{27} (.95) = 40.11

Lower bounds to the rank of R_3: s_1 = 3, s_3 = 7

Matrix of the principal component coefficients (in columns; the columns are arranged in descending order of magnitude of the respective eigenvalues).

	1.	2.	3.	4.	5.	6.	7.	8.
SO4C	.08	-.75	.25	-.52	.08	-.03	-.30	-.04
NA	.90	.06	-.05	-.15	-.15	.10	-.00	.35
CA	.68	-.09	.17	.40	.51	.26	-.10	-.06
K	.73	-.48	.02	-.14	.06	-.28	.34	-.13
U850	.64	.47	-.36	-.19	-.29	.20	-.11	-.26
V850	.28	.13	.80	.30	-.39	-.14	-.08	-.05
SUDE	-.17	.57	.54	-.47	.18	.27	.19	-.02
HUMI	.23	.77	-.03	-.12	.30	-.48	-.15	.02

**... α = .01 significance level

Table B11: Analysis of R_3

Test of the hypothesis $H_1 : \lambda_j^3 = \lambda_{j+1}^3 = \ldots = \lambda_p^3$ for various j

j	p-j-1	$\sum_{i=j}^{p} l_j$	log Λ_1	m = - {m- (2p+5)/6 - 2(j-1)/3}	T_1 = m log Λ_1	D.F.	critical value $\chi^2_{D.F.}$ (.95)	critical value $\chi^2_{D.F.}$ (.99)
7	2	.511	-.027	86.500	2.335	2	5.991	6.635
6	3	1.040	-.214	87.167	18.675 **	5	11.071	15.086
5	4	1.686	-.374	87.833	32.867 **	9	16.919	21.666
4	5	2.525	-.586	88.500	51.852 **	14	23.685	29.141
3	6	3.681	-.934	89.167	83.282 **	20	31.410	37.566

** .. α = -01 significance level

Table 812:
Characteristics of data from Røros

	MEAN	STD.DEV.	VARIANCE	SKEWNESS	KURTOSIS	MAXIMUM	MINIMUM
MM4	1.47	2.92	8.55	3.03**	11.71**	20.90	0.0
PRES	1013	12.61	159.0	.20	.08	1048	972
PTEN	.03	.97	0.95	.08	4.27**	5.50	-4.48
TEMP	-.29	11.74	137.8	-.49**	-.41	20.1	-32.3
HUMI	76.64	12.83	164.5	-.61**	-.20	99	40.6
USUR	-.03	1.82	3.33	.36**	1.97**	8.28	-5.26
VSUR	.56	1.99	3.94	.84**	2.68**	8.27	-6.36
WSPE	1.94	1.95	3.81	1.22**	1.02**	8.85	0.00

* ... α = .10 significance level
** .. α = .02 significance level

TABLE B13: DATA FROM RØROS

Correlation matrix R:

	MM4	PRES	PTEN	TEMP	HUMI	USUR	USUR
MM4	1.00						
PRES	-.26	1.00					
PTEN	-.08	.04	1.00				
TEMP	.20	-.24	-.11	1.00			
HUMI	.35	-.30	.08	-.42	1.00		
USUR	.10	-.06	.22	-.01	.13	1.00	
VSUR	-.07	-.09	-.19	.17	-.21	-.38	1.00
WSPE	.03	-.31	.01	.31	-.18	.04	.45

Eigenvalues of R: 2.02 1.68 1.24 1.04 .73 .65 .37 .27
Eigenvalues of S_3: 1.85 1.47 1.08 .70 .34 .21 0.00 0.00

Number of observations: 365
Value of the determinant of R: $|R| = 2.070E-1$

Test of hypothesis H_0: R=I $\quad T_0 = 567.87^{**}$ D.F. = 28

χ^2_{28} (.95) = 41.34

Test of hypothesis H_2 : $R = (1-\varrho)\ I + \varrho ee'$
ϱ... the common value of correlation coeff.,
H_2: only one non-trivial principal component.

T_{21} = 1484.87** D.F. = 27

χ^2_{27} (.95) = 40.11

Lower bounds to the rank of R: $s_1 = 4$, $s_2 = 6$

Matrix of the principal component coefficients (in columns: the columns are arranged in descending order of magnitude of the respective eigenvelues).

