

Overembedding Method for Modeling Nonstationary Systems

P. F. Verdes,¹ P. M. Granitto,² and H. A. Ceccatto³

¹Heidelberg Academy of Sciences, c/o Institute of Environmental Physics, Im Neuenheimer Feld 229, D-69120 Heidelberg, Germany

²Istituto Agrario di San Michelle a/A, Via E. Mach 2, I-38010 San Michelle a/A, Italy

³Instituto de Física Rosario, CONICET and Universidad Nacional de Rosario,
Boulevard 27 de Febrero 210 Bis, S2000EYP Rosario, Argentina

(Received 9 February 2005; published 24 March 2006)

We propose a general overembedding method for modeling and prediction of nonstationary systems. It basically enlarges the standard time-delay-embedding space by inclusion of the (unknown) slow driving signal, which is estimated simultaneously with the intrinsic stationary dynamics. Our method can be implemented with any modeling tool. Using, in particular, artificial neural networks, its application to both synthetic and real-world time series shows that it is highly efficient, leading to much more accurate results and longer prediction horizons than other existing overembedding methods in the literature.

DOI: 10.1103/PhysRevLett.96.118701

PACS numbers: 05.45.Tp, 84.35.+i, 89.75.Hc, 95.75.Wx

Real-world dynamical systems are often too complex to be modeled from first principles. In such cases, an alternative path is to collect a set of observations $\{x_t\}_{t=1}^N$ at regular intervals of time, and estimate the function f that is supposed to generate the data via the equation $x_{t+\tau} = f(\mathbf{x}_t)$, where $\mathbf{x}_t = (x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau})$ for some properly chosen τ and d [1]. The standard approach proposes some parameterized model for f , implicitly assuming that these parameters remain constant during the measurement process. However, most real-world time series have some degree of nonstationarity due to external perturbations or internal changes of the system. Furthermore, natural dynamics are often complex enough to comprise multiple time scales, so that for short observational periods the effective degrees of freedom with the largest scales act as external perturbations for the fastest observed modes. In spite of this, the vast literature on linear and nonlinear time-series analysis [2] mostly relies on the stringent condition of stationarity.

In recent years, however, an increasing effort has been devoted to extend delay-embedding ideas to cope with nonstationarity [3,4]. In [3], the proposed solution is to enlarge the pseudo phase space by including further lagged copies of the measured data. In [4], t is explicitly incorporated into the description of the system to encompass time-dependent dynamics. In this Letter, we overembed the system according to the natural description $x_{t+1} = f(\mathbf{x}_t, \alpha_t)$ [5], where α_t accounts for the nonstationary effects. However, unlike [5], we do not assume that the signal α_t is known *a priori*. Instead, we propose a very general algorithm—that we term α overembedding—which estimates α *simultaneously* with f [6]. Here we show that, besides providing a more natural decomposition into intrinsic dynamics and external component, this approach yields a remarkable improvement in modeling performance in comparison with other embedding strategies in the literature.

A general formulation of the α -overembedding algorithm is the following: (i) consider the data set \mathcal{D} whose

elements are patterns (\mathbf{x}_i, x_{i+1}) built from the original record, and randomly split it in disjoint learning \mathcal{L} and validation \mathcal{V} sets. (ii) Use *any* modeling approach to learn from \mathcal{L} a global stationary model $f(\bullet, \alpha = 0)$, with the extra input α momentarily set to zero. (iii) Switch on α starting from $\alpha = 0$, and readjust $f(\bullet, \alpha)$ by minimizing with respect to f 's internal parameters *and* α_t the error

$$E_{\mathcal{L}} = \sum_{i \in \mathcal{L}} [x_{i+1} - f(\mathbf{x}_i, \alpha_i)]^2 + \lambda \sum_{i \in \mathcal{L}} (\alpha_{i+1} - \alpha_i)^2.$$

Here the first term is the standard mean square error (MSE) of model $f(\bullet, \alpha)$. The second term enforces the basic assumption of a smooth α behavior by penalizing strong fluctuations [on $O(\tau)$ scales] of this quantity. The hyperparameter λ fixes the appropriate scale between both terms and can be determined as indicated below. (iv) While learning $f(\bullet, \alpha)$, monitor the behavior of

$$E_{\mathcal{V}} = \frac{1}{2} \sum_{j \in \mathcal{V}} \{ [x_{j+1} - f(\mathbf{x}_j, \alpha_{j-})]^2 + [x_{j+1} - f(\mathbf{x}_j, \alpha_{j+})]^2 \}.$$

Since for patterns j in \mathcal{V} the inputs α_j are unknown, $\alpha_{j\pm}$ indicate here the α values in the (past and future) nearest-neighbor points of j in \mathcal{L} . (v) Stop the learning process for $f(\bullet, \alpha)$ at $E_{\mathcal{V}}^{\min}$, the minimum of $E_{\mathcal{V}}$, which corresponds to the optimal model's generalization capability. (vi) For the determination of λ , repeat the above steps for different values of this hyperparameter and choose $\lambda_{\text{opt}} = \text{argmin}_{\lambda} E_{\mathcal{V}}^{\min}(\lambda)$.

