Anais XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia, Brasil, 16-21 abril 2005, INPE, p. 3495-3502.

C-NLPCA: Extracting Non-Linear Principal Components of Image Datasets

Silvia Silva da Costa Botelho[1], Rodrigo Andrade de Bem[1],
Ígor Lorenzato Almeida[1], Mauricio Magalhães Mata[1]

[1]Fundação Universidade Federal do Rio Grande - FURG
Av. Itália, Km 8 - Carreiros - 96201-900 Rio Grande, RS
{silviacb, debem, igor}@ee.furg.br, mauricio.mata@furg.br

Abstract. In this paper we apply a Neural Network (NN) to reduce image dataset, distilling the massive datasets down to a new space of smaller dimension. Due to the possibility of these data have nonlinearities, traditional multivariate analysis, like the Principal Component Analysis (PCA), may not represent reality. Alternatively, Nonlinear Principal Component Analysis (NLPCA) can be performed by a NN model to fulfill that deficiency. However, when the dimension of the image increases, NN may easily saturate . This work presents an original methodology associated with the use of a set of cascaded multi-layer NN with a bottleneck structure to extract nonlinear information of the large set of image data. We illustrate its good performance with a set of tests against comparisons using this methodology and PCA in the treatment of oceanographic data associated with mesoscale variability of an oceanic boundary current.
Keywords: neural network, image processing, PCA, cascaded-NLPCA.

1. Introduction

A general problem faced in computer science is to reduce the dimensions of a large datasets, like image datasets, in order to make sense of the bulk information contained in them.

A classical approach to go about this problem is to use linear solvers based on multivariate statistics like the Principal Component Analysis (PCA), Preisendorfer (1988). The PCA finds the eigenmodes of the data covariance matrix and, with this result, one is able to reduce dimensionality and analyze the main patterns of variability present in the dataset. The fact that PCA solves the eigenmode problem using a linear approach may lead the result to be an oversimplication of the variability contained in the dataset, especially if the processes ruling this variability have a nonlinear nature. The artificial Neural Network approach, called Nonlinear Principal Component Analysis (NLPCA), has been applied by several authors as a tool to try to overcome the limitations imposed by linear PCA, Lek(1999), Monahan (2000). The main advantage besides being able to take nonlinearity into account is that the computational process can occur unbiased by our knowledge about the variability aspects that ultimately control the study case.

However, NLPCA brings together an important limitation: saturation phenomena presented in Neural Networks (NN) prevents the use of this approach to handle large datasets. Hence, dataset forming large matrices (or data units), like images, can not be treated by Neural-based NLPCA.

When the dimension of image is much bigger with respect to the number of temporal samples available to analyse, a pre-processing stage is thus necessary to extract the relevant information backbone prior to the NN run. For instance, a PCA can be used as a dimension reductor, leaving the NN to work over a few modes only, Hsieh (2001). In this case, NLPCA runs with linearly reduced input patterns, thus limiting the method's potentials.

Hence, this work presents an original approach, called Cascaded-NLPCA (C-NLPCA), whose main purpose is to eliminate the pre-filtering stage, allowing the nonlinear PCA of the whole image. Our approach does not impose limitations associated with the original dimensions

of the image, allowing important result gains. The C-NLPCA can be used in a set of different domains. Particularly, we are interested in evaluating the potential of the approach to investigate the satellite image variability in oceanic areas dominated by strong mesoscale dynamics. The results presented here are compared with the classical PCA technique and NLPCA, highlighting the advantages and disadvantages of C-NLPCA from the computational and physical sense.

The text is structured with the following section presenting feature extraction problem of large dataset. Next, the theory of PCA and NLPCA is mentioned. The paper core is presented in the fourth section, which details our approach, giving a formal description of the Cascaded Nonlinear Principal Component Method. It is followed by section 5, which presents our satellite images used to validate our approach, testing C-NLPCA against the PCA. Finally, last section contains the general conclusions and suggestions for follow-up studies.

