Multivariate Statistical Methods in Atmospheric Science

Ian Jolliffe

University of Aberdeen

itj@maths.abdn.ac.uk

Abstract

Principal component analysis (PCA) is widely used in atmospheric science where it is commonly known as empirical orthogonal function (EOF) analysis. This article describes a number of methods which extend or modify PCA in three main ways, namely to simplify and hence improve interpretation, to consider more than one group of variables, or to consider more than two 'modes'. Because of lack of space, the techniques are presented in outline only, and no examples are given. However, references are provided where further details and applications may be found.

1 PCA/EOF analysis

Principal component analysis (PCA) is possibly the most widely used of all multivariate statistical techniques (Jolliffe, 2002), and it has been a valuable tool in atmospheric sciences since the development of electronic computers made its computations possible for problems of non-trivial size (Preisendorfer and Mobley, 1988). The technique has a number of different but equivalent definitions. The most common, and probably the simplest, is as follows.

Suppose that we have measurements on $\mathbf{x}$, a vector of p variables with covariance matrix $\mathbf{S}$. The kth principal component is defined as the linear function $y_k = \mathbf{a}_k'\mathbf{x}, k = 1, 2, \ldots, p$, whose sample variance $\mathbf{a}_k'\mathbf{S}\mathbf{a}_k$ is maximized, subject to $\mathbf{a}_k'\mathbf{a}_k = 1$ and (for $k > 1$) $\mathbf{a}_h'\mathbf{a}_k = 0, h < k$. It turns out that

- The vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ are the eigenvectors of $\mathbf{S}$ corresponding to the largest, second largest, ... , smallest eigenvalues respectively. In atmospheric science these eigenvectors are commonly known as empirical orthogonal functions (EOFs).

- The kth largest eigenvalue of $\mathbf{S}$ is equal to the sample variance of the kth PC, y_k.

In atmospheric science the most frequent, but by no means only, context for PCA is where the p variables correspond to measurements of a meteorological variable at p different spatial locations. In these circumstances the EOFs are often presented as contours on a map. In practice, it is common to work with the correlation, rather than covariance, matrix, corresponding to using variables in $\mathbf{x}$ that are standardised to each have unit variance.

The basic PCA/EOF technique has been extended or modified in many ways. In the three sections that follow we outline some modifications which lead to simplification (Section 2), which deal with relationships between two groups of variables (Section 3) and which look at more than two 'modes' in a data set(Section 4). The final section (5) contains a few concluding remarks.

2 Simplification

EOFs are often difficult to interpret (though probably less so in atmospheric science than in many other disciplines). To aid in interpretation, various approaches to simplification have been proposed, including the following:

- Rotation - orthogonal or oblique
- Restriction of loadings to a discrete set of values
- Combining variance maximization and simplification criteria
- LASSO-based approach
- Truncation of loadings
- Empirical orthogonal teleconnections

2.1 Rotation

Recall that EOFs are the vectors of loadings or coefficients (we use the two words synonomously here) defining the linear combinations of $\mathbf{x}$ that are the PCs. For simple interpretations we would like the loadings all to be large or small with no intermediate values, or to take only a small number of distinct values. The idea of using rotation of the EOFs to simplify their interpretion is well-known and widely used, but controversial (Richman, 1986, 1987; Jolliffe, 1987, 1995; Mestas-Nuñez 2000.) Among the questions that need to be addressed if rotation is used are whether it should be orthogonal or oblique, which of the large number of simplicity criteria should be used, how many EOFs to rotate, and the choice of normalization constraint.

2.2 Discrete values for loadings

Rotation attempts to simplify loadings by making them either small or large. An alternative simplification strategy limits the number of available values for the loadings to a small set. The idea dates back to Hausmann (1982) who allowed only -1, 0, +1. 'Simple' components defined by Vines (2000) – see also Jolliffe *et al.* (2002) – allow more integers. Chipman and Gu (2001) find ordinary EOFs, and then truncate loadings to -1, 0, +1 or to -$c_1, 0, +c_2$ for two chosen constants c_1, c_2.

2.3 Combining variance maximization and simplification

Jolliffe and Uddin (2000) successively maximize a criterion which combines variance and simplicity in their Simplified Component Technique (SCoT). Kiers (1993) goes further and suggests techniques that concentrate on maximizing simplicity, largely ignoring variance maximization.