Notice that for large λ values, $E_{\mathcal{V}}^{\min}(\lambda)$ reduces to the error of the stationary $f(\bullet, \alpha = 0)$ model, which is non-optimal because of the nonstationary character of the record. On the other hand, for very small λ the inputs α_j become highly fluctuating, producing a perfect fitting to targets in \mathcal{L} . In this last case, the generalization performance $E_{\mathcal{V}}^{\min}(\lambda)$ will again deteriorate since $\alpha_{j\pm}$ will be completely uncorrelated with the correct α_j value (we

are assuming, as usual, that targets are corrupted by random noise). In between, there is a range of λ that leads to relatively smaller $E_{\mathcal{V}}^{\min}(\lambda)$ and, consequently, a better modeling of the nonstationary dynamics.

To test the performance of this algorithm we will first apply it to a controlled synthetic problem, the forced logistic map with observational noise: $y_{t+1} = r_t y_t (1 - y_t)$, $x_t = y_t + \varepsilon_t$. Here ε_t is a zero-mean Gaussian additive noise and the forced parameter $r_t = r_0 + C \cos(2\pi t/T) \times \exp(-2\pi t/T)$. A constant displacement $r_0 = 3.8$ was chosen as to keep the system in the interesting chaotic regime, while the force strength $C = 0.045$ was taken as large as possible without collapsing the attractor to a trivial structure nor producing divergent behavior. The extra input α is then equivalent to r up to this amplitude and the displacement r_0 . We have set $T = N/2$, so that the profile of this driving parameter is the same independently of the record length considered ($N = 100, 500$, and 1000 in this study). The noise level was characterized in terms of the ratio between the standard deviations of noise and map forcing, and set to four different values: $\sigma_\varepsilon/\sigma_r = 0, 1, 5$, and 10% .

For each noise level and record length we have trained feedforward artificial neural networks (ANNs), mapping inputs $X_t \equiv (\mathbf{x}_t, \alpha_t)$ to targets x_{t+1} [7]. The only novelty here is that the ANNs are also trained with respect to the extra input α . Regarding the hyperparameter λ , we considered 8 different values spanning 4 orders of magnitude to find its optimal range. In all cases we have used tenfold validation; the 10 models so generated were later evaluated on 10 different test sets containing also N patterns and generated like \mathcal{D} but from different seeds y_0 . In this way, the results presented next are averages over 100 different experimental realizations. Notice that the ANN modeling can be further optimized (for instance, by choosing the best suited architecture for each noise level and record length). Consequently, these results are not intended to be the most accurate ones one can obtain; they only exemplify the efficacy of the α -overembedding approach to treat nonstationary time series. Furthermore, they provide a basis for comparison with other overembeddings proposed in the literature [3,4]. For completeness, in addition to these two other methods we have also considered two simple treatments: (i) ignoring the nonstationarity altogether and (ii) using a restricted data set with a limited number of past records to produce a local model less affected by the nonstationary forcing. In this last case we proceeded as follows: first, we split the data set \mathcal{D} in 10 parts of $N/10$ data points. Second, an optimal model for each interval was generated using the last M points in the record previous to the interval under consideration. The ANN training and the optimization of M were validated using points in the same interval being modeled [8]. Third, once the optimal value of M for each interval was determined, 10 different ANNs were trained on this restricted data set and

the corresponding models tested on the equivalent intervals in the 10 test sets. Thus, the results obtained with this approach are again averages over 100 realizations.

