Neural Processing Letters, Vol. 2, No. 4, 6-10, 1995

Combining Singular-Spectrum Analysis and neural networks for time series forecasting

F. Lisi[1], O. Nicolis[2], Marco Sandri[2]

[1] *Dipartimento di Scienze Statistiche, Università di Padova*
Via San Francesco 33, I-35121 Padova, Italy
[2] *Istituto di Scienze Economiche, Università di Verona*
Via dell'Artigliere 19, I-37129 Verona, Italy

Abstract. In this paper, we propose a combination of an adaptive noise-reduction algorithm based on Singular-Spectrum Analysis (SSA) and a standard feedforward neural prediction model. We test the forecast skill of our method on some short real-world and computer-generated time series with different amounts of additive noise. The results show that our combined technique has better performances than those offered by the same network directly applied to raw data, and therefore is well suited to forecast short and noisy time series with an underlying deterministic data generating process (DGP).

1. Introduction

For many years the field of time series forecasting has been largely analysed by statisticians (see e.g. [1]) and, only recently, has developed as an area of increasing importance for nonlinear dynamical system and artificial intelligent theory [2].

In this paper we address the difficult and challenging task of predicting short and noisy time series. Basically, the algorithm which we propose in this paper consists of two different steps: the preprocessing of data, based on the SSA filtering method (recently proposed by [3], and generalized by [4] and [5]) followed by the real forecasting step, in which we use a simple feedforward multilayer neural network with backpropagation learning. If the noise-reduction procedure successfully removes most of the noise, then the processed data set should have better short- and long- term predictability than the raw data set.

We compare the performances of our approach to those of an analogous neural network predictor trained with raw data. As benchmarks we have chosen a 'classic' chaotic time series, to which we have added two different amount of noise, and the SOI (Southern Oscillation Index) geophysical time series (see [6]). In section 2 we consider the basic assumptions of our approach, while in section 3 we shortly explain the theoretical background of singular-spectrum analysis (SSA) and the projection-reconstruction method for dynamics. In section 4 we show the results obtained from our simulations. A brief conclusion is given in 5.

2. Basic assumptions

Let us denote by Z_t the data generating processing (DGP) and let $\{z_t\}_{t=1...N}$, $z_t \in \Re^l$, be the series of N observations. We assume that $Z_t = Y_t + W_t$, where Y_t is a deterministic dynamics (i.e. governed by an evolutionary deterministic law), and W_t is an uncorrelated stochastic component (white noise). In reality, however, we do not necessarily have direct and complete access to all the state variables of the underlying deterministic system. We typically observe only one or few variables through a certain (deterministic) 'viewer' $h : \Re^m \rightarrow \Re^l$, that is $Y_t = h(X_t)$, where $dX/dt = F(X_t)$, $F : \Re^m \rightarrow \Re^m$, is the 'true' unknown dynamical system. Under weak regularity conditions on h and F, and for sufficently large positive integer d ($d \geq 2m + 1$), the Takens' theorem (see [7] and the multidimensional corollary in [8]) guarantees that the time evolution of the reconstructed vector $Y_t^d = (Y_t, \ ..., \ Y_{t-d+1})$ is diffeomorphically equivalent to the true dynamics X_t. Implicitly, this means that there exists a d-dimensional deterministic map G^d: $\Re^{d.l} \rightarrow \Re^{d.l}$, in general dependent on d, such that $Y_{t+1}{}^d = G^d(Y_t^d)$, or equivalently there exists a map $g^d : \Re^{d.l} \rightarrow \Re^l$ such that:

$$y_{t+1} = g^d(y_t, \ldots, y_{t-d+1}). \qquad (1)$$

For our forecasting purposes, under the given assumptions, as W_t is not clearly predictable, we should concentrate our attention only on the Y_t component and try to approximate as better as possible the map g^d. This, in turn, requires to preprocess Z_t with a 'good' filtering method which allow us to eliminate the noise W_t, and hence to get good estimates of the Y_t dynamics.

3. An overview of singular-spectrum analysis

For the sake of simplicity, we consider here only the unidimensional case $l = 1$, since the generalization to $l > 1$ is straightforward (see [5]). The first step in the implementation of single-channel SSA (S-SSA) is to construct the so-called 'trajectory matrix' in the d-dimensional reconstructed space:

$$Z = \frac{1}{\sqrt{N}} \begin{bmatrix} z_1 & \cdots & z_d \\ z_2 & \cdots & z_{d+1} \\ \vdots & & \vdots \\ z_{N-d+1} & \cdots & z_N \end{bmatrix} = \frac{1}{\sqrt{N}} \begin{bmatrix} z_1^d \\ z_2^d \\ \vdots \\ z_{N-d+1}^d \end{bmatrix} \tag{2}$$