2. Feature Extraction Problem

2.1. Image Vector

Independently of their nature, temporal data samples can be viewed as a vector. For instance, we can consider the dataset as a set of images, whose width and height associated are *w* and *h* pixels respectively. Thus, the number of components (pixels) of this vector will be *w* * *h*. Each pixel is coded by one vector component. The construction of this vector, called *image vector* $\vec{X}$, from an image is performed by a simple concatenation - the rows of the image are placed each beside one another, Romdhani (1999).

2.2. Image Space

The image vector belongs to a space, called image space, which is the space of all images whose dimension is *w* * *h* pixels. Thus, when plotting the *image vectors* they tend to group together to form a narrow cluster in the common image space. This is shown in Figure 2, where hypothetic image series with 3 pixels are showed. Similar images with common features (i.e. land masses, clouds, equal temperature zones, etc in oceanographic domain) are grouped together.

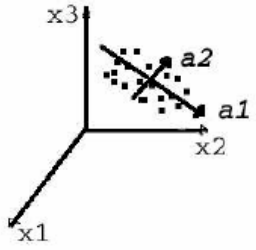

Figure 1: A set of Image Vectors, $\vec{X}(t) = x_1 + x_2 + x_3$, in a Image space with principal orthogonal axes (a1 and a2).

3. Principal Component Analyses Methods

3.1. Principal Component Analysis

Due to the cluster feature containing our image vectors, the full image space may be not an optimal space for our data description. This happens because we can have similarities (redundant information) between same components of different images. Thus, for several domains it can be

interesting to build a new space, lower in size than the original, which better describes the patterns contained in the image dataset.

The base vectors of this new space are called the Principal Components. Of course, using every pixel may bring redundant information, as each pixel depends on its neighbors. So if the dimension of the original image space is *w* * *h*, then the dimension of the new space is less than the dimension of the original image space. Thus, the goal of the PCA is to reduce the dimension of the original set or space so that the new basis better describes the typical îmodelsî of the set. PCA aims to catch the total variation in the set of images, and to explain this variability by few modes. The new basis vectors (axes) will be constructed by a linear combination (thus they are essentially orthogonal). Components in this new basis will be uncorrelated and will maximize the variance accounted for in the original variable. Those axes are shown in Figure 2. We can see that the variance of the data is a maximum in the direction **a1** and so defined as first principal component of variability. The second direction that yields the largest variance of the data, provided that it is orthogonal to **a1**, is **a2**.

Theory of PCA. Let $\overset{\rho}{X}(t) = x_1,...,x_p$ be a dataset, with dimension $\boldsymbol{p}$, where each variable $x_i, (i = 1,..., p)$. is a time series containing **n** observations. PCA transformation is given by a linear combination of $\boldsymbol{x_i}$, time function $\boldsymbol{u}$, and an associated vector $\boldsymbol{a}$:

$$u(t) = a * \overset{\rho}{X}(t), \tag{1}$$

so that

$$\left\langle \left\langle \left\| \overset{\rho}{X}(t) - au(t) \right\|^2 \right\rangle \right\rangle, \tag{2}$$

is minimized ($\langle\langle ... \rangle\rangle$ denotes a sample or time mean). Here u , called the first principal component (PC), is a time series, while $\boldsymbol{a}$, the first eigenvector of the data covariance matrix, often describes a spatial pattern. From the residual, $\overset{\rho}{X} - au$, the second PCA mode can be obtained, and so on for higher modes (see Hsieh (2001) for more details).

3.2. The theory of NLPCA

PCA only allows a linear mapping from $\overset{\rho}{X}$ to *u*. On the other hand, NLPC is obtained using a multi layer Neural Network, see figure 2, Kirby (1990). To perform NLPCA, the NN contains 3 hidden layers of neurons between the input and output layers. Hidden layers have nonlinear activation functions between the input and bottleneck layers and between the bottleneck and output layers. Hence, the network models a composition of functions. The five-layer NLPCA network has *p* nodes in the input layer, *r* nodes in the third (bottleneck) layer, and *p* in the output layer. Output layer must reproduce the input signals presented to the network. The nodes in layer 2 and 4 must have nonlinear activation functions, and the nodes in layer 1, 3 and 5 use linear activation functions. NLPCA network allows data compression/reduction because the p-dimensional inputs must pass through the r-dimensional bottleneck layer before reproducing the inputs. Once the network has been trained, the bottleneck node activation values give the scores.