2.4 A LASSO-based approach

The LASSO (Least Absolute Shrinkage and Selection Operator) approach was developed in multiple regression analysis as a means of dealing with multicollinearity (Tibshirani, 1996). It is a compromise between variable selection and biased regression. It shrinks some regression coefficients exactly to zero. To achieve this an extra constraint is added to the usual least squares regression procedure. The constraint places an upper bound on the sum of absolute values of the regression coefficients.

Jolliffe and Uddin (2002) adapt the idea to PCA. Adding an extra constraint $\sum_{j=1}^{p} |a_{kj}| < t$ to the usual PCA optimization problem, where a_{kj} is the jth coefficient (loading) in the kth EOF and t is a tuning parameter, drives some coefficients a_{kj} to zero as t decreases.

2.5 Truncation of Loadings

Setting to zero all loadings whose absolute value is less than some threshold is a fairly common informal procedure; Chipman and Gu (2001) and Richman and Gong (1999) give more formal procedures for doing so. Truncation should only be done with great caution - Cadima and Jolliffe (1995, 2001) demonstrate that those variables with the smallest loadings are not necessarily the least important in representing a component.

2.6 Empirical Orthogonal Teleconnections

The patterns in empirical orthogonal teleconnections (van den Dool *et al.*, 2000) are defined by regressions of the measured variable at each spatial location on its value at a single chosen location. The technique chooses such locations successively, based on how well the location can explain (residual) variation at all other locations. A complicating factor is that 'correlations' and 'regressions' are calculated on uncentred data, making the results of the analysis more difficult to interpret in statistical terms.

3 Relationships between variables in two (or more) groups

Sometimes we wish to relate two or more sets of variables, for example spatial fields of sea surface temperature and mean sea level pressure. Again a variety of techniques is available, including:

- Canonical Correlation Analysis
- Maximum Covariance Analysis
- Redundancy Analysis
- Principal Predictors
- Reduced Rank Regression
- PCA of Instrumental Variables

As we shall see, there are connections between several of these techniques and others.

3.1 Canonical Correlation Analysis (CCA)

This is the oldest of the techniques just listed. Like PCA it was developed by Hotelling in the 1930s - see Gittins (1985). Given two sets of variables $\mathbf{y}, \mathbf{x}$ of dimensions p_1, p_2, CCA successively finds pairs of linear combinations of the variables $\{\mathbf{a}'_{k1}\mathbf{y}, \mathbf{a}'_{k2}\mathbf{x}\}$ with maximum correlation, subject to being uncorrelated with previous pairs. It turns out that we need to solve the eigen-equation

$$\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{a}_{k2} = l_k\mathbf{R}_{xx}\mathbf{a}_{k2},$$

where $\mathbf{R}_{xy}, \mathbf{R}_{yx}, \mathbf{R}_{xx}, \mathbf{R}_{yy}$ are correlation matrices for appropriate sets of variables.

The solution is invariant to changes in scales of variables (unlike PCA) so the corresponding equation with correlation matrices replaced by covariance matrices leads to an equivalent solution.

3.2 Maximum Covariance Analysis

This technique has been 'rediscovered' many times. It dates back to Tucker (1958), where it was called *inter-battery factor analysis*. In atmospheric science it was popularised by Bretherton *et al.* (1992) as singular value decomposition (SVD) analysis. This name arises as the technique can be derived via a singular decomposition of the $(p_1 \times p_2)$ matrix of covariances between $\mathbf{x}$ and $\mathbf{y}$. Many other multivariate techniques also emerge from singular value decompositions of (other) matrices, so to minimize confusion Bretherton *et al.*'s unfortunate choice of nomenclature is best avoided.

