ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

SOM+EOF for Finding Missing Values

Antti Sorjamaa[1], Paul Merlin[2], Bertrand Maillet[2] and Amaury Lendasse[1]

1- Helsinki University of Technology - CIS
P.O. Box 5400, 02015 HUT - Finland
2- Variances and Paris-1 University CES/CNRS - A.A.Advisors-QCG
106 bv de l'hôpital F-75647 Paris cedex 13 - France

Abstract. In this paper, a new method for the determination of missing values in temporal databases is presented. This new method is based on two projection methods: a nonlinear one (Self-Organized Maps) and a linear one (Empirical Orthogonal Functions). The global methodology that is presented combines the advantages of both methods to get accurate candidates for missing values. An application of the determination of missing values for fund return database is presented.

1 Introduction

The presence of missing values in the underlying time series is a recurrent problem when dealing with databases. Number of methods have been developed to solve the problem and fill the missing values. The methods can be classified into two distinct categories: deterministic methods and stochastic methods.

Self-Organizing Maps [1] (SOM) aim to ideally group homogeneous individuals, highlighting a neighborhood structure between classes in a chosen lattice. The SOM algorithm is based on unsupervised learning principle where the training is entirely stochastic, data-driven. The SOM algorithm allows projection of high-dimensional data to a low-dimensional grid. Through this projection and focusing on its property of topology preservation, SOM allows nonlinear interpolation for missing values.

Empirical Orthogonal Functions (EOF) [2] are deterministic, enabling linear projection to a high-dimensional space. They have also been used to develop models for finding missing data [3]. Moreover, EOF models allow continuous interpolation of missing values, but are sensitive to the initialization.

This paper presents a new methodology, which combines the advantages of both the SOM and the EOF. The nonlinearity property of the SOM is used as a denoising tool and then the continuity property of the EOF method is used to efficiently recover missing data.

2 Self-Organizing Map

The SOM algorithm is based on an unsupervised learning principle, where training is entirely data-driven and no information about the input data is required [1]. Here we use a 2-dimensional network, compound in c units (or code vectors) shaped as a square *lattice*. Each unit of a network has as many weights as the length T of the learning data samples, $\mathbf{x}_n$, $n = 1, 2, ..., N$. All units of a network

can be collected to a weight matrix $\mathbf{m}(t) = [\mathbf{m}_1(t), \mathbf{m}_2(t), ..., \mathbf{m}_c(t)]$ where $\mathbf{m}_i(t)$ is the T-dimensional weight vector of the unit i at time t and t represents the steps of the learning process. Each unit is connected to its neighboring units through neighborhood function $\lambda(\mathbf{m}_i, \mathbf{m}_j, t)$, which defines the shape and the size of the neighborhood at time t. Neighborhood can be constant through the entire learning process or it can change in the course of learning.

Learning starts by initializing the network node weights randomly. Then, for randomly selected sample $\mathbf{x}_{t+1}$, we calculate a Best Matching Unit (BMU), which is the neuron whose weights are closest to the sample. BMU calculation is defined as $\mathbf{m}_{BMU(\mathbf{x}_{t+1})} = \arg\min_{\mathbf{m}_i, i \in I} \{\|\mathbf{x_{t+1}} - \mathbf{m}_i(t)\|\}$, where $I = [1, 2, ..., c]$ is the set of network node indices, BMU denotes the index of the best matching node and $\|.\|$ is standard Euclidean norm.

If the randomly selected sample includes missing values, the BMU cannot be solved outright. Instead, an adapted SOM algorithm, proposed by Cottrell and Letrémy [4], is used. The randomly drawn sample $\mathbf{x_{t+1}}$ having missing value(s) is split into two subsets $\mathbf{x}_{t+1}^T = NM_{\mathbf{x_{t+1}}} \cup M_{\mathbf{x_{t+1}}}$, where $NM_{\mathbf{x_{t+1}}}$ is the subset where the values of $\mathbf{x_{t+1}}$ are not missing and $M_{\mathbf{x_{t+1}}}$ is the subset where the values of $\mathbf{x_{t+1}}$ are missing. We define a norm on the subset $NM_{\mathbf{x_{t+1}}}$ as