Let $f : \Re^p \to \Re^r$ denotes the function modeled by layers 1, 2 and 3, and let $s : \Re^r \to \Re^p$ denotes the function modeled by layers 3, 4 and 5. Using this notation, the weights in the NLPCA network are determined under the following objective function:

$$\min \sum_{l=1}^{n} \left\| \vec{X}_l - \vec{X}_l' \right\| \tag{3}$$

, where $\vec{X}'$ is the output of the network. The relation 1 is now generalized to $u = f\left(\vec{X}\right)$, where f can be any nonlinear function explained by a feed-forward NN mapping from input layer to the bottleneck layer and instead of 2, $\left\| \vec{X}_l - \vec{X}_l' \right\|$ is minimized by nonlinear mapping functions, $\vec{X}' = s(u)$. The residual $\left\| \vec{X}_l - \vec{X}_l' \right\|$, can be input into the same network to extract the second NLPCA mode, and so on for the higher modes, Monahan (2000).

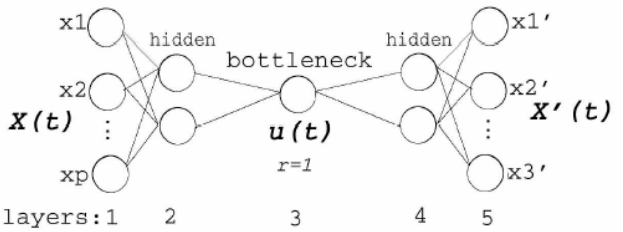

Figure 2: NLPCA: Neural Network to map Nonlinear Components.

4. C-NLPCA: Cascaded Nonlinear Principal Component Analyses

When it is necessary to run NLPCA with large dimension datasets, like images, there is a noticeable increase of parameters (weights) associated with the neurons of the NN, thus leading to the necessity to have a bigger number of temporal samples, so that this value can be near to the parameters of the NN, Hsieh (2001). It is known that sometimes it is not possible to attend this requirement even if one accepts the saturation and poor dimensional reduction risks. Moreover, the addition of more samples increases in an expressive way the computational overhead to conclude the analyses.

Thus, when the original dataset have many dimensions, several authors opt to filter the data before the NLPCA analyses, like the use of PCA reduction techniques, Botelho et al. (2003), Hsieh (2001). Using the former approach, the simplification introduced by the use of linear PCA analyses can lead to erroneous outputs or, at least, can produce coarser results.

Our C-NLPCA has the aim to allow the direct and totally nonlinear analyses of high dimension dataset, using a cascaded set of successive NLPCAs, see figure 3. The architecture is compose by two main stages: reduction and expansion. Images are decomposed into a set o small windows, which will be reduced and grouped by successive NLPCAs at reduction stage. A bottleneck NLPCA gives the final principal component. This value is expanded by the second stage (expansion stage), resulting in a output of the same dimension of the original dataset (spacial expansion).

4.1. Obtaining the C-NLPC - The Reduction Process

C-NLPCA assumes that *p'* is the ideal dimension for the input data. The ideal concept is associated with the relationship between parameters number (weights) and the number of temporal samples. Thus, we divide the original input image with dimension *p* into smaller windows with dimension *p'*. These windows are used directly as input of a first layer of NLPCAs. Each C-NLPCA, constrained by the saturation requirements, finds a local principal component (local reduction) of one window. The resulting patterns (reductions) are used as input to a new layer of

C-NLPCA. This process is repeated *n* times until only one pattern is left, thus giving the final reduction of whole original dataset. Despite during the first step the windows are independently analysed, in the second step the neighbor relations are considered, ensuing that the results are grouped in succession.

