AN ACE-BASED NONLINEAR EXTENSION TO TRADITIONAL EMPIRICAL ORTHOGONAL FUNCTION ANALYSIS

KLAUS BAGGESEN HILGER AND ALLAN AASBJERG NIELSEN

Informatics and Mathematical Modelling, Technical University of Denmark, Building 321
DK-2800 Kgs. Lyngby, Denmark, Telephone +45 4525 3422, Telefax +45 4588 1397
{kbh,aa}@imm.dtu.dk, www.imm.dtu.dk/~{kbh,aa}

OLE BALTAZAR ANDERSEN AND PER KNUDSEN

National Survey and Cadastre, Rentemestervej 8
DK-2400 Copenhagen NV, Denmark, Telephone +45 3587 5050, Telefax +45 3587 5051
{oa,pk}@kms.dk, research.kms.dk/~{oa,pk}

This paper shows the application of the empirical orthogonal functions/principal component transformation on global sea surface height and temperature data from 1996 and 1997. A nonlinear correlation analysis of the transformed data is proposed and performed by applying the alternating conditional expectations algorithm. New canonical variates are found that indicate that the highest correlation between ocean temperature and height is associated with the build-up of the El Niño during the last half of 1997.

1 Introduction

The aim of this paper is to present an extension to the traditional empirical orthogonal functions (EOF) analysis [10,17]. EOF is often applied in geophysical sciences to analyze temporal sequences. In the two-set case the traditional EOF analysis can be extended by the means of linear canonical correlation analysis (CCA) [8,4,1,14]. This paper presents a nonlinear correlation analysis of two-set data using the alternating conditional expectations (ACE) transform. ACE was originally proposed for nonlinear multiple regression [2]. The ACE transform is applied using the EOF analysis as a preprocessor. Two temporal sequences of scalar image data representing global sea surface temperature (SST) and sea surface height (SSH) are analyzed as these quantities are physically related. The data period covers 1996 and 1997 on a monthly basis. This particular period was chosen, as the build-up of one of the largest El Niño events [11] ever recorded occurred during the last half of 1997.

2 Proposed method

The EOF analysis is closely related to the principal components analysis (PCA) [7,1]. Often the usual PCA assumption on variables with mean zero is replaced by an assumption of temporal means of zero. A traditional CCA analysis involves a

joint assessment of two sets to produce pairs of canonical variates (CVs) that are linear combinations of the original variables with maximal correlation. All the CV pairs are restricted to be orthogonal/uncorrelated. Generalizations of CCA to handle multiset scenarios can be made [9,13]. In this paper we address the problem of finding the pair of canonical variates that maximize the correlation in a more general manner. We do not restrict transformations to be simple linear combinations of the original variables, but we allow for nonlinear mappings of the data. The nonlinear CV transformations are found as conditional expectations of the original data, applying the ACE algorithm. Traditionally ACE is used to handle data that naturally separates into two sets often with only one variable in one of the sets. The algorithm can however be generalized to handle several variables in both sets and even multiset scenarios [6]. By allowing for nonlinear transformations, it is possible to determine CVs with higher correlations than in the linear CCA analysis. However, it becomes more challenging to interpret the CV sets, as the degree of freedom for the transformations is very high.

Consider a multivariate data set of P variables with gray levels $r_i(\boldsymbol{x})$, $i = 1,\ldots,P$ where $\boldsymbol{x}$ is the coordinate vector denoting a grid point of the sample. Let $\boldsymbol{r}(\boldsymbol{x}) = [r_1(\boldsymbol{x}) \ \ldots \ r_P(\boldsymbol{x})]^T$ represent a random signal variable and assume first and second order stationarity such that $\mathbf{E}\{\boldsymbol{r}(\boldsymbol{x})\} = 0$ and $\mathbf{E}\{\boldsymbol{r}(\boldsymbol{x})\,\boldsymbol{r}(\boldsymbol{x})^T\} = \boldsymbol{\Sigma}$.