Now let be $\bar{N} = N - d + 1$. Using the well known singular value decomposition (SVD), the $(\bar{N} \times d)$ matrix Z can be decomposed as $Z = S\Sigma C^T$, where S is a $(\bar{N} \times d)$ matrix whose columns s_i are the eigenvectors of ZZ^T; C is a $(d\times d)$ matrix whose columns c_i are the eigenvectors of Z^TZ; Σ is a $(d\times d)$ diagonal matrix whose elements σ_i are the positive square roots of the eigenvalues of Z^TZ. The vectors s_i and c_i are also called 'singular vectors' and the σ_i are called 'singular values' of Z. A moment's reflection will show that the matrix Z^TZ is the covariance matrix of the series $\{z_t\}$.

The singular vectors are orthogonal, i.e. $c_i^Tc_j = s_i^Ts_j = \delta_{ij}$. Thus, the vectors c_i (called 'empirical orthogonal functions', EOFs) can be used as an orthonormal basis of the space $\Re^d$ on which d-dimensional points z_i^d can be projected. The columns of the matrix $A \equiv ZC$ are called 'principal components' (PCs) and represent the coefficients of projection of vectors z_i^d onto the EOFs. Finally, the singular values σ_i can be ordered as $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d \geq 0$.

As we have just said, the matrix ZC represents the trajectory matrix projected onto the basis c_i. If we think of a trajectory as exploring on average an ellipsoid of dimension d, the vectors c_i correspond to the directions of the principal axes of the ellipsoid, and the values σ_i associated with them correspond to the lengths of those axes. Each σ_i^2 can be looked at as the variance of the i-th principal component.

Then, recalling our basic assumption that $\{z_t\}$ is the result of a deterministic DGP contaminated by an observational white noise, if the series is stationary and sufficiently long, the perturbed covariance matrix Z^TZ can be approximated as follows:

$$Z^TZ = Y^TY + (\sigma_w^2/d)\, \mathrm{I} \tag{3}$$

where σ_w^2 denotes the variance due to noise, I is the $(d\times d)$ identity matrix and Y is the trajectory matrix of the deterministic part Y_t. Clearly, in this case the singular vectors c_i of the perturbed trajectory matrix will be the same as those of the unperturbed one, whereas, owing to noise, the singular values will be uniformly increased by an amount σ_w^2/d.

For any given value of the window length d and the level of noise σ_w^2, the signal-to-noise ratio (s/n) associated with each direction - measured by the quantity $d\sigma_i^2/\sigma_w^2$ - will decrease as the order of the corresponding singular value increases and clearly noise can entirely dominate the signal for 'higher order directions'. While, in general, the singular values of a noise-free series will be uniformely declining with the order i, in the presence of sufficiently strong white noise we should observe a plateau - the 'noise floor' - in the spectrum. If this is the case, only $p \leq d$ singular values will be above the noise floor.

Thefore, the basic idea is to consider only the first p principal directions, because they describe the largest fraction of the total variance that one can obtain using a projection onto p orthogonal vectors, and then to go back to the 'filtered' trajectory matrix $\tilde{Z}$ (and hence to the filtered series $\{\tilde{z}_t\}_{t=1}^N$) by the relation $\tilde{Z} = \tilde{A}\tilde{C}^T$where $\tilde{A}$ and $\tilde{C}$ are the 'truncated' versions of A and C respectively, obtained by dropping the last d-p columns from the original matrices.

In this process of reconversion one can easily recognize that the values assigned to elements of the series $\{\tilde{z}_t\}$ with the same index will generally be different, that is, the filtered series $\{\tilde{z}_t\}$ is not unique. In order to overcome this problem we follow the procedure suggested by [3] p. 105, and find a new series $\{u_t^p\}$ which is the closest, in the sense of least squares, to the d different reconverted series $\{\tilde{z}_t\}$.

Our main problem is of course the optimal choice of the p directions which allow to eliminate as much as possible the noise, without 'deforming' the underlying low-dimensional dynamics. Some authors [8] suggest to take p as the number of singular values above the noise floor, while others [3] propose to evaluate p by a method based on resampling technique. However, we prefer to adopt the '*best prediction*' method developed by [9] since it is strictly connected to our forecasting problem:

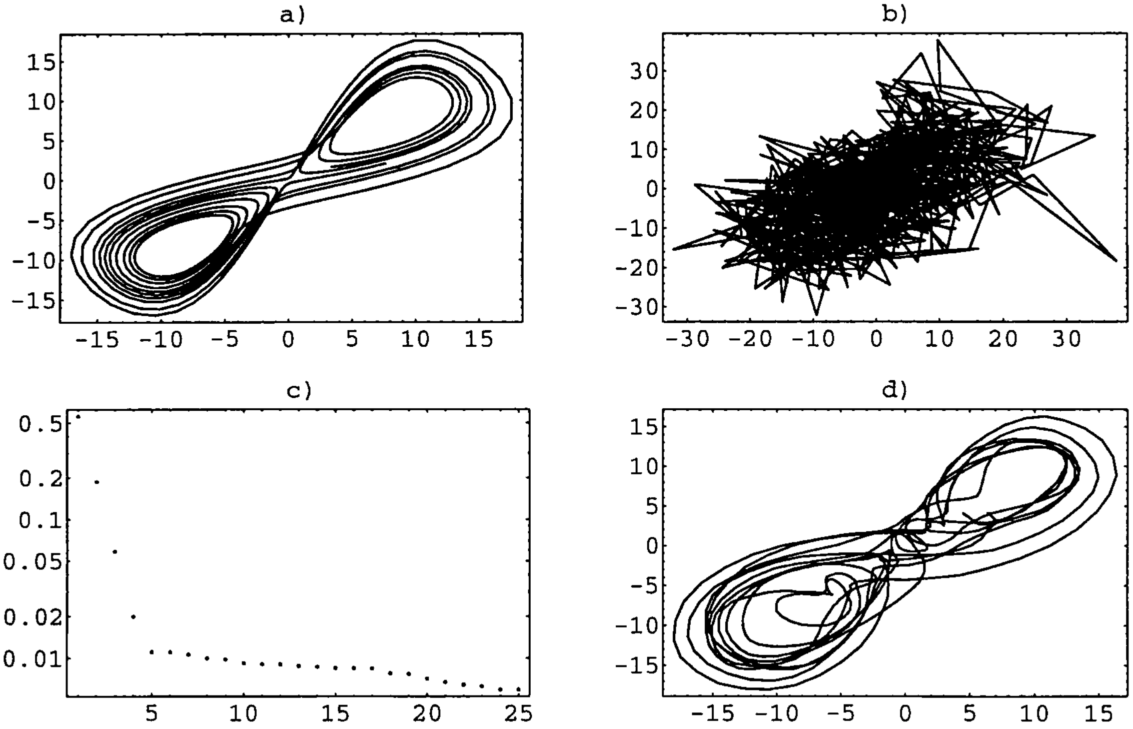

Fig. 1. a) 2-D phase portrait of a noise-free time series generated by the Lorenz model. b) The same trajectory with an additive white noise (n/s = 50%). c) The singular spectrum (M = 25). d) The filtered series (p = 4).

starting from the time series under analysis, we first build N_C training sets of length N_T ($N_T + N_C = N$). Then, for all p ($p = 1, \ldots, d$) and for all the N_C subsets, we filter the data using the first p PCs and perform T-step ahead neural forecasts $\hat{z}_i$, $i = N_T + 1, \ldots, N$. Finally, we evaluate the normalized mean square prediction error (NMSE) given by (see [2], p. 63):

$$NMSE = \frac{\sum_{i=N_T+1}^{N} (\tilde{z}_i - \hat{z}_i)^2}{\sum_{i=N_T+1}^{N} (\tilde{z}_i - \bar{z})^2}, \tag{4}$$

where

$$\bar{z} = \frac{1}{N_C} \sum_{i=N_T+1}^{N} \tilde{z}_i\,, \tag{5}$$

and choose the number p^* corresponding to the minimum NMSE.

4. Empirical results

We start our tests considering a time series of 650 observations (the first 400 as sample values and the following 250 as testing set) obtained by the integration of the well-known Lorenz chaotic model, contaminated by adding different amounts (50%, 100%) of white noise, measured in terms of the ratio between noise and signal variance. Using the sample data, we build 150 training sets containing 250 elements each. The forecasting model here adopted is a single hidden-layer neural network with 25 inputs, 30 hidden units, and 15 outputs.

Tables 1 and 2 show the results we have obtained on the two Lorenz series (n/s ratio = 50% and 100%, respectively), for the different prediction steps T ($T = 1, \ldots, 15$). Note that the optimal number p^* of directions, written in the first row of each table, decreases as the prediction step increases. This means that, in order to forecast the more distant values of the series, it is necessary to retain only the first PCs, that is the underlying trend. In the second and third rows, we list the normalized mean square prediction error corresponding to the different steps T, reached by our combined method (NMSE$_{FILT}$), and by the same neural network applied to the raw data (NMSE$_{RAW}$). The last row (REDUCT. %) shows the percentage error reduction obtained with our method respect to NMSE$_{RAW}$. It is worth pointing out that the improvement is more consistent in the higher noise case.