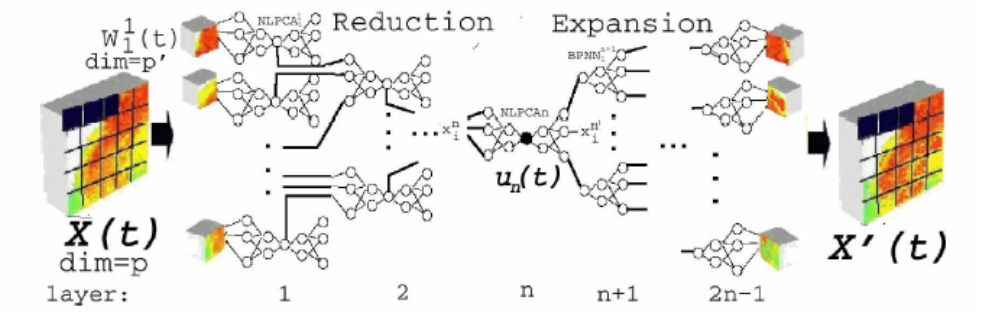

Figure 3: C-NLPCA: a set of layers grouped in Reduction and Expansion Stages. The pointed out neuron gives the pattern associated to the reduction of the original dataset.

We consider the window $vecW(t)$ as a subset of consecutive pixels of the image vector, $\vec{X}(t)$:

$$\vec{X}(t) = \left[\vec{W}_i(t) \mid 1 < i < m_1 = p / p'\right], \tag{4}$$

with:

$$\vec{W}_i(t) \cap \vec{W}_j(t) = 0, \forall i \neq j \tag{5}$$

For each component $\vec{W}_i(t)$ of $\vec{X}(t)$, we process the nonlinear analyses (section 3.2) with a standard NLPC network. Each network associated with $\vec{W}_i(t)$ is called NLPCA. They compose the first layer with *m1 = p/p'* networks. The bottleneck results $u_i(t)$ of each window $NLPC_i$ are grouped into a new second layer of NLPCAs. Notice that, the number of NLPCAs of the second layer is *m2=m1/p'*[1].

The process is repeated up to layer *n* composed by only one NLPCA network. The output of the bottleneck neuron of this only NLPCA is the nonlinear principal component, $u_i(t)$of the original large dataset.

4.2. Obtaining C-NLPCAs - The expansion Stage

The second role of the Principal Component Analyses, called expansion, is to obtain the data associated with each principal component reduction (PC, C-NLPC) in the original dimension of the image (PCA, C-NLPCA). Moreover, each time when a principal component *k* is calculated and we desire to obtain the next (*k+1*), the expansion process is also necessary to calculate the residues associated with *k*, which will be the input to the (*k+1*) C-NLPC analyses.

Hence, due to the cascading process, we have lost the original dimension of the input image, it is then necessary a method to obtain the expansion of the reduction. Expansion Stage is trained

[1] We use sub-index 1,2 to describe, respectively the first and the second layer of the cascade).

to expand the nonlinear principal component, $u_n(t)$, in a set of nonlinear expanded images. In fact, we propose a bottleneck layered structure to obtain the reduction/expansion of NLPCs.

Expansion layers are symmetric with reduction layers, resulting in a total of (2*n-1) layers. They are composed by simple backpropagation networks *BPNN* (without bottleneck neuron). The input of each *BPNN* is an output of the last layer. We use the original propagated image to train the desired outputs of *BPNN* networks.

Training BPNNs: Each resulting output x_i^n of the bottleneck NLPC layer *n* is used as input to train its respective $BPN(I^{(n+1)}$ in the next layer *(n+1)*. The desired output is the input of the respective layer *(n-1)* in reduction stage. This process is repeated up to *(2n-1)* layer, which has as desired output each pixel of original image.

Thus, the expanded image represent the original input taking into account only the current principal component. We use all components of the original dataset, their neighbors relations and temporal variabilities. The method can be applied independently of the dataset dimension size. It also maintains the nonlinearity associated with NN, avoiding the saturation restriction associated with them.

5. Applying C-NLPCA

We have tested our approach in a set of Sea Surface Temperature (SST). In order to analyze these data, researchers have generally adopted classical multivariate statistical methods, like PCA. However, related methods may produce an oversimplification of the dataset being analyzed by assuming that linear phenomena are dominant. Thus, if the data contain nonlinear lowerdimensional structure, it can not and will not be detectable by the PCA. Moreover, these images compose a large dataset, saturating and preventing the usual NLPCA methods. Thus, SST satellite images seem to be an ideal application to justify and test C-NLPCA approach.