Determining the direction of maximum variation means finding the direction $\boldsymbol{a}$, with $\boldsymbol{a}^T\boldsymbol{a} = 1$, such that the linear combination $y(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{r}(\boldsymbol{x})$ possesses maximum variance. The principal components transform chooses P linear transformations $y_i(\boldsymbol{x}) = \boldsymbol{a}_i^T\boldsymbol{r}(\boldsymbol{x})$, $i = 1,\ldots,P$ such that the variance for $y_i(\boldsymbol{x})$ is maximum among all linear transforms orthogonal to $y_j(\boldsymbol{x})$, $j = 1,\ldots, i\text{-}1$. The variance is given by $\mathbf{V}\{\boldsymbol{a}_i^T\boldsymbol{r}(\boldsymbol{x})\} = \boldsymbol{a}_i^T\boldsymbol{\Sigma}\,\boldsymbol{a}_i$. Thus, we see that the basis for the principal components is identified as the conjugate eigenvectors of the dispersion matrix. Let $v_1 = \mathbf{V}\{\boldsymbol{a}_1^T\boldsymbol{r}(\boldsymbol{x})\} \geq \ldots \geq v_P$ be the eigenvalues with the corresponding conjugate eigenvectors $\boldsymbol{a}_1, \ldots \boldsymbol{a}_P$, then $y_j(\boldsymbol{x})$ is the i'th principal component.

Consider two sets of random signal variables $\boldsymbol{r1}(\boldsymbol{x})$ and $\boldsymbol{r2}(\boldsymbol{x})$, each consisting of P1 respectively P2 variables. The alternating conditional expectations (ACE) algorithm chooses P = P1+P2 transformations $\{\varphi 1_1, \ldots, \varphi 1_{P1}\}$ and $\{\varphi 2_1, \ldots, \varphi 2_{P2}\}$ of the original variables such that

$$e^2 = \mathbf{E}\{[\,\Sigma\,\varphi 1_i(r1_i(\boldsymbol{x})) - \Sigma\,\varphi 2_j\,(r2_j(\boldsymbol{x}))\,]^2\} \,/\, \mathbf{E}\{[\,\Sigma\,\varphi 1_i\,(r1_i(\boldsymbol{x}))\,]^2\}$$

is minimized.

Applying ACE under the constraints that the new canonical variates must have variance one, ACE determines the transformations that maximize the correlation, $\mathbf{Corr}\{z1(\boldsymbol{x}), z2(\boldsymbol{x})\} = 1 - e^2/2$, between the new components $z1(\boldsymbol{x}) = \Sigma\varphi 1_i(r1_i(\boldsymbol{x}))$ and $z2(\boldsymbol{x}) = \Sigma\,\varphi 2_j\,(r2_j(\boldsymbol{x}))$. The new transformations need not be linear but can be even non-monotonic mappings. The ACE algorithm can be implemented to handle arbitrary mixtures of continuous ordered variables and categorical variables.

Optimal transformations exist and satisfy a complex system of integral equations [2]. Convergence to the optimal solutions is obtained through an iterative algorithm using only bivariate conditional expectations. Thus the ACE algorithm can be applied using a wide range of bivariate smoothers; we apply Friedman's local nonlinear supersmoother [4]. Note that suboptimal solutions can also be found applying multiple regression as suggested in [3].

3 Data and results

The data used are global monthly mean values of 1996-1997 SST data from the NOAA/NASA Oceans Pathfinder AVHRR SST database [16] and global monthly mean values of 1996-1997 SSH data from the NASA/GSFC Ocean Altimeter Pathfinder database [12,16]. The SSH data are interpolated point observations from the TOPEX/Poseidon radar altimeter mission. The SST data come as 360 rows by 720 columns half degree data starting at 180° longitude, the SSH data come as 179 rows by 360 columns one degree data starting at 0° longitude. The SST data have been resampled to the SSH grid. The AVHRR instrument is influenced by cloud coverage whereas the radar altimeter provides uninterrupted data. For a temporal analysis of the same data, see [15]. Statistics for the analysis are calculated where SST and the SSH have non-missing values for all 24 months.

In Figure 1 the first three principal components (PCs) are shown for both the SST and the SSH data. The first three SST-PCs explain approximately 99% of the total variation of the data. The first three SSH-PCs explain 59% of the variation in the data. By studying the principal components one can obtain valuable information about the dynamics of the ocean. In Figure 2 is included the correlations between the original SST and the SSH variables, and the PCs.

The PC transformed SST and SSH data are used as input to ACE. The orthogonal transformation is helpful when we wish to evaluate the results of the nonlinear correlation analysis. Inference problems are avoided that can occur when the original data are cross-correlated. The ACE algorithm applied generates nonlinear mappings of each input variable. The sum of all the transformed variables in each set determines the first ACE CV pair. In Figure 3 the first two ACE CV pairs are shown. The correlation for the first pair is 0.88 and for the second pair 0.63. The two highest correlations found using a linear approach are 0.76 and 0.72 [14]. In the figure is also shown the squared differences for each CV pair. The ACE analysis is performed on a complete set using all PCs for both SST and SSH. In Figure 4, for the first ACE pair, is shown the first three ACE transformations of the PCs together with a bar plot of the variances of all the ACE transforms. The bar plot may be useful when determining which transforms dominate each ACE CV.