	Prediction steps														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
p^*	4	4	3	3	3	3	3	3	3	3	3	3	3	2	2
NMSE_{FILT}	.257	.265	.293	.321	.340	.356	.354	.386	.408	.419	.466	.520	.579	.633	.650
NMSE_{RAW}	.362	.311	.347	.339	.329	.378	.389	.436	.469	.484	.487	.661	.652	.715	.783
REDUCT.%	41	18	19	6	-3	7	10	13	15	16	5	28	13	13	21

Table 1. *Noisy data from Lorenz model - n/s = 50%. Out-of-sample NMSE for filtered and raw data.*

	Prediction steps														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
p^*	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1
NMSE_{FILT}	.667	.675	.709	.743	.784	.850	.880	.903	.970	1.00	1.01	1.03	1.07	1.11	1.15
NMSE_{RAW}	1.13	1.33	1.41	1.47	1.53	1.89	2.09	2.03	1.93	2.18	1.96	2.11	2.40	2.17	2.35
REDUCT.%	42	50	50	50	49	56	58	56	50	55	49	52	56	49	52

Table 2. *Noisy data from Lorenz model - n/s = 100%. Out-of-sample NMSE for filtered and raw data.*

	Prediction steps														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
p^*	3	3	3	3	3	3	3	1	1	1	1	1	1	1	1
NMSE_{FILT}	565	.639	.693	.748	.833	.907	.997	1.04	1.06	1.08	1.08	1.09	1.10	1.10	1.10
NMSE_{RAW}	.884	1.22	1.21	1.33	1.51	1.50	1.68	2.17	2.43	2.15	1.94	2.06	2.22	2.53	2.21
REDUCT.%	37	48	43	44	45	40	41	53	57	50	45	48	51	57	51

Table 3. *SOI index. Out-of-sample NMSE for filtered and raw data.*

Table 3 shows the outcomes of our simulation for the SOI real-world series composed by 588 observations (the first 400 as sample values and the remaining 188 as testing set).

5. Conclusion and further work

Our preliminary analysis displays that the method we have developed, based on the combination of the projection-reconstruction algorithm of [3] with a very simple neural network, is able to offer significant improvements for short- and long-term predictions on series which are relatively short and very noisy. In order to further improve our findings, we are going to concentrate our future efforts in two directions:

- using more sophisticated neural architectures, e.g. the Time-Delay Neural Networks of [10], which give accurate long term predictions of noiseless chaotic time series;
- using more sophisticated filtering techniques, e.g. the Multi-Channel SSA discussed in [5].

References

[1] S. Makridakis, M. Hibon. Accuracy of forecasting: an empirical investigation. *Journal of the Royal Statistical Society A*, vol. 142, pp. 97-145, 1979.

[2] A.S. Weigend, N.A. Gershenfeld (eds.). *Time Series Prediction. Forecasting the Future and Understanding the Past,* Addison-Wesley Publ. Co., Reading, MA, 1994.

[3] R. Vautard, P. Yiou, M. Ghil. Singular-spectrum analysis: a toolkit for short, noisy and chaotic signals, *Physica D*, vol. 58, pp. 95-126, 1992.

[4] C.L. Keppenne, M. Ghil. Adaptive filtering and prediction of noisy multivariate signals: an application to subannual variability in atmospheric angular momentum, *Internation Journal of Bifurcation and Chaos*, vol. 3 (3), pp. 625-634, 1993.

[5] F. Lisi, A. Medio, M. Sandri. Noise filtering with multi-channel singular-spectrum analysis, *Working paper No. 94.21 University of Venice*}, Ca' Foscari, Italy, 1995.

[6] C.L. Keppenne, M. Ghil. Adaptive filtering and prediction of the Southern Oscillation Index, *Journal of Geophysical Research*, vol. 97 (D18), pp. 449-454, 1992.

[7] T. Sauer, J.A. Yorke, M. Casdagli. Embedology,

[7] T. Sauer, J.A. Yorke, M. Casdagli. Embedology, *Journal of Statistical Physics*, vol. 65 (3/4), pp. 579-616, 1991.

[8] D.S. Broomhead, D.S., G.P. King. On the qualitative analysis of experimental dynamical systems, in: *Nonlinear Phenomena and Chaos*, S. Sarkar (ed.), Adam Hilger, Bristol, pp. 113-144, 1986.

[9] F. Lisi. Statistical dimension estimation in singular spectrum analysis. *Working Paper No. 1994.10, Dept. of Statistics, University of Padua*, Italy, 1994.

[10] E.A. Wan. Time series prediction by using a connectionist network with internal delay lines, in: *Time Series Prediction. Forecasting the Future and Understanding the Past*, A.S. Weigen, N.A. Gershenfeld (eds.), Proceedings Vol. XV, Santa Fe Institute, Addison Wesley, pp. 195-217, 1994