	1.	2.	3.	4.	5.	6.	7.	8.
MM4	.05	.71	.19	.38	.32	.42	.06	-.17
PRES	.23	-.76	-.07	.18	.08	.53	.06	.20
PTEN	.33	.04	-.57	-.49	.57	-.01	-.06	-.02
TEMP	-.64	.24	-.36	.49	.21	-.16	-.13	.28
HUMI	.55	.56	.43	-.28	-.02	.04	-.06	.34
USUR	.39	.35	-.64	.07	-.47	.22	-.23	-.03
VSUR	-.74	-.07	.28	-.37	.02	.27	-.39	-.04
WSPE	-.67	.32	-.27	-.40	-.18	.22	.37	.08

* ... α = .05 significance level
** .. α = .01 -"- -"-

Table B14: Data from Røros

Test of the hypothesis $H_1 : \lambda_j = \lambda_{j+1} = .. = \lambda_p$ for various j

j	p-j-1	$\sum_{i=j}^{p} l_j$	log Λ_1	m = - {m- (2p+5)/6 - 2(j-1)/3}	T_1 = m log Λ_1	D.F.	critical value $\chi^2_{D.F.}$ (.95)	$\chi^2_{D.F.}$ (.99)
7	2	.641	-.025	356.500	9.017 **	2	5.991	6.635
6	3	1.290	-.201	357.167	71.955 **	5	11.071	15.086
5	4	2.018	-.314	357.833	112.322 **	9	16.919	21.666
4	5	3.056	-.552	358.500	197.819 **	14	23.685	29.141
3	6	4.295	-.793	359.169	284.952 **	20	31.410	37.566

**.. α = .01 significance level

TABLE B15:
Characteristics of data from Tynset

	MEAN	STD.DEV	VARIANCE	SKEWNESS	KURTOSIS	MAXIMUM	MINIMUM
MM4	1.18	2.50	6.23	3.15**	12.14**	19.00	0.0
PRES	1014	12.55	157.4	.21*	.22	1049	973
PTEN	.05	1.08	1.17	-.42**	1.69**	2.93	-4.64
TEMP	-.65	12.97	168.3	-.60**	-.38	21.03	-35.05
HUMI	78.4	12.33	152.3	-.56**	-.47	98.64	41.42
USUR	.05	1.28	1.64	1.39**	6.58**	7.64	-3.33
VSUR	.71	1.37	1.89	1.41**	6.71**	9.96	-4.69
WSPE	1.43	1.40	1.96	2.40**	9.99**	11.81	0.00

* ... α = .10 significance level
** .. α = .02 significance level

TABLE B16: DATA FROM TYNSET

Correlation matrix R:

	MM4	PRES	PTEN	TEMP	HUMI	USUR	VSUR
MM4	1.00						
PRES	-.29	1.00					
PTEN	-.05	.08	1.00				
TEMP	.25	-.27	-.06	1.00			
HUMI	.27	-.19	-.07	-.45	1.00		
USUR	.02	-.11	.03	.01	.00	1.00	
VSUR	-.02	-.07	-.22	.09	-.05	.02	1.00
WSPE	.05	-.20	-.14	.18	-.11	.26	.63

Eigenvalues of R: 1.97 1.50 1.33 1.09 .84 .68 .31 .28
Eigenvalues of S3: 1.76 1.42 1.18 1.02 .73 .64 .24 .00

Number of observations: 365
Value of the determinant of R: $|R| = 2.131E-1$

Test of hypothesis H_0: R=I $\quad T_0 = 557.28^{**}$ D.F. = 28

χ^2_{28} (.95) = 41.34

Test of hypothesis H_2: $R = (1-\varrho)I + \varrho ee'$
ϱ... the common value of corr. coeff. $T_{21} = 1467.09^{**}$ D.F.=27
H_2: only one non-trivial principal component $\quad \chi^2_{27}$ (.95) = 40.11

Lower bounds to the rank of R: $s_1 = 4$, $s_2 = 7$

Matrix of the principal component coefficients (in columns: the columns are arranged in descending order of magnitude of the respective eigenvalues).