The technique has been developed from a number of different viewpoints, one of which leads to the better name, *maximum covariance analysis* (von Storch and Zwiers, 1999, Section 14.1.7; Cherry, 1997 suggests *canonical covariance analysis*.) This intepretation shows that the procedure is similar in what it does to canonical correlation analysis but has two differences:

- it successively maximizes covariance rather than correlation
- its vectors of loadings are orthogonal, rather than its derived variables being uncorrelated

The optimization problem is solved by finding the SVD of $\mathbf{S}_{yx}$, or equivalently from an eigenanalysis of $\mathbf{S}'_{yx}\mathbf{S}_{yx} = \mathbf{S}_{xy}\mathbf{S}_{yx}$, where $\mathbf{S}_{yx}, \mathbf{S}_{xy}$ are matrices of covariances between $\mathbf{y}$ and $\mathbf{x}$.

3.3 Redundancy Analysis

In canonical correlation and maximum covariance analyses, the two sets of variables are treated on an equal footing. In redundancy analysis $\mathbf{y}$, $\mathbf{x}$ consist of response and predictor variables respectively. We try to successively maximize the proportion of variance in the response variables that can be accounted for by orthogonal linear combinations of the predictor variables (van den Wollenberg, 1977). This leads to the eigen-equation

$$\mathbf{R}_{xy}\mathbf{R}_{yx}\mathbf{a}_{k2} = l_k\mathbf{R}_{xx}\mathbf{a}_{k2}.$$

To find $\mathbf{a}_{k1}$, the rôles of $\mathbf{x}$ and $\mathbf{y}$ in this equation are reversed.

Redundancy analysis is equivalent to *PCA of instrumental variables* (Rao, 1964), and to a version of *reduced-rank regression* (Davies and Tso, 1982), as well as having links with multivariate regression.

3.4 Principal Predictors

Thacker (1999) proposes *principal predictors*. The technique is defined similarly to redundancy analysis, except that the derived variables are uncorrelated rather than the vectors of loadings being orthogonal. This leads to solving the eigen-equation

$$\mathbf{S}_{xy}[diag(\mathbf{S}_{yy})]^{-1}\mathbf{S}_{yx}\mathbf{a}_{k2} = l_k\mathbf{S}_{xx}\mathbf{a}_{k2}.$$

3.5 Two groups of variables – other possibilities

Our brief descriptions of techniques above are by no means exhaustive. Other possibilities include:

- Multivariate Regression (Davies and Tso, 1982)
- Regression based on PCs in each group (Preisendorfer and Mobley, 1988, Chapter 9)
- CCA based on PCs in each group (Muller, 1982)
- Combined PCA of all $p_1 + p_2$ variables (Bretherton *et al.*, 1992)
- Partial Least Squares (Frank and Friedman, 1993)
- Softly Shrunk Reduced Rank Regression (Aldrin, 2000)
- Latent Variable Multivariate Regression (Burnham *et al.*, 1999)
- Extensions to more than two groups - see Section 4.4

4 Extensions to three (or more) modes

A typical atmospheric science application has the two 'modes' of time and space. There may be a third mode, corresponding to different meteorological variables measured at each time-space location, or to different levels of the atmosphere. We may even have all four modes available. There are a number of extensions of PCA that deal with three-mode or multi-way data, which we now describe briefly.

4.1 Three modes - O-mode . . . T-mode analysis

It is arguable whether this idea is really an extension. If there are 3 modes (space, time, meteorological variables), we can choose 2 out of 3 to represent the variables and observations respectively in a PCA, and analyse those two, fixing the third. This can be done in 6 ways, and O-mode, P-mode, Q-mode, R-mode, S-mode, T-mode correspond to these 6 possibilities (Cattell, 1978; Richman, 1986). S-mode (locations = variables, times = observations) is most common in atmospheric science, but T-mode (times = variables, locations = observations) and occasionally others are also sometimes used.

4.2 Three modes - Extended EOF analysis

If there are n times, s locations, and p variables measured at each space-time location, we can combine locations and variables to give a total of sp variables in a $(n \times sp)$ data matrix, and perform PCA on that matrix. It is also possible to combine two of the three modes in two other ways. A further extension is to incorporate data at different time lags, leading to Multivariate Extended EOF analysis (Mote *et al.*, 2000). The latter also extends multichannel singular spectrum analysis (MSSA - Plaut and Vautard, 1994).

