9.3 ERROR PROPAGATION THROUGH PRINCIPAL COMPONENTS

G. Louis Smith

National Institute for Aerospace
Hampton, Virginia

ABSTRACT

The propagation of measurement errors through the computation of principal components (PC) is treated. In atmospheric sciences applications, typically a principal component analysis is applied to time variations and the associated empirical orthogonal functions (EOF) describe the spatial patterns. The errors in the EOFs are also treated in this paper.

The measurement errors are modeled as consisting of a bias plus uncorrelated errors. The bias errors appear in the mean distribution and do not appear in the PCs or EOFs. The variances of the error in an eigenvalue is proportional to the variance of the measurement errors times the eigenvalue and inversely proportional to the number of regions. The variances of the errors of the PCs are expressed in terms of the PCs, and the variances of errors in the PCs are proportional to the variance of the measurement errors and inversely proportional to the number of regions times the square of the spacing of the eigenvalues.

1. INTRODUCTION

Masssive data sets of climatological parameters have been compiled and are being developed. In particular, satellites provide daily coverage of the Earth for many parameters. To derive information and understanding from these data is a major problem. One tool for studying data is principal component analysis. Often one has measurements at a set of locations and times, which constitute a set of maps describing a time-varying field. One approach to understanding these maps is to extract from them temporal and geographic patterns which describe as much variance as possible by use of principal component analysis (PCA). These descriptions are statistical, but are useful because the underlying correlations are due to the physics of the problem and thus provide insight into the physics. In this paper, the term empirical orthogonal functions (EOFs), which is simply another term for principal components, will be used to denote the geographical patterns.

Principal components, or EOFs, are the eigenvectors of the covariance matrix of the data set. They are statistics based on the data set and as such are subject to sampling errors. The effects of sampling errors on the principal components or EOFs have been studied by North et al. (1982). Errors in the measurements will also result in errors in the principal components or EOFs. The present paper analyses the errors in the computed principal components, which describe the temporal variations, and the EOFs, which describe the geographical variations, due to measurement errors. First a linearized analysis of the propagation of errors through the computation of the covariance matrix into the principal components and EOFs is presented. This is a straightforward exercise in linear algebra. Next, the application of the analysis is demonstrated by computing the errors of the principal components representing the time-variations of the annual cycle of net longwave radiation at the surface over the Earth and the corresponding EOFs which describe the geographic distributions.

2. ANALYSIS

In an application of principal component analysis or EOFs to analysis of atmospheric sciences, one has measurements at a number of locations for a set of times (e.g. Preisendorfer and Mobley, 1988). These values constitute a sequence of maps. One begins by defining the sequence of measurements at a given point as a vector, so that each location on the map has a time-vector for the field. One then computes the temporal mean at each location and subtracts this mean to produce a sequence of anomalies at each location. The outer product of each vector with itself is formed and these

products are summed over the map to produce a covariance matrix for the time-variations. The eigenvalues and eigenvectors, or principal components, of this covariance matrix are then computed, and ordered by eigenvalue, with the first PC having the largest eigenvalue. Each PC is then projected onto the sequence of maps to produce an empirical orthogonal function (EOF), which is the corresponding map. This analysis includes an error with the measurement and follows the above procedure. Alternatively, one may form a vector for each map at a given time, so that one has a sequence of spatial vectors with which to compute a covariance matrix in space. The principal of duality affirms that the results are the same from either path.

2.1 Preliminaries

A set of measurements of a parameter r_{mg} is given, where $m \in [1, M]$ denotes time and $g \in [1, N]$ denotes the geographical location or grid number. In this paper it is assumed that the set is complete, i.e. there are no data voids or gaps. The parameter is partitioned into its mean value r_g and an anomaly z_{mg}. The *measured* values of r_{mg} contain errors, which will be partitioned into a bias error b_g for each location and a random contribution w_{mg} which varies with each measurement and has population mean of zero. Thus

$$r_{mg} = r_g + z_{mg} + b_g + e_{mg} \qquad (1)$$

The first step of a principal component analysis is to compute the mean of the parameter for each grid location and to subtract this mean from each measurement. The computed mean is thus

$$\langle r_g \rangle = r_g + b_g + e_g \qquad (2)$$

where $e_g = \sum w_{mg}/M$ is the error of the mean for region g due to the random error of the measurements and is has a standard deviation of $\sigma_{wg}/\sqrt{M-1}$. The computed anomaly at time m for region g is

$$x_{mg} = r_{mg} - \langle r_g \rangle = z_{mg} + y_{mg}$$

where $y_{mg} = w_{mg} - e_{mg}$ is a random variable with sample mean of zero over the M measurements and standard deviation

$$\sigma_{yg} = \sigma_{wg} \sqrt{\frac{M-2}{M-1}}\,.$$

The bias errors of the measurements thus appear in the mean of the computed field and not in the computed anomalies. The computed anomalies contain only the random component of the measurement errors. The next step is to compute the covariance matrix of the anomalies.