	1.	2.	3.	4.	5.	6.	7.	8.
MM4	-.29	.67	-.34	.11	.13	-.53	-.15	-.12
PRES	.48	-.54	.27	.07	-.02	-.61	.10	.11
PTEN	.35	-.11	-.29	-.51	.72	.03	.05	-.01
TEMP	-.53	-.18	-.73	.14	-.06	-.03	.30	.17
HUMI	.17	.80	.45	.00	.10	.02	.27	.21
USUR	-.26	.06	.06	-.85	-.40	-.14	.11	-.10
VSUR	-.70	-.21	.47	.18	.30	-.04	.21	-.28
WSPE	-.81	-.16	.31	-.20	.18	-.05	-.22	.31

** .. α = .01 significance level

Table B17: Data from Tynset
Test of the hypothesis $H_1 : \lambda_j = \lambda_{j+1} = .. = \lambda_p$ for various j

j	p-j-1	$\sum_{i=j}^{p} l_j$	log Λ_1	m = - {m- (2p+5)/6 - 2(j-1)/3}	T_1 = m log Λ_1	D.F.	critical value $\chi^2_{D.F.}$ (.95)	critical value $\chi^2_{D.F.}$ (.99)
7	2	.590	-.003	356.500	.906	2	5.991	6.635
6	3	1.274	-.256	357.167	91.291 **	5	11.071	15.086
5	4	2.111	-.446	357.833	159.576 **	9	16.919	21.666
4	5	3.203	-.688	358.500	246.648 **	14	23.685	29.141
3	6	4.536	-.948	359.167	340.499 **	20	31.410	37.566

**.. α = .01 significance level

APPENDIX C

Tables of the distribution of the determinant $|R|$ of sample correlation matrix

SOURCE: NAGARSENKER (18)

Table C1

1% Points of $|R|$

N \ P	3	4	5	6	7	8
4	$.0^{4}37877$					
5	$.0^{2}51559$	$.0^{5}83414$				
6	.027706	$.0^{2}15806$	$.0^{5}24479$			
7	.065713	.010309	$.0^{3}51475$	$.0^{6}79738$		
8	.11136	.028229	$.0^{2}39086$	$.0^{3}17298$	$.0^{6}23717$	
9	.15898	.053352	.012072	$.0^{2}14901$	$.0^{4}59233$	$.0^{7}69806$
10	.20548	.082925	.025106	$.0^{2}51096$	$.0^{3}56880$	$.0^{4}20544$
11	.24940	.11470	.042152	.011580	$.0^{2}21402$	$.0^{3}21704$
12	.29018	.14713	.062115	.020879	$.0^{2}52458$	$.0^{3}88792$
13	.32772	.17922	.083991	.032659	.010106	$.0^{3}23389$
14	.36214	.21038	.10697	.046453	.016726	$.0^{2}47941$
15	.39366	.24026	.13044	.061784	.024976	$.0^{2}83719$
16	.42254	.26870	.15395	.078223	.034651	.013099
17	.44902	.29564	.17719	.095402	.045521	.018933
18	.47336	.32109	.19994	.11302	.057356	.025788
19	.49576	.34508	.22207	.13085	.069945	.033555
20	.51644	.36769	.24348	.14870	.083097	.042115
25	.59944	.46263	.33884	.23436	.15243	.092832
30	.65864	.53421	.41592	.30979	.22016	.14891
35	.70280	.58956	.47824	.37417	.28185	.20406
40	.73695	.63344	.52919	.42883	.33661	.25563
45	.76411	.66899	.57142	.47539	.38480	.30277
50	.78622	.69834	.60690	.51534	.42718	.34543
60	.82000	.74389	.66303	.57999	.49763	.41857
70	.84459	.77757	.70532	.62981	.55336	.47821
80	.86328	.80346	.73827	.66925	.59832	.52737
90	.87796	.82397	.76464	.70120	.63524	.56839
100	.88980	.84061	.78620	.72758	.66606	.60306