5.1. SST Satellite Images Series

The data used are a series of three and a half years Sea Surface Temperature (SST) satellite images (from 1991 to 1994) of the southwestern Pacific Ocean. These images have been derived from the full resolution images (1 km x 1 km) recorded by the Advanced Very High Resolution Radiometer (AVHRR) on board of the National Oceanic Atmospheric Administration (NOAA) polar orbiting satellites. Data have a 9 km x 9 km spatial and 10 days temporal resolution (enough for the study of mesoscale dynamics). The dataset dimension is 60 x 60 pixels, Mata (2000).

5.2. Tests and Discussion

The dataset has (60x60) 3600 spatial variables and 128 time points. In this area, one can expect that the first 2 or 3 principal components would be enough to explain almost the totality of the data variance. Indeed, that is confirmed by computing the PCA modes from the dataset, which revealed that the eigenvalues associated to spatial modes 1, 2 and 3. Thus, we are going to search/analyse only the firsts mode 1 and 2.

The first 3 PCs (time series) computed from the dataset are shown in Figure 4. Looking at this figure, one can perceive the ample dominance of the first mode variability over the higher ones. Having a frequency of about 1 year, the first mode has a clear physical meaning in oceanography, which is related to the seasonal heating and cooling of sea surface waters following the

annual cycle of solar radiation input. The second and third modes are not that straightforward to interpret. They seem to be dominated by a higher than seasonal frequency, however, this signal seems also contaminated by a long term component. This is a clear sign of the inability of the PCA to separate in different modes signals that either are typically nonlinear, not dominant in the series or have similar energy levels (contribute equally to the total variance). The result is a blend of processes in a single mode and thus making almost impossible for one to extract any physical sense out of this mode.

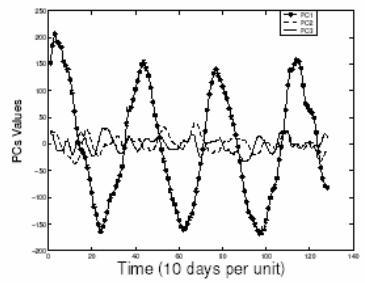

Figure 4: First three PC from the PCA: PC1 (with *), PC2 (dashed line) and PC3.

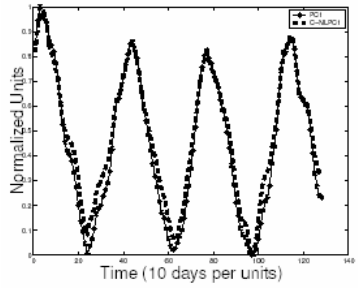

Figure 5: First PC from the PCA (with *) and from C-NLPCA (dashed line). Due to linearityof the first mode, both method present the same results

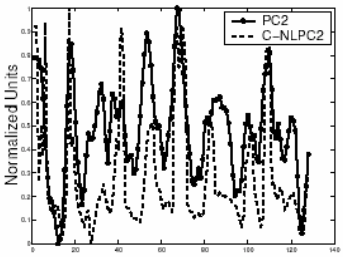

Figure 6: PC2 from the PCA (with*) and C-NLPC2 (dashed line) computed using the C-NLPCA network

C-NLPCA method was implemented in c++. To avoid local minimum, a set of 20 CNLPCAs process the same dataset. We eliminate the worst results. Each C-NLPCA is initialized with random weights, which are changed in 40.000 iterations. Each temporal image is decomposed into a subset of windows to feed C-NLPCA network (*p'=24*, NLPCAs with 2 hidden neurons). It needs 5 layers (2 for reduction and 2 for expansion).

As the solution of the first PCA mode is linear, and it dominates unquestionably the dataset variability (the seasonal oscillation pattern of the sea surface temperature), one would expect that the first principal component of the C-NLPCA network would have quite a similar pattern. Indeed, that is confirmed in Figure 5, where it is clear the excellent agreement between the PC1 and C-NLPC1 functions.

