Measurement noise reduction by local projection methods

M. E. Mera and M. Morán

Departamento de Análisis Económico I, Universidad Complutense
Campus de Somosaguas, 28223 Madrid. Spain
Email: mera@ccee.ucm.es, mmoranca@ccee.ucm.es

Abstract—A measurement noise reduction problem occurs when a sequence of states (time series) of a system governed by a low dimensional dynamics is recorded using a measurement process subject to error. In this paper we review those algorithms designed for noise reduction with unknown dynamics and that take a non-parametric approach, i.e. such that the only assumption on the underlying dynamics is its smoothness. These algorithms are always based on the local analysis of the data, which are a scalar time series in some cases and a multivariate time series in other cases. We show that all of the algorithms can be understood in a common framework: the projection of neighbourhoods on suitable linear manifolds. In this context the essential distinctive features of each algorithm may be explained in terms of the metric considered, the neighbourhoods chosen, and the goal pursued.

1. Introduction

Many problems relating to measurement noise reduction can be described by the equations

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k) \text{ and } \mathbf{X}_k = g(\mathbf{x}_k) + \mathbf{e}_k.$$

where $\mathbf{x}_k \in \mathbb{R}^d$ is the unobservable state vector of a system at period k; f is a smooth unknown dynamics, $g : \mathbb{R}^d \to \mathbb{R}^p$ is a smooth vectorial function or *observable;* $\mathbf{e}_k$ is an independent and identically distributed (*i.i.d.*) multivariate stochastic process; and $\mathbf{X}_k$, $k = 1, 2, ..., N$, is the available time series. In this scheme $\mathbf{e}_k$ is referred to as the *measurement noise*. Further, we adopt a non-parametric approach, which does not assume any specific form for the dynamics. Instead, it is based on a local analysis which tries to exploit the smoothness of f. There is a rich literature on algorithms that fall into this category (see [1]-[8]). They can be classified as those algorithms for which the function g is the identity and those designed for scalar time series, i.e. $g : \mathbb{R}^d \to \mathbb{R}$. These last algorithms require the reconstruction of a multidimensional state space equivalent in some sense to the original one. Takens's theorem guarantees [9] that in absence of noise, for generic observable g, and if $m \geq 2d + 1$, the m-dimensional reconstructed time series $\mathbf{x}_k^m := (x_k, x_{k+1}, ..., x_{k+m-1})$, $k = 1, 2, ..., N-m+1$, provides a diffeomorphic image of the attractor associated with f and that the main properties of the dynamics f can be inferred from those of the shift dynamics $f^*(\mathbf{x}_k^m) := \mathbf{x}_{k+1}^m$.

One difference between noise reduction for scalar and for multivariate time series is the structure of the error term. In scalar time series the error is unidimensional, so the degrees of uncertainty of the m coordinates of the reconstructed time series $\mathbf{X}_k^m := \mathbf{x}_k^m + \mathbf{e}_k^m$ are identical, i.e., $\mathbf{e}_k^m$ is an *uncorrelated* and *homoskedastic error* (it has covariance matrix $\Sigma = \sigma^2 I$ where I is the identity matrix). The algorithms designed for scalar time series can be used for multivariate time series corrupted by uncorrelated and homoskedastic noise. However, the results can be improved signicantly [10] if we incorporate into the reduction scheme information about the structure of the error term when it is correlated and/or heteroskedastic.

All of the algorithms we describe in this paper are based on local projections. The idea is that in absence of noise the reconstructed time series $\{\mathbf{x}_k^m, k = 1, ..., N - m + 1\}$ lies on a d-dimensional smooth submanifold M of $\mathbb{R}^m$. Therefore, for points $\mathbf{x}_j^m$ within a small ball U_i centered at $\mathbf{x}_i^m$, the m-dimensional difference vectors $\mathbf{x}_j^m - \left\langle \mathbf{x}_i^m \right\rangle_{U_i}$, where $\left\langle \mathbf{x}_i^m \right\rangle_{U_i}$ is the center of mass of U_i, span only a d-dimensional linear subspace, which is in fact an estimate of the tangent space of M at $\left\langle \mathbf{x}_i^m \right\rangle_{U_i}$. However, these difference vectors span $\mathbb{R}^m$ due to the noise, so the deterministic part of the time series can be (partially) recovered by projecting such vectors onto suitable d-dimensional linear subspaces.