$$\|\mathbf{x_{t+1}} - \mathbf{m}_i(t)\|_{NM_{\mathbf{x_{t+1}}}} = \sum_{k \in NM_{\mathbf{x_{t+1}}}} (\mathbf{x}_{t+1,k} - \mathbf{m}_{i,k}(t))^2, \quad (1)$$

where $\mathbf{x}_{t+1,k}$ for $k = [1, ..., T]$ denotes the k^{th} value of the chosen vector and $\mathbf{m}_{i,k}(t)$ for $k = [1, ..., T]$ and for $i = [1, ..., c]$ is the k^{th} value of the i^{th} code vector.

Then the BMU is calculated with

$$\mathbf{m}_{BMU(\mathbf{x}_{t+1})} = \arg \min_{\mathbf{m}_i, i \in I} \left\{ \|\mathbf{x_{t+1}} - \mathbf{m}_i(t)\|_{NM_{\mathbf{x_{t+1}}}} \right\}. \quad (2)$$

When the BMU is found the network weights are updated as

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) - \varepsilon(t)\lambda\left(\mathbf{m}_{BMU(\mathbf{x}_{t+1})}, \mathbf{m}_i, t\right)[\mathbf{m}_i(t) - \mathbf{x_{t+1}}], \forall i \in I, \quad (3)$$

where $\varepsilon(t)$ is the adaptation gain parameter, which is $]0, 1[$-valued, decreasing gradually with time. The number of neurons taken into account during the weight update depends on the neighborhood function $\lambda(\mathbf{m}_i, \mathbf{m}_j, t)$. The number of neurons, which need the weight update, usually decreases with time.

After the weight update the next sample is randomly drawn from the data matrix and the procedure started again by finding the BMU of the sample. The recursive learning procedure is stopped when the SOM algorithm has converged.

Once the SOM algorithm has converged, we obtain some clusters containing our data. Cottrell and Letrémy proposed to fill the missing values of the dataset by the coordinates of the code vectors of each BMU as natural first candidates for missing value completion:

$$\pi_{(M_{\mathbf{x}})}(\mathbf{x}) = \pi_{(M_{\mathbf{x}})}\left(\mathbf{m}_{BMU(\mathbf{x})}\right), \quad (4)$$

where $\pi_{(M_{\mathbf{x}})}(.)$ replaces the missing values $M_{\mathbf{x}}$ of sample $\mathbf{x}$ with the corresponding values of the BMU of the sample. The replacement is done for every data sample and then the SOM has finished filling the missing values in the data.

3 Empirical Orthogonal Functions

This section presents Empirical Orthogonal Functions (EOF) [2, 5]. In this paper, EOF are used as a denoising tool and for finding the missing values at the same time [3].

The EOF are calculated using standard and well-known Singular Value Decomposition (SVD), $\mathbf{X} = \mathbf{UDV}^* = \sum_{k=1}^{K} \rho_k \mathbf{u}_k \mathbf{v}_k$, where $\mathbf{X}$ is 2-dimensional data matrix, $\mathbf{U}$ and $\mathbf{V}$ are collections of singular vectors $\mathbf{u}$ and $\mathbf{v}$ in each dimension respectively, $\mathbf{D}$ is a diagonal matrix with the singular values ρ in its diagonal and K is the smaller dimension of $\mathbf{X}$ (or the number of nonzero singular values if $\mathbf{X}$ is not full rank). The singular values and the respective vectors are sorted to decreasing order.

When EOF are used to denoise the data, not all singular values and vectors are used to reconstruct the data matrix. Instead, it is assumed that the vectors corresponding to larger singular values contain more data with respect to the noise than the ones corresponding to smaller values [2]. Therefore, it is logical to select q largest singular values and the corresponding vectors and reconstruct the denoised data matrix using only them.

In the case where $q < K$, the reconstructed data matrix is obviously not the same than the original one. The larger q is selected, the more original data, which also includes more noise, is preserved. The optimal q is selected using validation methods, for example [6].