Table C2

5% Points of $|R|$

N \ P	3	4	5	6	7	8
4	$.0^{2}10183$					
5	.026873	$.0^{3}24025$				
6	.084781	$.0^{2}86215$	$.0^{4}66687$			
7	.15341	.033427	$.0^{2}29095$	$.0^{4}20033$		
8	.22052	.070026	.013279	$.0^{2}10067$	$.0^{5}61052$	
9	.28177	.11213	.031479	$.0^{2}52641$	$.0^{3}35344$	$.0^{5}19002$
10	.33621	.15558	.055468	.013903	$.0^{2}20781$	$.0^{3}12528$
11	.38420	.19807	.083017	.026714	$.0^{2}60434$	$.0^{3}81667$
12	.42644	.23847	.11237	.042905	.012570	$.0^{2}25909$
13	.46372	.27630	.14231	.061570	.021567	$.0^{2}57984$
14	.49674	.31143	.17202	.081894	.032721	.010585
15	.52614	.34393	.20101	.10321	.045639	.016936
16	.55243	.37393	.22896	.12502	.059925	.024731
17	.57605	.40162	.25572	.14692	.075219	.033798
18	.59738	.42720	.28121	.16866	.091212	.043943
19	.61672	.45087	.30542	.19002	.10765	.054971
20	.63433	.47278	.32837	.21089	.12431	.066703
25	.70291	.56146	.42600	.30562	.20651	.13089
30	.75003	.62530	.50067	.38381	.28099	.19598
35	.78432	.67320	.55892	.44781	.34564	.25658
40	.81038	.71038	.60536	.50057	.40112	.31112
45	.83083	.74003	.64315	.54456	.44876	.35959
50	.84732	.76421	.67444	.58167	.48986	.40253
60	.87223	.80125	.72317	.64066	.55677	.47444
70	.89017	.82827	.75930	.68529	.60863	.53174
80	.90369	.84883	.78714	.72017	.64986	.57820
90	.91425	.86501	.80923	.74815	.68335	.61650
100	.92272	.87806	.82718	.77108	.71106	.64856

APPENDIX_D

Tables of the Chi-square distribution

SOURCE: OWEN, D.B.: HANDBOOK OF STATISTICAL TABLES, Addison - Wesley, Reading 1962

The value tabled is $\chi^2_f(\gamma)$.

$\chi^2_f(\gamma)$: Prob $\{\chi^2$ r.v. with f degrees of freedom $<$ tabled value$\} = \gamma$

Significance level $\alpha = 1-\gamma$

Table D1

Critical Values for the Chi-Square Distribution

$\Pr\{\chi^2$ r.v. with f degrees of freedom $\leq$ tabled value$\} = \gamma$

	γ					
f	0.005	0.01	0.025	0.05	0.10	0.25
1	-	-	0.001	0.004	0.016	0.102
2	0.010	0.020	0.051	0.103	0.211	0.575
3	0.072	0.115	0.216	0.352	0.584	1.213
4	0.207	0.297	0.484	0.711	1.064	1.923
5	0.412	0.554	0.831	1.145	1.610	2.675
6	0.676	0.872	1.237	1.635	2.204	3.455
7	0.989	1.239	1.690	2.167	2.833	4.255
8	1.344	1.646	2.180	2.733	3.490	5.071
9	1.735	2.088	2.700	3.325	4.168	5.899
10	2.156	2.558	3.247	3.940	4.865	6.737
11	2.603	3.053	3.816	4.575	5.578	7.584
12	3.074	3.571	4.404	5.226	6.304	8.438
13	3.565	4.107	5.009	5.892	7.042	9.299
14	4.075	4.660	5.629	6.571	7.790	10.165
15	4.601	5.229	6.262	7.261	8.547	11.037
16	5.142	5.812	6.908	7.962	9.312	11.912
17	5.697	6.408	7.564	8.672	10.085	12.792
18	6.265	7.015	8.231	9.390	10.865	13.675
19	6.844	7.633	8.907	10.117	11.651	14.562
20	7.434	8.260	9.591	10.851	12.443	15.452
21	8.034	8.897	10.283	11.591	13.240	16.344
22	8.643	9.542	10.982	12.338	14.042	17.240
23	9.260	10.196	11.689	13.091	14.848	18.137
24	9.886	10.856	12.401	13.848	15.659	19.037
25	10.520	11.524	13.120	14.611	16.473	19.939
26	11.160	12.198	13.844	15.379	17.292	20.843
27	11.808	12.879	14.573	16.151	18.114	21.749
28	12.461	13.565	15.308	16.928	18.939	22.657
29	13.121	14.257	16.047	17.708	19.768	23.567
30	13.787	14.954	16.791	18.493	20.599	24.478
31	14.458	15.655	17.539	19.281	21.434	25.390
32	15.134	16.362	18.291	20.072	22.271	26.304
33	15.815	17.074	19.047	20.867	23.110	27.219
34	16.501	17.789	19.806	21.664	23.952	28.136
35	17.192	18.509	20.569	22.465	24.797	29.054
36	17.887	19.233	21.336	23.269	25.643	29.973
37	18.586	19.960	22.106	24.075	26.492	30.893
38	19.289	20.691	22.878	24.884	27.343	31.815
39	19.996	21.426	23.654	25.695	28.196	32.737
40	20.707	22.164	24.433	26.509	29.051	33.660
41	21.421	22.906	25.215	27.326	29.907	34.585
42	22.138	23.650	25.999	28.144	30.765	35.510
43	22.859	24.398	26.785	28.965	31.625	36.436
44	23.584	25.148	27.575	29.787	32.487	37.363
45	24.311	25.901	28.366	30.612	33.350	38.291