For multivariate d-dimensional time series the algorithms first proceed to embed the time series in a higher dimensional space [3], [10], considering for instance $\mathbf{Z}_i = (\mathbf{X}_{i-1}, \mathbf{X}_i, \mathbf{X}_{i+1})$, $i = 2, ..., N - 1$ or $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{X}_{i+1})$, $i = 1, ..., N-1$. Since in absence of noise, these data lie also on a d-dimensional smooth submanifold, the algorithms recover the deterministic part of the time series projecting again on suitable d-dimensional subspaces.

Therefore, the first distinctive feature of each (scalar or vectorial) algorithm is what embedding space is to be used (i.e. the *working space)*. This determines what neighbourhoods are to be chosen. The second feature is the metric to be considered in the embedding space. We show below how the estimations given by each algorithm can be understood as a local projection with respect to a given metric.

The third distinctive feature of the algorithms is their goal. All of the algorithms are iterative: the noise reduction scheme produces a new time series that will be the input for the following iteration of the algorithm. In order to decide the end of the process we need to measure the

level of reduction of noise. Some authors [1]-[3], [6]-[8] try to reduce the pointwise distance between the cleaned time series $\widehat{\mathbf{x}}_i, i = 1, ..., N$ and the clean one,

$$E(\mathbf{x}, \widehat{\mathbf{x}}) := \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \widehat{\mathbf{x}}_i - \mathbf{x}_i \right\|^2 \right)^{1/2}.$$

This is consistent with the goal of recovering the original, uncorrupted time series. Other authors [5] try to maximize the local self-consistency of the data by reducing

$$E_{dyn}(\widehat{\mathbf{x}}) := \left(\frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \widehat{\mathbf{x}}_{i+1} - \widehat{f_i}(\widehat{\mathbf{x}}_i) \right\|^2 \right)^{1/2}, \qquad (1)$$

where $\widehat{f_i}$ is a local estimation of the dynamics f at $\mathbf{x}_i$. This provides a *dynamical noise reduction.*

Mera and Morán have recently proposed an alternative approach [10]. They try to recover the geometric properties and the long-run statistical regularity of the underlying dynamics by minimizing the mean distance to the attractor

$$d_m(\widehat{\mathbf{x}}, \mathbf{x}) := \frac{1}{N} \sum_{j=1}^{N} \min_{i=1,...,N} \left\| \widehat{\mathbf{x}}_j - \mathbf{x}_i \right\|.$$

Consider that any true orbit of the dynamical system conveys (with probability one) rich information on the geometry of the attractor and on the statistical long-run behaviour of the dynamics f. The d_m distance between any such orbit and the original clean orbit is very small. In contrast, the pointwise distance between both orbits is, in general, large. The noise reduction measure associated with d_m is $\langle R_h \rangle := 100\left(1 - \frac{d_m(\widehat{\mathbf{x}}, \mathbf{x})}{d_m(\mathbf{X}, \mathbf{x})}\right)$. We shall refer to this noise reduction measure as *statistical noise reduction.*

Creating a common framework for the different noise reduction algorithms considered here is made possible by the following theorem, which shows how to get the linear subspace which best fits data in $\mathbb{R}^n$ with respect to a given metric. Let A be an $n \times n$ symmetric positive definite matrix and let $\mathcal{L}_p$ be the set of p-dimensional linear subspaces of $\mathbb{R}^n$. The orthogonal projection of $\mathbf{v} \in \mathbb{R}^n$ onto $T \in \mathcal{L}_p$ with respect to the metric $\delta_{A^{-1}}$ is given by $\mathrm{P}_T\mathbf{v} := \arg\min_{\mathbf{w} \in T} \mathbf{v}A^{-1}\mathbf{w}^t$. We say that the linear subspace $\mathcal{T}_p$ is the best linear subspace in $\mathcal{L}_p$, for the points $\{\mathbf{Z}_k, k = 1, ..., NV\} \in \mathbb{R}^n$, with respect to the metric $\delta_{A^{-1}}$, if it minimizes over all $T \in \mathcal{L}_p$

$$\mathcal{E}(T) := \frac{1}{NV} \sum_{i=1}^{NV} (\mathbf{Z}_k - P_T\mathbf{Z}_k)^t A^{-1} (\mathbf{Z}_k - P_T\mathbf{Z}_k).$$