EOF (or SVD) cannot be directly used with databases including missing values. The missing values must be replaced by some initial values in order to use the EOF. This replacement can be for example the mean value of the whole data matrix $\mathbf{X}$ or the mean in one direction, row wise or column wise. The latter approach is more logical when the data matrix has some temporal or spatial structure in its columns or rows.

After the initial value replacement the EOF process begins by performing the SVD and the selected q singular values and vectors are used to build the reconstruction. In order not to lose **any** information, only the missing values of $\mathbf{X}$ are replaced with the values from the reconstruction. After the replacement, the new data matrix is again broken down to singular values and vectors with the SVD and reconstructed again. The procedure is repeated until convergence criterion is fulfilled.

4 Global Methodology

The two methodologies presented in the previous two sections are combined and the global methodology is presented. The SOM algorithm for missing values is first ran through performing a nonlinear projection for finding the missing

values. Then, the result of the SOM estimation is used as initialization for the EOF method.

For The SOM we must select the optimal grid size c and for the EOF the optimal number of singular values and vectors q to be used. This is done using validation, using the same validation set for all combinations of the parameters c and q. Finally, the combination of SOM and EOF that gives the smallest validation error is used to perform the final filling of the data.

Even the SOM as well as the EOF are able to fill the missing values alone, the experimental results demonstrate that together the accuracy is better. The fact that these two algorithms suit well together is not surprising. Two approaches can be considered to understand the complementarity of the algorithms.

Firstly, SOM algorithm allows nonlinear projection. In this sense, even for dataset with complex and nonlinear structure, the SOM code vectors will succeed to capture the nonlinear characteristics of the inputs. However, the projection is done on a low-dimensional grid (in our case two-dimensional) with the possibility of losing the intrinsic information of the data.

EOF method is based on a linear transformation using the Singular Value Decomposition. Because of the linearity of the EOF approach, it will fail to reflect the nonlinear structures of the dataset, but the projection space can be as high as the dimension of the input data and remains continuous.

5 Experimental Results

For illustration, we use a dataset of North American fund returns[1] composed with 679 funds on a 4-year period of 219 weekly values. This gives us a dataset $\mathbf{X}$ of the size 219×679 with a total of 148 701 values. The fund return correspond to the yield of asset values between two consecutive dates as $r_t = \frac{v_{t+1}}{v_t} - 1$, where v_t is the value of the considered asset at time t.

Figure 1 shows 10 rescaled fund values $\left(v_t^{'} = 100 \prod_{i=1}^{t} (1 + r_i)\right)$. The fund values are correlated time series including first order trends. There are no missing values contained in the database.

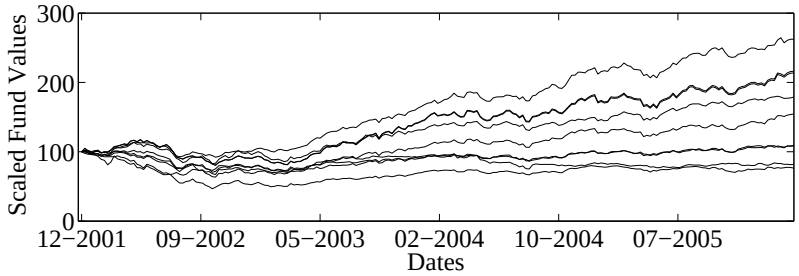

Fig. 1: Rescaled asset values of 10 funds present in the database.

[1]Data provided by Lipper, A Reuters Company.

Before running any experiments, we randomly remove 7.5 percent of the data to a test set. The test set contains 11 152 values. For validation, the same amount of data is removed from the dataset. Therefore, for the model selection and learning we have a database with total of 15 percent missing values.

The Monte Carlo Cross-Validation with 10 folds is used to select the optimal parameters for the SOM, the EOF and the SOM+EOF method. The 10 selected validation sets are the same for each method. All validation errors are shown in Figure 2. In the case of the SOM+EOF, the errors shown are minimum errors after EOF with different SOM sizes.