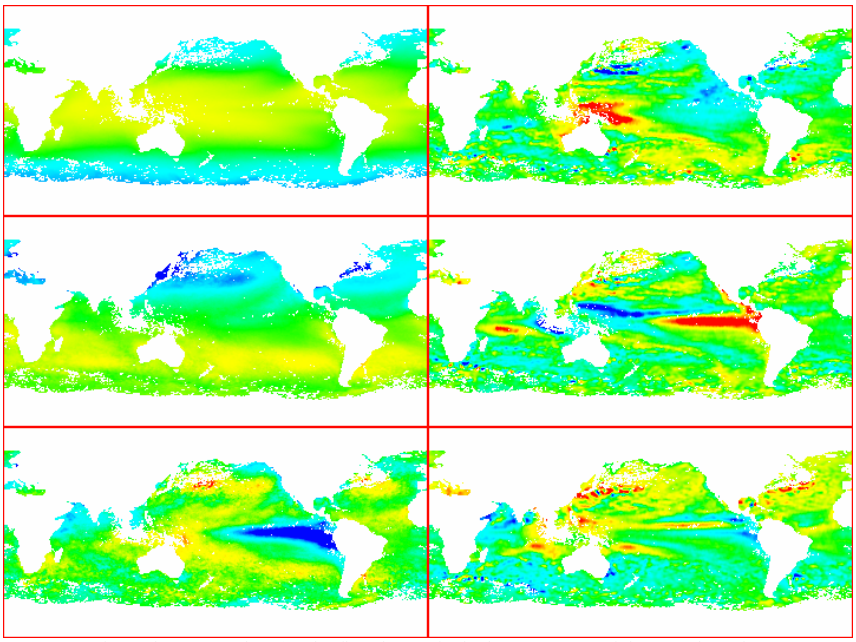

Figure 1. First column (top-down): The first three principal components of the SST data, the percentage of the variance explained by each component is 95%, 4%, and 0.2%. Second column: The first three principal components of the SSH data. The percentage of variance explained by each component is 29%, 23%, and 7%. The images are stretched linearly from mean ± three standard deviations and shown in pseudocolor, blue is minimum and red maximum.

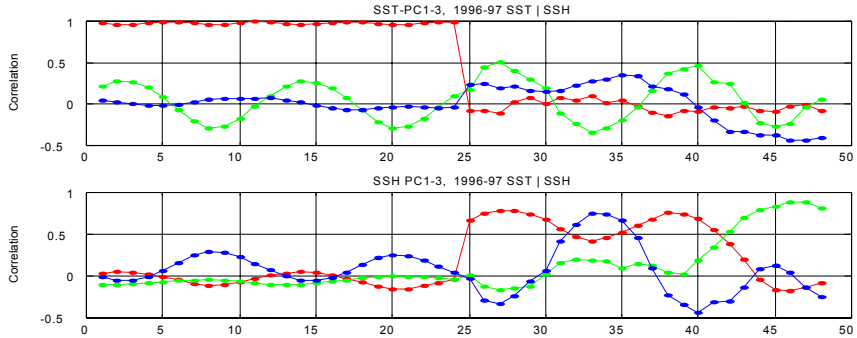

Figure 2. Top: The correlations between original SST (labeled 1-24) and SSH variables (labeled 25-48), and SST-PCs 1 through 3 (shown as red, green, and blue). Bottom: The correlations between original SST (labeled 1-24) and SSH variables (labeled 25-48), and SSH-PCs 1 through 3 (shown as red, green, and blue).