Such best linear subspace can be obtained through the eigenvectors of the $n \times n$ matrix $M_{\mathbf{ZZ}} := \frac{1}{NV} \Sigma_{k=1}^{NV} \mathbf{Z}_k \mathbf{Z}_k^t$ in the metric δ_A. A system of vectors $\{\mathbf{w}_1, ..., \mathbf{w}_n\} \subset \mathbb{R}^n$ is called *orthonormal system of eigenvectors of* $M_{\mathbf{ZZ}}$ *in the metric* δ_A if they satisfy: (i) there exist real numbers λ_i (called eigenvalues in the metric δ_A) such that $M_{\mathbf{ZZ}}\mathbf{w}_i = \lambda_i A\mathbf{w}_i, 1 \leq i \leq n$, and (ii) $\mathbf{w}_i^t A \mathbf{w}_j = \delta_{ij}$, for all i, j.

Theorem (see [10]) *Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $M_{\mathbf{ZZ}}$ in the metric δ_A, let $\{w_i, i = 1, ..., n\}$ be the corresponding orthonormal eigenvectors and, for $1 \leq p \leq n$, let B be the $n \times (n - p)$ matrix whose columns are the last $n - p$ eigenvectors $\{w_{p+1}, ..., w_n\}$. Then the best linear subspace in $\mathcal{L}_p$* with respect to the metric $\delta_{A^{-1}}$ is given by $T_p := span\{Aw_1, ..., Aw_p\}$ and $P_{\mathcal{T}_p} Z := (I - ABB^t)Z, Z \in R^n$.

2. Algorithms for scalar time series

2.1. Cawley-Hsu-Sauer method [1],[6]

These algorithms use as working space $\mathbb{R}^m$ for some suitable m, and they consider the Euclidean metric, i.e. they take the $m-$dimensional identity matrix I as the matrix A in the above theorem. Let U_i be a neighbourhood of $\mathbf{X}_i^m$ and let $\left\langle \mathbf{X}_i^m \right\rangle_{U_i}$ be the center of mass of the points within U_i. The algorithms reduce the noise by projecting the data $\mathbf{Z}_j := \mathbf{X}_j^m - \left\langle \mathbf{X}_i^m \right\rangle_{U_i}$, $\mathbf{X}_j^m \in U_i$ onto the best d-dimensional subspace $\mathcal{T}_d$. The estimation of $\mathbf{x}_j$ they compute is $\widehat{\mathbf{x}}_j = \left\langle \mathbf{X}_i^m \right\rangle_{U_i} + \alpha \mathbf{Z}_j + (1 - \alpha)\mathrm{P}_{\mathcal{T}_d}\mathbf{Z}_j$ where $0 \leq \alpha < 1$. Notice that this procedure ties up the cleaned time series to the original data through the term $\alpha\mathbf{Z}_j$. This is in concordance with the goal of a pointwise reduction. In fact, the authors justify this step as a useful means to soften the effects of rare statistical outliers. Since each point of the scalar time series appears as a component of m consecutive delay vectors, the algorithm provides multiple estimations for almost all of the points of the scalar time series. This problem is settled by averaging such estimations.

2.2. Schreiber-Grassberger method [8]

After a previous embedding of the data in $\mathbb{R}^m$ for a suitable (odd) value of $m > 2d + 1$, this algorithm proceeds to assign as the estimate of the central coordinate of each m-vector $\mathbf{X}_i^m$ a linear function of its $m - 1$ remaining coordinates. Such linear function is estimated using a neighbourhood U_i of $\mathbf{X}_i^m$. This gives a single estimate of each data point of the scalar time series. It can be proved that this procedure is equivalent to the projection of $\mathbf{X}_i^m$ on the best $m - 1$ dimensional subspace with respect to the metric $\delta_{A^{-1}}$ corresponding to a diagonal matrix A with all its entries equal to zero, with the exception of the central coordinate which is set equal to one. Since such a matrix is not invertible, the authors take a diagonal matrix whose diagonal entries are all very small (in the numerical results they take a value 0.001) and a single 1 in the central entry of the diagonal. Here the central coordinate plays an special role: If the coordinate computed as a linear function were to be the last coordinate, it would be ill-defined along the unstable manifold and a correction of the first coordinate would be ill-defined along the stable manifold. Only the central coordinate correctly takes into account information about both the past and the future. It should be mentioned here the very simple but quite efficient noise reduction algorithm [7] which proceeds in the same way as described above, but takes as an estimate of the central coordinate the

average of the central coordinates for all the points in the neighbourhood.