4.3 Three-mode PCA

Suppose that our data are of the form x_{ijk}, where i, j, k index three modes and $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, $k = 1, 2, \ldots, t$. Then three-mode PCA approximates the data by

$$\tilde{x}_{ijk} = \sum_{h=1}^{m} \sum_{l=1}^{q} \sum_{r=1}^{s} a_{ih} b_{jl} c_{kr} g_{hlr},$$

where the values of m, q, s are less, and if possible very much less, than n, p, t respectively, and the parameters $a_{ih}, b_{jl}, c_{kr}, g_{hlr},\ i = 1, 2, \ldots, n,\ h = 1, 2, \ldots, m,\ j = 1, 2, \ldots, p,\ l = 1, 2, \ldots, q,\ k = 1, 2, \ldots, t,\ r = 1, 2, \ldots, s$ are chosen to give a good fit of $\tilde{x}_{ijk}$ to x_{ijk} for all i, j, k. There are a number of methods for solving this problem and, like ordinary PCA, they involve finding eigenvalues and eigenvectors of cross-product or covariance matrices, in this case by combining two of the modes as in extended EOF analysis, before finding cross-products - see Kroonenberg (1983).

4.4 Extensions to more than three modes (groups of variables)

One type of multiway data corresponds to having different groups of variables, which gets us back to the topic of Section 3. Both Casin (2000) and van der Geer (1984) consider a number of techniques applicable to three or more groups.

5 Concluding Remarks

In all the techniques described above, especially those concerned with simplification, there is a desire to find physical interpretations for the results. This is frequently controversial – often there is a failure to recognise that the criteria underlying the techniques are not designed to look for physically meaningful results. If the 'physical modes' are already known, the analysis may be redundant. The confusion that exists is illustrated by the conflicting arguments that 'physical modes' are likely to be correlated (oblique rotation – Richman, 1986), or that they should be independent (independent component analysis – Aires *et al.* (2000)).

One large class of EOF-related techniques not covered in this review are those that are designed to explore oscillatory behaviour in space and time, for example SSA (Golyandina *et al.*, 2001), MSSA/EEOF analysis (Plaut and Vautard, 1994), POP analysis (von Storch *et al.*, 1988), Hilbert EOFs (Cai and Baines, 2001), MTM-SVD (Mann and Park, 1999).

Acknowledgements

This article is based on a talk entitled ‘What’s in a name? ACP, PCA, EOFs and a few related techniques’ given at the ‘Journée ‘classification’ et ‘analyse spatiale” on February 8th, 2002. I am grateful to Philippe Naveau for the invitation to contribute to the ‘Journée’, and for his suggestion that I write up the talk as a review article.

References

Aires, F., Chedin, A. and Nadal, J. P. (2000). Independent component analysis of multivariate time series: application to tropical SST variability. *J. Geophys. Res. - Atmos.*, **105 (D13)**, 17437–17455.

Aldrin, M. (2000). Multivariate prediction using softly shrunk reduced-rank regression. *Amer. Statistician*, **54**, 29–34.

Bretherton, C. S., Smith, C. and Wallace, J. M. (1992). An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5**, 541–560.

Burnham, A. J., MacGregor, J. F. and Viveros, R. (1999). Latent variable multivariate regression modelling. *Chemometr. Intell. Lab. Syst.*, **48**, 167–180.

Cadima, J. and Jolliffe, I. T. (1995). Loadings and correlations in the interpretation of principal components. *J. Appl. Statist.*, **22**, 203–214.

Cadima, J. F. C. L. and Jolliffe, I. T. (2001). Variable selection and the interpretation of principal subspaces. *J. Agri. Biol. Environ. Statist.*, **6**, 62–79.

Cai, W. and Baines, P. (2001). Forcing of the Antarctic Circumpolar Wave by ENSO teleconnections. *J. Geophys. Res.-Oceans*, **106**, 9019–9038.

Casin, Ph. (2001). A generalization of principal component analysis to K sets of variables. *Computat. Statist. Data Anal.*, **35**, 417–428.

Cattell, R. B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. New York: Plenum Press.

Cherry, S. (1997). Some comments on singular value decomposition analysis. *J. Climate*, **10**, 1759–1761.

Chipman, H. A. and Gu, H. (2001). Interpretable dimension reduction. Submitted for publication.