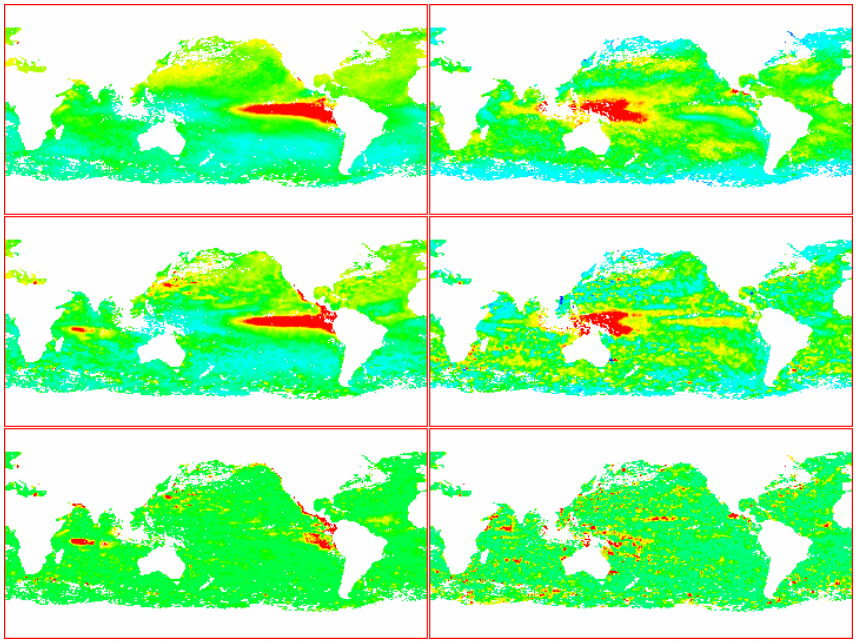

Figure 3. The columns contain the first respectively the second ACE CV pairs, and their squared differences. Each column (top-down): The ACE CV of the SST-PCs, the ACE CV of the SSH-PCs, and the squared differences. The images are stretched linearly from mean ± three standard deviations and shown in pseudocolor, blue is minimum and red maximum.

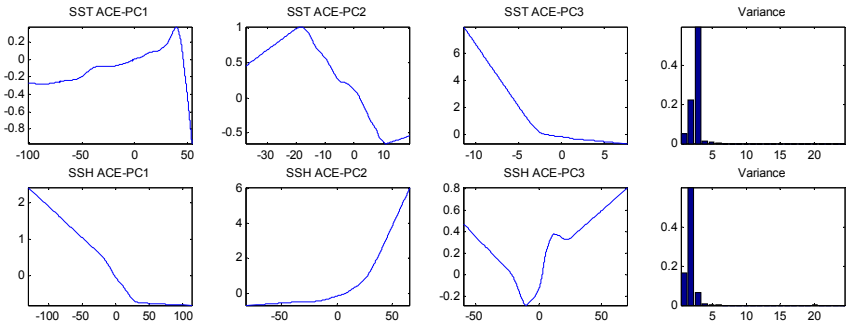

Figure 4. Top row: The first three ACE transformations of the SST-PCs - transformed value vs. input value. The last column contains a bar plot of the variances of all the ACE transformations for the SST PC1-24. Bottom row: The first three ACE transformations of the SSH-PCs - transformed value vs. input value, and a bar plot of the variances of the ACE transformations for the SSH PC1-24.

4 Discussion and conclusions

Inspecting the correlations between the SST-PCs and the sea surface temperature in Figure 2 reveals that the annual cycle is very dominant. SST-PC1 represents a temporal mean signal of the SST and has high positive semi-annually oscillating correlation with all of the SST months. The component represents the fact that it is warm at the Equator and colder near the poles. There is no apparent correlation with the SSH field because the mean of this field has been removed prior to the analysis.

In SST-PC2 we see highs in the Southern Hemisphere and lows in the Northern Hemisphere. In the correlations the annual cycle [9] is clearly present with positive correlations in the months from December to May and negative correlations from June to November. In SST-PC3 a strong signal off the equatorial west coast of South America is present. The component does contain some annual signal but has correlations very close to zero to the SST data. Near the end of 1997 where the El Niño is starting to build up we see a slight divergence from the zero into negative correlations. The negative correlations indicate that during the El Niño the temperature of the west coast of South America is high. Focusing on the El Niño event we see that especially the signal in SSH-PC2 is related to the phenomenon with high correlations to SSH in the last half of 1997. SSH-PC1 appears to be the mean height signal and is highly correlated with the SSH expect during the El Niño where it has negative correlations closer to zero. SSH-PC3 seems related to the same annual cycle as found for SST-PC2.