2.3. Grassberger et al. method [2]

This algorithm is a modification of the previous one but instead of correcting just the central coordinate for each point $\mathbf{X}_i^m$ it corrects several more central coordinates. How many of these will be corrected and the dimension of the projection subspaces are parameters of the algorithm. Since each data point of the scalar time series appears as a one central component of several delay vectors, the average of such estimations gives the final estimate of the point. In our setting, the metric $\delta_{A^{-1}}$ is given by a diagonal matrix A with all very small entries, with the exception of the central coordinates which are equal to ones.

3. Algorithms for multivariate time series

3.1. Kostelich and Yorke's method[5]

The smoothness of f implies that, and for any point $\mathbf{x}$ in a small neighbourhood of a point $\mathbf{x}_i$ in the attractor $f(\mathbf{x}) \sim \mathbf{B}_i\mathbf{x} + \mathbf{d}_i$ holds, where $\mathbf{B}_i$ is some $d \times d$ matrix and $\mathbf{d}_i$ is a m-dimensional vector. Thus, the algorithm takes a neighbourhood U_i for each point $\mathbf{X}_i$ of the noisy d-dimensional time series, and then gives as estimates of $\mathbf{B}_i$ and $\mathbf{d}_i$ the solution to the least squares problem

$$\min_{\mathbf{B},\mathbf{d}} \sum_{j\in U_i} \left\|\mathbf{X}_{j+1} - (\mathbf{B}\mathbf{X}_j + \mathbf{d})\right\|^2 .$$

A naive estimate $\widehat{\mathbf{x}}_i$ of $\mathbf{x}_i$ would be $\widehat{\mathbf{x}}_i := \mathbf{B}_i\mathbf{X}_i + \mathbf{d}_i$, but such an estimate is too influenced by the noise in $\mathbf{X}_i$ and by the error in the estimation of $\mathbf{B}_i$ and $\mathbf{d}_i$. For this reason the algorithm seeks a new time series which is consistent with the estimated linear maps. This is done by selecting windows of p consecutive points $\{\mathbf{X}_i, \mathbf{X}_{i+1}, ..., \mathbf{X}_{i+p}\}$ and then taking $\{\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_{i+1}, ..., \widehat{\mathbf{x}}_{i+p}\}$ as the solution of the optimization problem

$$\min_{\widehat{\mathbf{x}}_i,...,\widehat{\mathbf{x}}_{i+p}} w \sum_{j=i}^{i+p} \left\|\widehat{\mathbf{x}}_j - \mathbf{X}_j\right\|^2 + \sum_{j=i}^{p-1} \left\|\widehat{\mathbf{x}}_{j+1} - (\mathbf{B}_j\widehat{\mathbf{x}}_j + \mathbf{d}_j)\right\|^2 \quad (2)$$

where w is a weighting factor. Notice that this minimization problem is coherent with a dynamical noise reduction (see (1)). The link with the projection methods comes from the fact that the least squares estimator is also a projection matrix.

3.2. Hegger and Schreiber's method[3]

This algorithm is conceptually similar to the previous one. It is also an adaptation to multivariate time series of Schreiber and Grassberger's algorithm [8] described in the previous section. In order to capture the information about the past and the future in the estimate of $\mathbf{x}_i$ the algorithm takes the time series $\mathbf{Z}_i := (\mathbf{X}_{i-1}, \mathbf{X}_i, \mathbf{X}_{i+1}) \in \mathbb{R}^{3d}$, $i = 2, ..., N-1$. Thus the working space is $\mathbb{R}^{3d}$, with d the dimension of the original data points. The hypothesis is that the clean time series satisfies, for points $\mathbf{z}_j$ near to $\mathbf{z}_i := (\mathbf{x}_{i-1}, \mathbf{x}_i, \mathbf{x}_{i+1})$, the linear relationship

$$\mathbf{x}_j \sim \mathbf{B}_i\mathbf{x}_{j-1} + \mathbf{C}_i\mathbf{x}_{j+1} + \mathbf{d}_i$$