2.2 Covariance Matrix with errors

The covariance matrix for the field is defined as

$$C_0(i,j) = \sum_{g=1}^{N} w_g z_{ig} z_{jg} \qquad (4)$$

and is thus of dimension MxM The time history of values of the parameter r_{mg} for a region g define a vector $\boldsymbol{v}_g$ of dimension M. The C_0 matrix for the field can be expressed as the sum of the area-weighted outer products of the $\boldsymbol{v}_g$:

$$C_0 = \sum_{g=1}^{N} w_g v_g^T v_g\,.$$

The principal components are the eigenvectors of C_0, given by the relation

$$Cu = lu \qquad (5)$$

The eigenvectors are defined to be normalized such that

$$u_i^T u_i = 1\,.$$

Because of measurement errors, the z_{ig} are replaced by x_{ig}, whence by eq. (3) the covariance matrix becomes

$$C(i,j) = \sum_{g=1}^{N} w_g \left(z_{ig} + y_{ig}\right)\left(z_{jg} + y_{jg}\right)$$

$$= C_0(i,j) + dC(i,j) \qquad (6)$$

where $dC(i.j)$ is the first order perturbation of C and

$$dC(i,j) = \sum_{g=1}^{N} w_g \left(z_{jg} y_{ig} + z_{ig} y_{jg}\right) \qquad (7)$$

Second order terms in the errors y_{ig} have been neglected.

(3)

2.3 Errors of eigenvalues and eigenvectors

The effect of a first order perturbation in C is to perturb the eigenvalues l_{i0} and

eigenvectors u_{i0} in accordance with eq. (5), whence

$$dCu_{0i} + C_0 du_i = dl_i u_{0i} + l_{0i} du_i \quad (8)$$

The perturbation to the eigenvector is normal to the eigenvector, i.e.

$$u_i^T du_i = 0 \quad (9)$$

Premultiplying eq. (8) by u_i^T and using eq. (9) gives

$$dl_i = u_{0i}^T dC u_{0i} \quad (10)$$

The perturbation of u_i can be expressed in terms of the eigenvectors as a basis set:

$$du_i = \sum_{j \neq i} a_{ij} u_j \quad (11)$$

where by eq. (9) the summation excludes the *i*-th eigenvector. Equation (10) is used in eq. (9) and the result is premultiplied by u_j^T, giving the result that

$$a_{ik} = \frac{u_{0k}^T dC u_{0i}}{l_{0i} - l_{0k}} \quad (12)$$

For the present case, the perturbation of the covariance matrix, *?C* is given by the summation term in eq. (6).

The mean of y_{ig} is zero, so that the mean of *?C* is zero, whence by eq. (9) the mean perturbations of the eigenvalues are zero. Likewise, by eqs. (10) and (11), the means of the eigenvectors, or principal components, are zero. The next question is what are the standard deviations of the perturbations of the eigenvalues and the principal components.

The standard deviations of the perturbations of the eigenvalues are defined as the expected value of $(??_i)^2$. Equation (10) is squared and the expected values taken for both sides. The following assumptions are now made for the measurement errors:

i. Measurements errors are independent of time or location.
ii. Measurement errors are uncorrelated in space, i.e. one grid point to another.
iii. Measurement errors are uncorrelated in time.

From these three assumptions,

$$E\left\{y_{ig} y_{jh}\right\} = s_e^2 d_{ij} d_{gh} \quad (13)$$

It is also assumed that the weightings for all regions are the same, thus $w_g = 1/N$. From eqs. (10) and (13) it follows that the variances of the changes in the eigenvalues are

$$s_{lk}^2 = \frac{4 s_e^2 l_k}{N} \quad (14)$$

The variance of the error of the *i*-th component of the *j*-th eigenvector is found by use of eq. (11), whence

$$s_{uki}^2 = \sum_{p=1}^{M} \sum_{q=1}^{M} E\left\{a_{ip} a_{iq}\right\} u_{kp} u_{kq} \quad (15)$$

The covariances of the $a_{ip} a_{iq}$ are computed by use of equations (12) and (14), giving

$$E\left\{a_{ip} a_{iq}\right\} = 0 \quad \text{if } p = i, \text{ or } q = i, \text{ or } p \neq q$$

$$= \frac{s_e^2 \left(l_i + l_p\right)}{N^2 \left(l_i - l_p\right)^2} \quad \text{if } p = q \quad (16)$$