The main interest in this work is to appraise the α -overembedding performance in modeling the local dynamics. For this, we estimated $E_{\mathcal{T}}$, the normalized MSE between predicted and observed values in the test set, for all the above discussed methods. In Fig. 1 we present the average results for different combinations of noise level and record length (results for $N = 500$, not shown, are very similar to those for $N = 1000$). As this figure shows, in all cases the α overembedding produces the most accurate modeling of the forced logistic dynamics, with errors relatively close to the optimal one (residual error given by the noise-to-signal variance ratio). In particular, in the noiseless case our approach is more than 1 order of magnitude better than the second best method. The error bars included in Fig. 1 indicate that the performance gain is independent on the particular realization. A further comparison of the robustness of the obtained models is provided by the iterated n -step-ahead predictions. These results are given in Fig. 2 for $N = 100$ and no noise, and show that the error gain of more than 1 order of magnitude persists until $n = 10$, where all other algorithms are already producing useless predictions [$E_{\mathcal{T}}(n = 10) \simeq 1$]. Alternatively, for any predetermined error level our approach is able to extend the prediction horizon nearly 3 steps ahead.

A further test on the validity of the α overembedding is provided by the normalized MSE between the original and reconstructed forces, $E_\alpha = (1/\sigma_r^2) \sum_{t=1}^N (r_t - \alpha_t)^2$. Here σ_r^2 is the forced-parameter variance and α is the reconstructed profile, conveniently rescaled to the same mean and variance of r . The corresponding results for E_α are given in Table I; they are highly satisfactory except perhaps for the worst scenario (largest noise and fewest points in the data set). We stress, however, that for $N = 100$ one is pushing the method to its limits, since the basic assumption of a slow α change along the record is hardly satisfied. For

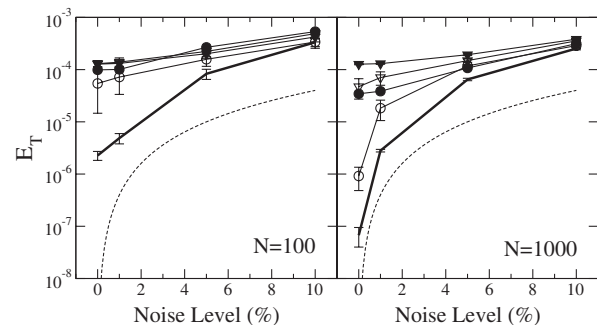


FIG. 1. Average test-set error $E_{\mathcal{T}}$ as a function of noise level. Full triangles: ignoring nonstationarity; open triangles: incorporating t as input [4]; full circles: local modeling; open circles: incorporating further lagged coordinates [3]; thick full line: α overembedding; dashed line: noise residual error.

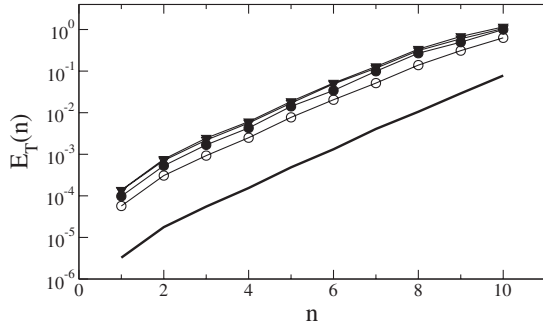


FIG. 2. Average n -step-ahead prediction error $E_{\mathcal{T}}(n)$ as a function of n for the logistic map and $N = 100$. Symbols follow the same convention as in Fig. 1.

this extreme situation these results also compare favorably with similar results obtained in [9] using an entirely different and very efficient reconstruction method. For instance, for the noise-free case the value $E_{\alpha} = 0.017$ obtained in this work should be compared with $E_{\alpha} = 0.034$ produced by the method in [9]. The advantage here is the use of a *global* strategy to estimate α , while the method in [9] was based on developing a series of local models.

In view of the excellent results obtained for the logistic map, two questions naturally arise: (i) is the sensible gain in $E_{\mathcal{T}}$ shown in Fig. 1 obtainable in other cases? and (ii) how much would a poor modeling of the intrinsic dynamics affect the α reconstruction and, consequently, the performance of the proposed method? In partial answer to these questions we have studied the Hénon map $x_{t+1} = 1 - 1.4x_t^2 + b_t x_{t-1}$, with the parameter b_t forced like in the logistic map case. We considered $N = 100$ (to be in the worst situation of scarcity of data) and trained ANNs with $h = 10$ hidden units. A study similar to the one performed before leads again to an error gain with respect to other overembedding methods of nearly 1 order of magnitude (left panel of Fig. 3). Regarding the effects of using poor modeling techniques, we considered less flexible ANNs by reducing the number of hidden units from $h = 10$ to 2. In the right panel of Fig. 3 we plot the ratio $e(h) = E_{\mathcal{T}}(h)/E_{\mathcal{T}}(h = 2)$ for both the stationary and nonstationary modeling approaches in the noiseless case, and also the reconstruction error E_{α} . These results show that the α -overembedding performance degrades gracefully when the modeling technique cannot completely capture the complexity of the local dynamics, until in the worst case ($h = 2$) no coherent smooth driving force can be recon-

TABLE I. Average error E_{α} in the reconstruction of the driving parameter r_t for the noisy logistic map.