Table D1 cont.

			γ			
f	0.75	0.90	0.95	0.975	0.99	0.995
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750
6	7.841	10.645	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.757
12	14.845	18.549	21.026	23.337	26.217	28.299
13	15.984	19.812	22.362	24.736	27.688	29.819
14	17.117	21.064	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801
16	19.369	23.542	26.296	28.845	32.000	34.267
17	20.489	24.769	27.587	30.191	33.409	35.718
18	21.605	25.989	28.869	31.526	34.805	37.156
19	22.718	27.204	30.144	32.852	36.191	38.582
20	23.828	28.412	31.410	34.170	37.566	39.997
21	24.935	29.615	32.671	35.479	38.932	41.401
22	26.039	30.813	33.924	36.781	40.289	42.796
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.559
25	29.339	34.382	37.652	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.194	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.993
29	33.711	39.087	42.557	45.722	49.588	52.336
30	34.800	40.256	43.773	46.979	50.892	53.672
31	35.887	41.422	44.985	48.232	52.191	55.003
32	36.973	42.585	46.194	49.480	53.486	56.328
33	38.058	43.745	47.400	50.725	54.776	57.648
34	39.141	44.903	48.602	51.966	56.061	58.964
35	40.223	46.059	49.802	53.203	57.342	60.275
36	41.304	47.212	50.998	54.437	58.619	61.581
37	42.383	48.363	52.192	55.668	59.892	62.883
38	43.462	49.513	53.384	56.896	61.162	64.181
39	44.539	50.660	54.572	58.120	62.428	65.476
40	45.616	51.805	55.758	59.342	63.691	66.766
41	46.692	52.949	56.942	60.561	64.950	68.053
42	47.766	54.090	58.124	61.777	66.206	69.336
43	48.840	55.230	59.304	62.990	67.459	70.616
44	49.913	56.369	60.481	64.201	68.710	71.893
45	50.985	57.505	61.656	65.410	69.957	73.166