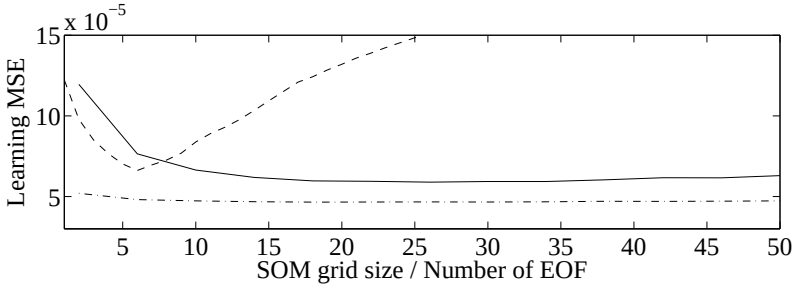

Fig. 2: Validation errors w.r.t. SOM size or number of EOF. SOM validation error (Solid line), EOF (dashed line) and SOM+EOF (dash-dotted line).

The optimal size of the SOM grid is found to be 28×28, which is a total of 784 units. Therefore, we have more code vectors in the SOM than observations (629). It means that we have nonlinear interpolation between observations and better approximation of the missing values with more units than data.

When the EOF is performed alone, initial values are substituted as the column means of the original matrix, calculated only with the known values. From the Figure 2 the smallest error with the EOF method is achieved with q equal to 6. This number of EOF is very small compared to the maximum of 219 EOF, which is the smaller dimension of the data. It suggests quite strong noise influence in the data and that there is only a small number of efficient EOF needed to represent the denoised data.

The smallest error achieved with the SOM+EOF method is with SOM grid size 27×27 and with EOF parameter q equal to 39. The number of selected EOF is much larger with the SOM initialization than with the column mean initialization. It suggests there are more efficient EOF to use in the approximation of the missing values than with the plain column mean initialization and that the SOM has already denoised the data. The optimal SOM grid size in the SOM+EOF method is found out to be roughly the same size than when performing the SOM alone. It is quite intuitive to think that the best possible filling achieved with SOM is enhanced with linear, high-dimensional projection of the EOF. From the Figure 2 it is clearly notable that with every SOM size the SOM+EOF method gives lower validation error than either SOM or EOF alone.

Table 1 contains the validation and test errors of all three methods.

Table 1: Learning and Test Errors for SOM, EOF and SOM+EOF.

10^{-5}	Learning Error	Test Error
SOM	5.83	5.57
EOF	6.61	6.13
SOM + EOF	4.63	4.34

From the Table 1, we can see that the SOM+EOF outperforms the SOM reducing the validation error by 21 percent and the test error by 22 percent. The EOF alone is not performing as well as the SOM alone.

6 Conclusion

In this paper, we have compared 3 methods for finding missing values in temporal databases. The methods are Self-Organizing Maps (SOM), Empirical Orthogonal Function (EOF) and the combination of the two SOM+EOF.

The advantages of the SOM include the ability to perform nonlinear projection of high-dimensional data to lower dimension with interpolation between discrete data points. For the EOF, the advantages include high-dimensional linear projection of high-dimensional data and the speed and the simplicity of the method. The SOM+EOF includes the advantages of both individual methods, leading to a new accurate approximation methodology for finding the missing values. The performance obtained in test show the accuracy of the new methodology.

It has also been shown experimentally that the optimal number of code vectors used in the SOM has to be larger than the number of observations. It is necessary in order to take the advantage of the self-organizing property of the SOM and the interpolation ability for finding the missing data.

For further work, the modifications and performance upgrades for the global methodology are fine-tuned for different types of datasets. The methodology will then be applied to datasets from climatology.

References

[1] Teuvo Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[2] R. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.

[3] J. Boyd, E. Kennelly, and P. Pistek. Estimation of eof expansion coefficients from incomplete data. *Deep Sea Research*.

[4] Marie Cottrell and Patrick Letrémy. Missing values: Processing with the kohonen algorithm. pages 489–496. Applied Stochastic Models and Data Analysis, Brest, France, 17-20 May, 2005.

[5] SOM+EOF Toolbox: http://www.cis.hut.fi/projects/tsp/?page=Downloads.

[6] Amaury Lendasse, V. Wertz, and Michel Verleysen. Model selection with cross-validations and bootstraps - application to time series prediction with rbfn models. In *LNCS*, number 2714, pages 573–580, Berlin, 2003. ICANN/ICONIP (2003), Springer-Verlag.