Noise level	0%	1%	5%	10%
$N = 100$	0.017	0.013	0.088	0.191
$N = 500$	0.002	0.003	0.013	0.031
$N = 1000$	0.001	0.002	0.015	0.025

structed ($E_{\alpha} = 1$) and the results become equivalent to those of the stationary approach. Furthermore, we have checked that for the case of misspecified models—for instance, a single input x_t instead of (x_t, x_{t-1}) for the Hénon map—no driving force can be reconstructed and the method becomes again equivalent to the stationary approach. The same overall results can be obtained for the notoriously more complex problem of modeling the Mackey-Glass equation. Details on these investigations will be given elsewhere.

As a real-world application of the α -overembedding approach we have considered the modeling of the sunspot number (SSN) record, a standard benchmark in time-series studies. In this case we have proceeded as follows: first, the 1700–2003 annual mean SSN record was optimally embedded in a time-delayed space with $\mathbf{x}_t = (x_t, x_{t-1}, x_{t-2}, x_{t-3}, x_{t-8}, x_{t-9})$ [10]. Then, we generated 100 different test sets containing 70 randomly selected patterns (\mathbf{x}_t, x_{t+1}) ($\sim 25\%$ of the data). In each case, with the remaining data we trained ensembles of 20 ANNs, both with and without the extra input α [11]. Without considering possible SSN nonstationarities (i.e., not using α_t), the mean $E_{\mathcal{T}}$ obtained averaging over the 100 test sets was 0.112 ± 0.027 ; with the α overembedding here proposed we obtained $E_{\mathcal{T}}(\alpha) = 0.076 \pm 0.012$. Plotting as a histogram the relative gains $\eta = [E_{\mathcal{T}} - E_{\mathcal{T}}(\alpha)]/E_{\mathcal{T}}$ one obtains a rather flat distribution with 90% of the counts falling in the interval [10%, 50%] and with a mean value $\bar{\eta} \simeq 30\%$ [13]. We are not aware of existing approaches in the vast literature on SSN capable of such impressive error reduction, particularly for an already very competitive modeling technique like ensembles of ANNs.

We have also considered iterated predictions for the SSN record. For this, we randomly selected 10 out of the 100 ensemble models already built, and iterated its predictions n steps ahead along the whole record. In this way, the average results obtained depend only on n and not on the starting prediction point. Figure 4 shows the obtained

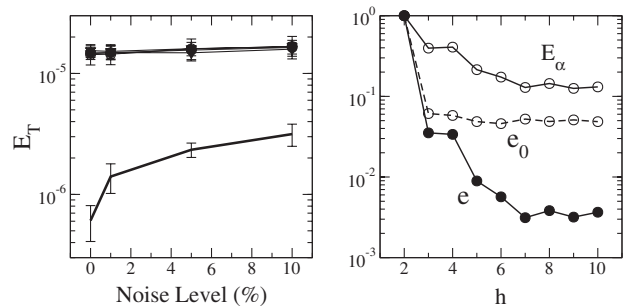


FIG. 3. Left panel: average test-set error $E_{\mathcal{T}}$ as a function of noise level for the Hénon map. Symbols follow the same convention as in Fig. 1. Right panel: normalized test-set errors for the stationary (e_0) and nonstationary (e) approaches as a function of h , the number of ANN hidden units. Also shown is the α reconstruction error E_{α} .

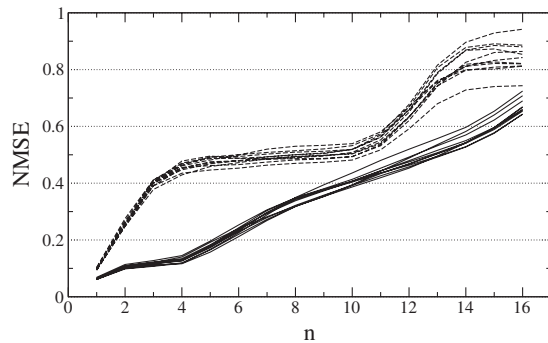


FIG. 4. Normalized MSE of 10 different models in n -step-ahead predictions for the SSN record. Stationary modeling: dashed lines; α -overembedding method: full lines.