Davies, P. T. and Tso, M. K.-S. (1982). Procedures for reduced-rank regression. *Appl. Statist.*, **31**, 244–255.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics tools. *Technometrics*, **35**, 109–148 (including discussion).

Gittins, R. (1985). *Canonical Analysis. A Review with Applications in Ecology*. Berlin: Springer.

Golyandina, N. E., Nekrutin, V. V. and Zhigljavsky, A. A. (2001). *Analysis of Time Series Structure. SSA and Related Techniques*. Boca Raton: Chapman and Hall.

Hausmann, R. (1982). Constrained multivariate analysis. In: *Optimisation in Statistics* (eds: S. H. Zanckis and J. S. Rustagi), 137–151. Amsterdam: North Holland.

Jolliffe, I.T. (1987). Rotation of principal components: some comments. *J. Climatol.*, **7**, 507–510.

Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints. *J. Appl. Statist.*, **22**, 29–35.

Jolliffe, I. T. (2002). *Principal Component Analysis*, 2nd edition. New York: Springer.

Jolliffe, I. T. and Uddin, M. (2000). The simplified component technique - an alternative to rotated principal components. *J. Computat. Graph. Statist.*, **9**, 689–710.

Jolliffe I. T. and Uddin, M (2002). A modified principal component technique based on the LASSO. Submitted for publication.

Jolliffe I. T., Uddin, M and Vines, S.K. (2002). Simplified EOFs - three alternatives to rotation. To appear in *Climate Res.*.

Kiers, H. A. L. (1993). A comparison of techniques for finding components with simple structure. In: *Multivariate Analysis: Future Directions 2*, (eds. C. M. Cuadras and C. R. Rao), 67–86. Amsterdam: North Holland.

Kroonenberg, P. M. (1983). *Three-Mode Principal Component Analysis*. Leiden: DSWO Press.

Mann, M. E. and Park, J. (1999). Oscillatory spatiotemporal signal detection in climate studies: a multi-taper spectral domain approach. *Adv. Geophys.*, **41**, 1–131.

Mestas-Nuñez, A. M. (2000). Orthogonality properties of rotated empirical modes. *Int. J. Climatol.*, **20**, 1509–1516.

Mote, P. W., Clark, H. L., Dunkerton, T. J., Harwood, R. S., and Pumphrey, H. C. (2000). Intraseasonal variations of water vapor in the tropical upper troposphere and tropopause region. *J. Geophys. Res.*, **105**, 17457–17470.

Muller, K. E. (1982). Understanding canonical correlation through the general linear model and principal components. *Amer. Statistician*, **36**, 342–354.

Plaut, G. and Vautard, R. (1994). Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere. *J. Atmos. Sci.*, **51**, 210–236.

Preisendorfer, R. W. and Mobley, C. D. (1988). *Principal Component Analysis in Meteorology and Oceanography*. Amsterdam: Elsevier.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya A*, **26**, 329–358.

Richman, M. B. (1986). Rotation of principal components. *J. Climatol.*, **6**, 293–335.

Richman, M. B. (1987). Rotation of principal components: a reply. *J. Climatol.*, **7**, 511–520.

Richman, M. B. and Gong, X. (1999). Relationships between the definition of the hyperplane width to the fidelity of principal component loading patterns. *J. Climate*, **12**, 1557–1576.

Thacker, W. C. (1999). Principal predictors. *Int. J. Climatol.*, **19**, 821–834.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, **23**, 111–136.

van de Geer, J. P. (1984). Linear relations among k sets of variables. *Psychometrika*, **49**, 79–94.

van den Dool, H. M., Saha, S. and Johansson, Å. (2000) Empirical orthogonal teleconnections. *J. Climate* **13**, 1421-1435.

van den Wollenberg, A. L. (1977). Redundancy analysis. An alternative for canonical correlation analysis. *Psychometrika*, **42**, 207–219.

Vines, S. K. (2000). Simple principal components. *Appl. Statist.*, **49**, 441–451.

von Storch, H., Bruns, T., Fischer-Bruns, I. and Hasselmann, K. (1988). Principal oscillation pattern analysis of the 30- to 60-day oscillation in general circulation model equatorial troposphere. *J. Geophys. Res.*, **93**, 11022–11036.

von Storch, H. and Zweirs, F. W. (1999). *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.