In Figure 4, for the first ACE CV pair, we see that the variance of the first two to three ACE-PCs dominate the SST and the SSH data. The remaining ACE transforms are all very close to zero. Inspecting SST-PC3 and SSH-PC2 in Figure 1, we would expect ACE to change the sign of one of the components when looking for maximum correlation. This is exactly what is happening with SST ACE-PC3 being transformed using a monotonely decreasing function and transforming SSH ACE-PC2 with a function that is monotonely increasing. Thus the ACE CVs focus on the El Niño event as the part of the signal containing the highest correlation between SST and SSH. The nonlinear transformations of the SST-PC1 and 2, and the SSH-PC1 and 3, appear to be included in the ACE CVs to explain additional spatial patterns present primarily within the Pacific Ocean. In the CVs in Figure 3 the pattern can be seen as a low signal in the western and southern part of the Pacific Ocean. Inspecting the ACE CVs for the first ACE pair the El Niño appears very strongly. Highs in the squared difference image show where the global model is less successful in correlating sea surface temperature to the sea surface height. This is particularly conspicuous near the west coast of Central and South America. There are also contributing regions in the Atlantic and Indian Oceans. Looking at the second ACE CV pair we notice an interesting signal in the western part of the Pacific Ocean. The signal is recognized as being related to the usual ocean temperature and height configuration. A full interpretation of the lower order ACE

pairs is beyond the scope of this paper. The ACE analysis seems to be able to separate relevant ocean configurations from the temporal data. It looks for high correlations between the involved variables through nonlinear mappings, and finds components with higher correlations in comparison to those found by a linear analysis, in effect producing a more detailed decomposition of the multivariate data. Furthermore, the ACE analysis is purely data-driven and thus constitutes a useful exploratory tool for a data analyst when looking for insight into the structure of data.

The analysis presented in this paper is in good agreement with the established oceanographic knowledge on the build-up of one of the largest El Niño events on record.

Acknowledgement

The Pathfinder SST data provision at JPL is due to J. Vazques, R. Sumagaysay and co-workers. The Pathfinder SSH data provision at GSFC is due to V. Zlotnicki and co-workers. This work is done as a part of the GEOSONAR project funded by the Danish National Research Councils under the Earth Observation Program. GEOSONAR is headed by Dr. Per Knudsen, National Survey and Cadastre, Denmark.

References

1. Anderson T. W., *An Introduction to Multivariate Statistical Analysis*, 2nd edition, J. Wiley, 1984.
2. Breiman L. and Friedman J. H., Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, 80(391), 580-598, 1985.
3. Buja A. and Kass R. E., Comment: Some observations on ACE methodology, *Journal of the American Statistical Association*, 80(391), 602-607, 1985
4. Cooley W. W. and Lohnes P. R., *Multivariate Data Analysis*, John Wiley and Sons, 1971.
5. Friedman J. H. and Stuetzle W., Smoothing of scatterplots, Technical Report ORION006, Department of Statistics, Stanford University, 1982.
6. Hilger K.B., *Exploratory analysis of multivariate single and multisets*, Ph.D. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, in prep.
7. Hotelling H., Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24, 417-441, 1933.
8. Hotelling H., Relations between two sets of variates, *Biometrika*, XXVIII, 321-377, 1936.

9. Kettenring J. R., Canonical analysis of several sets of variables, *Biometrika*, 58, 433-451, 1971.
10. Knudsen P., Andersen O. B. and Knudsen T., ATSR sea surface temperature data in a global analysis with TOPEX/Poseidon altimetry, *Geophysical Research Letters,* 23(8), 821-824, 1996.
11. NASA Facts, *El Niño*, The Earth Science Enterprise Series, NF-211, 1999.
12. NASA/GSCF, http://neptune.gsfc.nasa.gov/ocean.html, Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, Maryland, USA.
13. Nielsen A. A., Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data, *IEEE Transactions of Image Processing*, accepted 2001.
14. Nielsen A. A., Hilger K. B., Andersen O. B. and Knudsen P., A bivariate extension to traditional empirical orthogonal function analysis, *MultiTemp 2001*.
15. Nielsen A. A., Hilger K. B., Andersen O. B. and Knudsen P., A temporal extension to traditional empirical orthogonal function analysis, *MultiTemp 2001*.
16. PO.DAAC, http://podaac.jpl.nasa.gov, Jet Propulsion Laboratory, National Aeronautics and Space Administration, Pasadena, California, USA.
17. Preisendorfer R. W., *Principal Component Analysis in Meteorology and Oceanography*, posthumously compiled and edited by C. D. Mobley. Developments in Atmospheric Science, 17, Elsevier, 1988.