Table 02

Critical Values for the Chi-Square Distribution

γ

f	0.005	0.01	0.025	0.05	0.10	0.25
46	25.041	26.657	29.160	31.439	34.215	39.220
47	25.775	27.416	29.956	32.268	35.081	40.149
48	26.511	28.177	30.755	33.098	35.949	41.079
49	27.249	28.941	31.555	33.930	36.818	42.010
50	27.991	29.707	32.357	34.764	37.689	42.942
51	28.735	30.475	33.162	35.600	38.560	43.874
52	29.481	31.246	33.968	36.437	39.433	44.808
53	30.230	32.018	34.776	37.276	40.308	45.741
54	30.981	32.793	35.586	38.116	41.183	46.676
55	31.735	33.570	36.398	38.958	42.060	47.610
56	32.490	34.350	37.212	39.801	42.937	48.546
57	33.248	35.131	38.027	40.646	43.816	49.482
58	34.008	35.913	38.844	41.492	44.696	50.419
59	34.770	36.698	39.662	42.339	45.577	51.356
60	35.534	37.485	40.482	43.188	46.459	52.294
61	36.300	38.273	41.303	44.038	47.342	53.232
62	37.068	39.063	42.126	44.889	48.226	54.171
63	37.838	39.855	42.950	45.741	49.111	55.110
64	38.610	40.649	43.776	46.595	49.996	56.050
65	39.383	41.444	44.603	47.450	50.883	56.990
66	40.158	42.240	45.431	48.305	51.770	57.931
67	40.935	43.038	46.261	49.162	52.659	58.872
68	41.713	43.838	47.092	50.020	53.548	59.814
69	42.494	44.639	47.924	50.879	54.438	60.756
70	43.275	45.442	48.758	51.739	55.329	61.698
71	44.058	46.246	49.592	52.600	56.221	62.641
72	44.843	47.051	50.428	53.462	57.113	63.585
73	45.629	47.858	51.265	54.325	58.006	64.528
74	46.417	48.666	52.103	55.189	58.900	65.472
75	47.206	49.475	52.942	56.054	59.795	66.417
76	47.997	50.286	53.782	56.920	60.690	67.362
77	48.788	51.097	54.623	57.786	61.586	68.307
78	49.582	51.910	55.466	58.654	62.483	69.252
79	50.376	52.725	56.309	59.522	63.380	70.198
80	51.172	53.540	57.153	60.391	64.278	71.145
81	51.969	54.357	57.998	61.261	65.176	72.091
82	52.767	55.174	58.845	62.132	66.076	73.038
83	53.567	55.993	59.692	63.004	66.976	73.985
84	54.368	56.813	60.540	63.876	67.876	74.933
85	55.170	57.634	61.389	64.749	68.777	75.881
86	55.973	58.456	62.239	65.623	69.679	76.829
87	56.777	59.279	63.089	66.498	70.581	77.777
88	57.582	60.103	63.941	67.373	71.484	78.726
89	58.389	60.928	64.793	68.249	72.387	79.675
90	59.196	61.754	65.647	69.126	73.291	80.625

Table D2 cont.

f	0.75	0.90	0.95	0.975	0.99	0.995
46	52.056	58.641	62.830	66.617	71.201	74.437
47	53.127	59.774	64.001	67.821	72.443	75.704
48	54.196	60.907	65.171	69.023	73.683	76.969
49	55.265	62.038	66.339	70.222	74.919	78.231
50	56.334	63.167	67.505	71.420	76.154	79.490
51	57.401	64.295	68.669	72.616	77.386	80.747
52	58.468	65.422	69.832	73.810	78.616	82.001
53	59.534	66.548	70.993	75.002	79.843	83.253
54	60.600	67.673	72.153	76.192	81.069	84.502
55	61.665	68.796	73.311	77.380	82.292	85.749
56	62.729	69.919	74.468	78.567	83.513	86.994
57	63.793	71.040	75.624	79.752	84.733	88.236
58	64.857	72.160	76.778	80.936	85.950	89.477
59	65.919	73.279	77.931	82.117	87.166	90.715
60	66.981	74.397	79.082	83.298	88.379	91.952
61	68.043	75.514	80.232	84.476	89.591	93.186
62	69.104	76.630	81.381	85.654	90.802	94.419
63	70.165	77.745	82.529	86.830	92.010	95.649
64	71.225	78.860	83.675	88.004	93.217	96.878
65	72.285	79.973	84.821	89.177	94.422	98.105
66	73.344	81.085	85.965	90.349	95.626	99.330
67	74.403	82.197	87.108	91.519	96.828	100.554
68	75.461	83.308	88.250	92.689	98.028	101.776
69	76.519	84.418	89.391	93.856	99.228	102.996
70	77.577	85.527	90.531	95.023	100.425	104.215
71	78.634	86.635	91.670	96.189	101.621	105.432
72	79.690	87.743	92.808	97.353	102.816	106.648
73	80.747	88.850	93.945	98.516	104.010	107.862
74	81.803	89.956	95.081	99.678	105.202	109.074
75	82.858	91.061	96.217	100.839	106.393	110.286
76	83.913	92.166	97.351	101.999	107.583	111.495
77	84.968	93.270	98.484	103.158	108.771	112.704
78	86.022	94.374	99.617	104.316	109.958	113.911
79	87.077	95.476	100.749	105.473	111.144	115.117
80	88.130	96.578	101.879	106.629	112.329	116.321
81	89.184	97.680	103.010	107.783	113.512	117.524
82	90.237	98.780	104.139	108.937	114.695	118.726
83	91.289	99.880	105.267	110.090	115.876	119.927
84	92.342	100.980	106.395	111.242	117.057	121.126
85	93.394	102.079	107.522	112.393	118.236	122.325
86	94.446	103.177	108.648	113.544	119.414	123.522
87	95.497	104.275	109.773	114.693	120.591	124.718
88	96.548	105.372	110.898	115.841	121.767	125.913
89	97.599	106.469	112.022	116.989	122.942	127.106
90	98.650	107.565	113.145	118.136	124.116	128.299