Conversely, one can expect different patterns when comparing the PCA solution for the higher modes (mode 2 for example) with the correspondent C-NLPCs.

After computing C-NLPC2, it is plotted together with PC2 in Figure 6. One can see how different they look and, taking advantage of the use of normalized units, it is clearly notable that PC2 contains a combination of a higher frequency phenomenon (about 140 days) with a lower frequency one (about 3 years). The 140 days signal should be basically composed by marine mesoscale phenomena while the 3 years one should be related to a residual of the interannual signal in the sea surface temperatures due to El Niño events, which can be quite intense in this part of the ocean. The blend of signals of such a different nature in only one PCA mode is related to the fact that those phenomena are essentially nonlinear and also contribute similarly to the total variance, hence making the PCA linear approach only a crude approximation of the observed variability..

In other hand, C-NLPC2 describes a pattern with only some isolated peaks that have frequency of about 120-180 days, and showing no clear evidence of another signal being superimposed (low frequency phenomena). The latter reinforces the hypothesis that NLPCA deals better with the higher mode variability due to an enhanced capability in isolating the signals of higher

modes, and thus suggests that C-NLPC2 may be representing a single oceanographic process. Indeed, several studies support the above assertion as they have found that besides the seasonal variability, the mesoscale dynamics is a quite important feature in that ocean area, Mata (2000). These authors also emphasize that the measoscale variability is basically due to the shedding of large eddies by the East Australian Current to the south of 33. S. During those times the Current would leave from its ìnormalî state and move about the study domain (mainly retract to the north), thus creating a sea surface temperature anomaly possible to be captured by the NLPCA analysis. The NLCP2 function depicts quite well this pattern, as it remains most of the time around the zero line and shows spikes that may well represent the eddy shedding events. The above studies about the East Australian Current also point out that the Current sheds between 2 - 4 eddies per year, but can also experience periods of lower activity, Mata (2000), pattern also matched by the NLPC2. Thus, we believe that C-NLPC is producing significantly better results than the linear PCA, further assessment of the higher modes is underway.

6. Conclusions

In the present study, we propose an original method to reduce the dimension of large image data-set, obtaining principal components of them. We use a cascaded Neural Network in a bottle-necked structured to obtain dimension reduction, giving the principal components of the data variability. The same structure is also used to expand the data from obtained principal component. The method is applied to study the mesoscale variability of an oceanic boundary current. As results, the PCA can not fully isolate those low frequency modes from others and the computation leads to time series containing more that one signal associated with distinct physical processes. On the other hand, the C-NLPCA network has demonstrated the capability of isolating the second mode of variability which seems to be related with the mesoscale variability of the oceanographic scenario, thus encouraging further investigation on others application domains.

References

[1] S. Botelho, R. de Bem, M. Mata, and I. Almeida. Applying neural networks to study the mesoscale variability of oceanic boundary currents. *Lectures Notes in Artificial Inteligence*, 2871:684-688, 2003.

[2] W. Hsieh. Nonlinear principal component analysis by neural network. *Tellus*, 53A:599-615, 2001.

[3] M. Kirby and L. Sirovich. Application of karhunen-loeve procedure for the caracterization of human faces. *IEEE On pattern analysis and machine intelligence, 1990.*

[4] S. Lek and J. Guegan. Artificial neural networks as a tool in ecological modelling an introduction. *Ecological Modelling*, 120:65-73, 1999.

[5] M. Mata. *On the mesoscale variability of the East Australian Current at subtropical latitudes.* PhD thesis, Flinders University, 2000.

[6] A. Monahan. *Nonlinear principal component analysis of climate data*. PhD thesis, University of British Columbia, 2000.

[7] R. W. Preisendorfer. *PCA in Metereology and Oceanography. Developments in Atmospherics Science*, volume 17. Elsevier, 1988.

[8] S. Romdhani, A. Psarrou, and S. Gong. Multi-view nonlinear active shape model using kernel pca. *In Tenth British Machine Vision Conference, 1999.*