NMSE as a function of n , again both with and without the extra input α . While the results for the stationary modeling accuse some structure related to the cycle length, the α -overembedding approach is able to capture the intrinsic characteristic of the SSN dynamics, producing an error that grows almost linearly with n . More importantly, the error level corresponding to $n = 3$ for the stationary model is reached in the nonstationary approach for a prediction horizon of nearly a whole cycle. Notice that in this study we have used the reconstructed α_t as input to the ANN for n -step predictions into the future, something that would not be possible in actual forecasting. However, α_t is in practice smooth enough to allow being simply extrapolated for time spans of at most a cycle length and then used as input without producing sensible errors in the SSN predictions.

A deeper study of errors, which cannot be reproduced here due to space limitations, shows that in a 4-step-ahead prediction the NMSE of a stationary model concentrates essentially in the ascendant phase of the (short) solar cycles. This localized distribution indicates the difficulties in predicting the early cycle behavior from the descendent part of the previous one, and also confirms the common observation that the maximum activity can be accurately predicted once the early solar cycle behavior is known. Instead, for the nonstationary model the errors are more uniformly distributed along the whole cycle, which supports the contention that the α -overembedding approach provides a more suitable modeling framework. This analysis and concrete predictions for the remaining part of the 23rd and the whole 24th cycles will be presented elsewhere.

We have proposed a general overembedding method for modeling and prediction of nonstationary systems, which can be implemented with any learning algorithm. Using, in particular, ANNs, its application to both synthetic and real-world time series has shown that it leads to much more accurate results and longer prediction horizons than other existing methods in the literature. Our approach is able to treat strong nonstationarities provided the forcing is slow

enough in comparison with the system's short-scale dynamics. This requirement can be relaxed to some extent with a gracious degrading in the method's performance, as exemplified on the forced logistic and Hénon maps. Future work will explore the potential of the proposed method to model secular and anthropogenic changes in climatic time series.

This work was supported by the Alexander von Humboldt Foundation (P.F.V.), PAT project SAMPPA (P.M.G.), and CONICET (H.A.C.).

- [1] M. Casdagli and S. Eubank, *Nonlinear Modeling and Forecasting*, Santa Fe Institute Studies in the Science of Complexity XII (Addison-Wesley, Reading, MA, 1992); A. Weigend and N.A. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past*, Santa Fe Institute Studies in the Science of Complexity XV (Addison-Wesley, Reading, MA, 1993).
- [2] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge Nonlinear Science Series 7 (Cambridge University Press, Cambridge, England, 1997).
- [3] R. Hegger, H. Kantz, L. Matassini, and T. Schreiber, *Phys. Rev. Lett.* **84**, 4092 (2000); J. Stark, *J. Nonlinear Sci.* **9**, 255 (1999).
- [4] M. Small, D. J. Yu, and R. H. Harrison, *Non-stationarity as an Embedding Problem*, edited by S. Boccaletti, H.L. Mancini, W. Gonzalez-Vinas, J. Burguete, and D.L. Valladares, *Space-Time Chaos: Characterization, Control and Synchronization* (World Scientific, Singapore, 2001), p. 3.
- [5] M. Casdagli, *Physica (Amsterdam)* **35D**, 335 (1989).
- [6] A preliminary exploration of these ideas first appeared in a congress proceedings; see M.I. Széliga, P.F. Verdes, P.M. Granitto, and H.A. Ceccatto, *Physica (Amsterdam)* **327A**, 190 (2003).
- [7] We considered ANNs with 2 input units, 5 sigmoidal hidden units, and a linear output unit. Weights and biases were adjusted using the standard backpropagation rule.
- [8] For simplicity, M was considered alternatively $N/10$, $2N/10$, etc. (i.e., including the points in the first, two first, etc. previous intervals).
- [9] P.F. Verdes, P.M. Granitto, H.D. Navone, and H.A. Ceccatto, *Phys. Rev. Lett.* **87**, 124101 (2001).
- [10] H. Pi and C. Peterson, *Neural Comput.* **6**, 509 (1994).
- [11] Ensembles of ANNs with 8 hidden units were built using all the methods described in [12]. The best results presented here correspond to the algorithm called W-SECA in this reference.
- [12] P.M. Granitto, P.F. Verdes, and H.A. Ceccatto, *Artif. Intell.* **163**, 139 (2005).
- [13] The reconstructed α_t profile is practically identical to the SSN trend obtained with a 33-year (~ 3 short cycles) low-pass Fourier filtering. Interestingly, including α_{t-1} as an additional input produces a further 2% error reduction with respect to the use of α_t alone. Lack of space prevents us from discussing here the implications of this result.