where $\mathbf{B}_i$ and $\mathbf{C}_i$ are $d \times d$ matrices and $\mathbf{d}_i$ is an d-dimensional vector. They are computed by solving the optimization problem

$$\min_{\mathbf{B},\mathbf{C},\mathbf{d}} \sum_{j:\mathbf{X}_j\in U_i} \left\|\mathbf{X}_j - (\mathbf{B}\mathbf{X}_{j-1} + \mathbf{C}\mathbf{X}_{j+1} + \mathbf{d})\right\|^2$$

where U_i is a neighbouhood of $\mathbf{Z}_i$. Let $\mathbf{X}_i^{corr} := \mathbf{B}_i\mathbf{X}_{i-1} + \mathbf{C}_i\mathbf{X}_{i+1} + \mathbf{d}_i$ be the estimate given by the linear model described above. The algorithm takes as estimate $\widehat{\mathbf{x}_i} = (1-\alpha)\mathbf{X}_i + \alpha\mathbf{X}_i^{corr}$ where $0 < \alpha \leq 1$. It can be proved that $\mathbf{X}_i^{corr}$ are the central coordinates of the orthogonal projection of $\mathbf{Z}_i$ on the best linear d-dimensional subspace with respect to the metric $\delta_{A^{-1}}$ where A is a diagonal matrix having all the entries almost null except for the d central coordinates, which are ones.

3.3. Mera and Morán's method[10]

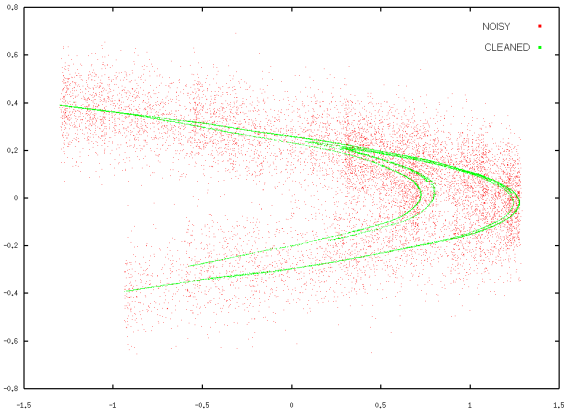
This algorithm pursues a statistical noise reduction, and it is the only algorithm in the literature able to exploit the available information that an uncorrelated and heteroskedastic noise contains. Its starting point is Kostelich and Yorke's algorithm [5]. In fact these authors already pointed out the unsuitability of the least squares fits. Indeed it is well known that in linear models where the independent variables are also measured with error, the least squares estimators are biased, being the bias an increasing function of the variances. Fuller [11] establishes the theory for these type of models showing how to get unbiased and consistent estimators, which are also the maximum likelihood estimators if the errors are Gaussian.

Mera and Moran's algorithm uses $\mathbb{R}^d$ as working space, with d the dimension of the data. It starts by taking a neighborhood U_i for each point $\mathbf{X}_i$ and it uses the theory developed in [11] to obtain an unbiased maximum likelihood estimate of the matrix $\mathbf{B}_i$ satisfying

$$\mathbf{X}_{j+1} - \langle\mathbf{X}_{i+1}\rangle_{U_i} \approx \mathbf{B}_i\left(\mathbf{X}_j - \langle\mathbf{X}_i\rangle_{U_i}\right), \mathbf{X}_j \in U_i$$

where $\langle\mathbf{X}_i\rangle_{U_i}$ denotes the center of mass of the points of the neighborhood U_i and $\langle\mathbf{X}_{i+1}\rangle_{U_i}$ is the center of mass of their images. Furthermore, this theory gives, for each $\mathbf{X}_j \in U_i$, the maximum likelihood estimate $\widehat{\mathbf{z}}_j$ of $\mathbf{z}_j := (\mathbf{x}_{j+1} - \langle\mathbf{x}_{i+1}\rangle_{U_i}, \mathbf{x}_j - \langle\mathbf{x}_i\rangle_{U_i})$. Using such estimation makes unnecessary the last step in the Kostelich and Yorke's algorithm (see (2)).