Table 03

Critical Values for the Chi-Square Distribution

γ

f	0.005	0.01	0.025	0.05	0.10	0.25
91	60.005	62.581	66.501	70.003	74.196	81.574
92	60.815	63.409	67.356	70.882	75.100	82.524
93	61.625	64.238	68.211	71.760	76.006	83.474
94	62.437	65.068	69.068	72.640	76.912	84.425
95	63.250	65.898	69.925	73.520	77.818	85.376
96	64.063	66.730	70.783	74.401	78.725	86.327
97	64.878	67.562	71.642	75.282	79.633	87.278
98	65.694	68.396	72.501	76.164	80.541	88.229
99	66.510	69.230	73.361	77.046	81.449	89.181
100	67.328	70.065	74.222	77.929	82.358	90.133
102	68.965	71.737	75.946	79.697	84.177	92.038
104	70.606	73.413	77.672	81.468	85.998	93.944
106	72.251	75.092	79.401	83.240	87.821	95.850
108	73.899	76.774	81.133	85.015	89.645	97.758
110	75.550	78.458	82.867	86.792	91.471	99.666
112	77.204	80.146	84.604	88.570	93.299	101.575
114	78.862	81.836	86.342	90.351	95.128	103.485
116	80.522	83.529	88.084	92.134	96.958	105.396
118	82.185	85.225	89.827	93.918	98.790	107.307
120	83.852	86.923	91.573	95.705	100.624	109.220
122	85.520	88.624	93.320	97.493	102.458	111.133
124	87.192	90.327	95.070	99.283	104.295	113.046
126	88.866	92.033	96.822	101.074	106.132	114.961
128	90.543	93.741	98.576	102.867	107.971	116.876
130	92.222	95.451	100.331	104.662	109.811	118.792
132	93.904	97.163	102.089	106.459	111.652	120.708
134	95.588	98.878	103.848	108.257	113.495	122.625
136	97.275	100.595	105.609	110.056	115.338	124.543
138	98.964	102.314	107.372	111.857	117.183	126.461
140	100.655	104.034	109.137	113.659	119.029	128.380
142	102.348	105.757	110.903	115.463	120.876	130.299
144	104.044	107.482	112.671	117.268	122.724	132.219
146	105.741	109.209	114.441	119.075	124.574	134.140
148	107.441	110.937	116.212	120.883	126.424	136.061
150	109.142	112.668	117.985	122.692	128.275	137.983
200	152.241	156.432	162.728	168.279	174.835	186.172
250	196.161	200.939	208.098	214.392	221.806	234.577
300	240.663	245.972	253.912	260.878	269.068	283.135
400	330.903	337.155	346.482	354.641	364.207	380.577
500	422.303	429.388	439.936	449.147	459.926	478.323
600	514.529	522.365	534.019	544.180	556.056	576.286
700	607.380	615.907	628.577	639.613	652.497	674.413
800	700.725	709.897	723.513	735.362	749.185	772.669
900	794.475	804.252	818.756	831.370	846.075	871.032
1000	888.564	898.912	914.257	927.594	943.133	969.484

Table D3 cont.