The coupling coefficients K_{ip} are defined as

$$K_{ip} = \frac{l_i + l_p}{(l_i - l_p)^2} \quad \text{for } i \neq p \quad (17)$$

$$= 0 \quad \text{for } i = p$$

These coefficients appear repeatedly in this analysis. The variances of errors of the eigenvectors are now written as

$$s_{uki}^2 = \frac{s_e^2}{N} \sum_{p=1}^{M} K_{kp} u_{pi}^2 \quad (18)$$

Equation (18) applies to the individual components of the eigenvector $\boldsymbol{u}_k$. For the average of the components of $\boldsymbol{u}_k$, the normality condition gives the average over the components of u_{pi}^2 as M^{-1}, so that the variance of the mean of the components is

$$s_{uk}^2 = \frac{s_y^2}{MN} \sum_{p \neq k} K_{kp} \quad (19)$$

2.4 Errors of EOFs:

Once the eigenvectors, or principal components, have been computed, the corresponding EOFs, or geographic patterns, can be computed as

$$F_{kg} = \sum_{m=1}^{M} u_{km} z_{mg} \quad (20)$$

The effects of measurement errors are to add an error to the z_{mg} and also to change the eigenvectors, so the perturbation in the EOF is

$$dF_{kg} = \sum_{m=1}^{M} (u_{km} e_{mg} + du_{km} z_{mg})$$

The variance of the perturbation is

$$s^2_{Fkg} = \sum_{p=1}^{M}\sum_{q=1}^{M}(u_{kp}E\{y_{pg}y_{qg}\}u_{kq} + z_{pg}E\{du_{kp}du_{kq}\}z_{qg} + u_{kp}E\{y_{pg}du_{kq}\}z_{qg} + u_{kp}E\{y_{pg}du_{kq}\}z_{qg})$$

which reduces to

$$s^2_{Fkg} = s^2_y[1+ N^{-2}\sum_m F^2_{jg}K_{kj}] \quad (21)$$

The first term is simply the error at the location due to the measurement error and the second term is due to the error which is induced in the EOFs by the measurement errors.

3. CONCLUSIONS

An analysis of effects of measurement errors on computed PCs and EOFs is presented. The measurement errors are modeled as consisting of a bias plus uncorrelated errors. The bias errors appear in the mean distribution and do not appear in the PCs or EOFs. The variances of the error in an eigenvalue is proportional to the variance of the measurement errors times the eigenvalue and inversely proportional to the number of regions. Errors in the PCs and EOFs are written in terms of contamination of PCs by each other due to the measurement errors, as quantified by interaction coefficients. These coefficients are proportional to the sum of the eigenvalues of the contaminating and contaminated PC and inversely proportional to the square of the difference of the eigenvalues.

The variances of the errors of the PCs are expressed in terms of the PCs, and the variances of errors in the PCs are proportional to the variance of the measurement errors and the interaction coefficients and inversely proportional to the number of regions. The variances of the errors in the EOFs, describing the spatial patterns, are due to the measurement errors directly and also to the contamination of EOFs by each other due to the measurement errors. This contamination in the spatial domain is also described by the interaction coefficients. The interaction coefficients increase quickly with increasing order of the contaminated PC or EOF, so that the growth of errors in PCs and EOFs with order is very rapid and the high order PCs and EOFs become overwhelmed by these errors.

4. ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of this research by the Surface Radiation Budget Program in the Science Directorate of the Langley Research Centre. This program is funded by the Office of Space of NASA as a part of the Global Energy and Water Experiment.

5. APPENDIX: List of Symbols

b_g bias of measurement error of region g
C_{ij} element of covariance matrix
$E\{\}$ expected value of
e measurement error
F_{kg} value of k-th empirical orthogonal function at region g
M number of times
N number of regions
r_{mg} measurement of region g at time m
$\boldsymbol{u_k}$ k-th normalized principal component vector
w_g area weighting of region g
x_{mg} random part of measurement error at region g and time m
y_{mg} measurement error of region g at time m with bias estimate removed
z_{mg} anomaly of region g at time m
$?_{ij}$ interaction coefficient for effect of uj on ui
$?_i$ i-th eigenvalue
$?_x$ standard deviation of error in x

6. REFERENCES

North, G. R., T. L. Bell, R. F. Cahalan and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions, Mon. Wea. Rev., 110, 699-706.

Preisendorfer and Mobley, 1988: *Principal Component Analysis in Meteorology and Oceanography*, Elsevier.

Papoulis, A., 1965: *Probability, Random Variables and Stochastic Processes,* McGraw-Hill Book Co., New York, NY.

Wilber, A. C., G. L. Smith, D. Rutan, C. H. Whitlock, T. P. Charlock, N. A. Ritchey and S. K. Gupta, 1996: Annual and Interannual Variations of the Surface Radiation Budget, Proc. Global Ocean, Atmospheric and Land System, Amer. Met. Soc.,, Atlanta, Ga. 29 Jan.-2 Feb.