The link between Mera and Moran's algorithm and the projection methods comes from the fact that $\widehat{\mathbf{z}}_j$ coincides [10] with the orthogonal projection of $\mathbf{Z}_j$ on the best d-dimensional linear subspaces in the metric Σ^{-1}, where Σ is the covariance matrix of the errors $\mathbf{e}_j$ in $\mathbf{Z}_j := \mathbf{z}_j + \mathbf{e}_j$.



Thus, this method takes advantage of the structure of Σ to reduce the noise. The authors prove that this method gives a time series having almost the same statistical and geometric properties as the true dynamics, even in the case of high noise amplitudes or when the variances of the components of the error are different and/or there are correlations. Since, in general, the matrix Σ is unknown, the identity matrix is taken as Σ at the first iteration of the algorithm, and it is updated in the next iterations with the empirical covariance matrix of the estimated errors.

The figure above shows a noisy time series of 10000 points generated by the Hénon map, with a highly heteroskedastic Gaussian noise ($\sigma_{e_x} = 0.01\sigma_s$ and $\sigma_{e_y} = 0.15\sigma_s$, where σ_s is the standard deviation of the clean signal). We have plotted the output of the algorithm together with the noisy time series. The levels of reduction are 83% in terms of pointwise distance, 97% in terms of dynamic consistency and 94% in terms of statistical regularity.

4. Conclusions

We have reviewed several algorithms addressed to measurement noise reduction and based on a non-parametric, local approach. All these algorithms can be understood as producing outputs through local orthogonal projections, with respect to suitable metrics, onto linear manifolds. They differ in the working spaces, the metrics and the goals pursued. Although these are the three essential distinctive features, there are also other special useful devices proposed by the authors of the algorithms whose discussion is beyond the scope of this short review. Interested readers can find these details in the quoted literature. In particular criteria are avaliable for an adaptive selection of the size of the neighbourhoods [13], and for a proper selection of the dimension of the subspace in which the neighbourhoods must be projected [12], both depending on the geometry of the time series at each point instead of remaining fixed for all the points.

All of the algorithms display a highly efficient performance in their respective goals. For instance for a time series of 5000 data points from Hénon map corrupted with an uncorrelated and heteroskedastic Gaussian noises with amplitudes between 1% and 10% they give a noise reduction in terms of pointwise distance between 65% and 73% and between 83% and 87% in terms of dynamical consistency. These results seem to indicate that, in spite of the stronger intuitive content of pointwise noise reduction and the appealing fact that pointwise noise reduction permits the improvement of the short run forecasting of the series, there exist intrinsic bounds for such goal. These bounds might be related to our limited capability for a pointwise, short run prediction of chaotic dynamical systems recorded with uncertainty, even if the dynamics is known. By contrast, the more realistic goal of statistical noise reduction could find its bounds only in the length of the data sets.

Acknowledgments

We would like to thank NOLTA2006 organizing committee members for their fruitful suggestions and comments.

This research has been supported by the Ministerio de Ciencia y Tecnología, research project BFM2003-08204/Mate.

References

[1] R. Cawley and G-H. Hsu. Phys. Rev. A 46, 6, 3057-3082 (1992).

[2] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath and T. Schreiber. Chaos 3, 2, 127-141 (1993)

[3] R. Hegger and T. Schreiber. Phys. Lett. A 170, 305-310 (1992).

[4] E. J. Kostelich and T. Schreiber. Phys. Rev. E 48, 3, 1752-1763 (1993).

[5] E. J. Kostelich and J. A. Yorke. Physica D 41, 183-196 (1990).

[6] T. Sauer. Physica D 58, 193-201 (1992).

[7] T. Schreiber. Phys. Rev. E 47, 4, 2401-2404 (1993)

[8] T. Schreiber and P. Grassberger. Phys. Lett. A 160, 411-418 (1991).

[9] F. Takens. "Dynamical Systems and turbulence." D. A. Rang and L. S. Young eds. Lecture Notes in Mathematics 898, 396 (1981).

[10] M.E. Mera and M. Morán. Chaos 16, 1, 13116 (2006).

[11] W. A. Fuller. "Measurement error models." John Wiley & Sons (1987).

[12] A. Leontitsis, T. Bountis and J. Pagge. Chaos 14, 1, 106-11, (2004).

[13] A. Kern, W.-H. Steeb and R. Stoop. Int. J. Mod. Phys. C 11, 1, 125-146 (2000).