f	γ 0.75	0.90	0.95	0.975	0.99	0.995
91	99.700	108.661	114.268	119.282	125.289	129.491
92	100.750	109.756	115.390	120.427	126.462	130.681
93	101.800	110.850	116.511	121.571	127.633	131.871
94	102.850	111.944	117.632	122.715	128.803	133.059
95	103.899	113.038	118.752	123.858	129.973	134.247
96	104.948	114.131	119.871	125.000	131.141	135.433
97	105.997	115.223	120.990	126.141	132.309	136.619
98	107.045	116.315	122.108	127.282	133.476	137.803
99	108.093	117.407	123.225	128.422	134.642	138.987
100	109.141	118.498	124.342	129.561	135.807	140.169
102	111.236	120.679	126.574	131.838	138.134	142.532
104	113.331	122.858	128.804	134.111	140.459	144.891
106	115.424	125.035	131.031	136.382	142.780	147.247
108	117.517	127.211	133.257	138.651	145.099	149.599
110	119.608	129.385	135.480	140.917	147.414	151.948
112	121.699	131.558	137.701	143.180	149.727	154.294
114	123.789	133.729	139.921	145.441	152.037	156.637
116	125.878	135.898	142.138	147.700	154.344	158.977
118	127.967	138.066	144.354	149.957	156.648	161.314
120	130.055	140.233	146.567	152.211	158.950	163.648
122	132.142	142.398	148.779	154.464	161.250	165.980
124	134.228	144.562	150.989	156.714	163.546	168.308
126	136.313	146.724	153.198	158.962	165.841	170.634
128	138.398	148.885	155.405	161.209	168.133	172.957
130	140.482	151.045	157.610	163.453	170.423	175.278
132	142.566	153.204	159.814	165.696	172.711	177.597
134	144.649	155.361	162.016	167.936	174.996	179.913
136	146.731	157.518	164.216	170.175	177.280	182.226
138	148.813	159.673	166.415	172.412	179.561	184.538
140	150.894	161.827	168.613	174.648	181.840	186.847
142	152.975	163.980	170.809	176.882	184.118	189.154
144	155.055	166.132	173.004	179.114	186.393	191.458
146	157.134	168.283	175.198	181.344	188.666	193.761
148	159.213	170.432	177.390	183.573	190.938	196.062
150	161.291	172.581	179.581	185.800	193.208	198.360
200	213.102	226.021	233.994	241.058	249.445	255.264
250	264.697	279.050	287.882	295.689	304.940	311.346
300	316.138	331.789	341.395	349.874	359.906	366.844
400	418.697	436.649	447.632	457.305	468.724	476.606
500	520.950	540.930	553.127	563.852	576.493	585.207
600	622.988	644.800	658.094	669.769	683.516	692.982
700	724.861	748.359	762.661	775.211	789.974	800.131
800	826.604	851.671	866.911	880.275	895.984	906.786
900	928.241	954.782	970.904	985.032	1001.630	1013.036
1000	1029.790	1057.724	1074.679	1089.531	1106.969	1118.948

NORSK INSTITUTT FOR LUFTFORSKNING

(NORGES TEKNISK-NATURVITENSKAPELIGE FORSKNINGSRÅD)
POSTBOKS 130, 2001 LILLESTRØM
ELVEGT. 52.

TLF. (02) 71 41 70

RAPPORTTYPE	RAPPORT NR.	
Oppdragsrapport	OR 16/84	ISBN--82-7247-479-4
DATO APRIL 1984	ANSV.SIGN. O.F.Skogvold	ANT. SIDER 66
TITTEL Tests of hypotheses in principal component analysis.		PROSJEKTLEDER J. Schaug NILU PROSJEKT NR. E-8414
FORFATTER(E) Alena Moldanova		TILGJENGELIGHET** A OPPDRAGSGIVERS REF.
OPPDRAGSGIVER		
3 STIKKORD (á maks. 20 anslag) Principal components	Statistical analysis	Test of hypotheses
REFERAT (maks. 300 anslag, 5-10 linjer)		
TITLE Tests of hypotheses in the principal component analysis		

ABSTRACT (max. 300 characters, 5-10 lines.

The principal component analysis (PCA) has been described and its applications have been indicated. The purpose was to present a self-contained account of the tests of statistical hypotheses in the PCA to enable their propper use. Several examples are included as well as a review of literature.

**Kategorier:	Åpen - kan bestilles fra NILU	A
	Må bestilles gjennom oppdragsgiver	B
	Kan ikke utleveres	C