Contribution for the book
"Nonlinear Deterministic Modelling and Forecasting
of Economic and Financial Time Series"
by A. Soofi and L. Cao
15 September 2000

State Space Local Linear Prediction

D. Kugiumtzis
Department of Statistics, University of Glasgow,
Glasgow G12 8QW, UK

Key word index: *time series, nonlinearity, chaos, local linear prediction, regularisation*

Abstract

Local linear prediction is one of several methods that have been applied to prediction of real time series including financial time series. The difference from global linear prediction is that, for every single point prediction, a different linear autoregressive (AR) model is estimated based only on a number of selected past scalar data segments. Geometrically, these data segments correspond to points close to the target point when the time series is viewed in a pseudo-state space with dimension equal to the order of the local AR model.

The parameters of the local linear model are typically estimated using ordinary least squares (OLS). Apart from potential linearisation errors, a drawback of this approach is the high variance of the predictions under certain conditions. It has been shown that a different set of so-called regularisation techniques, originally derived to solve ill-posed regression problems, gives more stable solutions (and thus better predictions) than OLS on noisy chaotic time series. Three regularisation techniques are considered, i.e. principal component regression (PCR), partial least squares (PLS) and ridge regression (RR). These methods reduce the variance compared to OLS, but introduce more bias. A main tool of this analysis is the Singular Value Decomposition (SVD), and a key to successful regularisation is to dampen the higher order SVD components. For the sake of completeness, truncated total least squares is discussed as well, which is designed to solve "error-in-variables" problems. Even though it would be expected that this method is more appropriate for noisy time series, it turns out to give the worst predictions.

This chapter will describe the general features of local linear prediction and particularly the OLS solution and the regularisations. The statistical properties of the methods will be highlighted and explained in the setting of local linear prediction. The superiority of the predictions using regularised solutions over OLS predictions will be demonstrated using simulated data and financial data.

1 Introduction

Nonlinear dynamics and chaos offer a different perspective in understanding random-like phenomena, ranging from heart function to monetary system evolution, as well as new computational tools for analysing the measured data, e.g. heart rates and exchange rates. The

rationale is that the apparently complex behaviour of the data may be due to low-dimensional nonlinear dynamics, which can possibly be identified and modelled using appropriate nonlinear tools. Methods based on nonlinear dynamics and chaos have been widely used for the analysis of real world time series with varying degrees of success [1, 2].

For some applications, especially in finance, the interest is in short term predictions rather than identification of the underlying system. A variety of more or less complex prediction techniques has been used for this scope, including neural networks and radial basis functions, but there is little variation in the quality of fit across the methods (see Chapters 8-14; for a comparative study see [3]). In this respect, local linear methods represent simple and attractive alternatives for prediction purposes. Moreover, when enriched with sophisticated techniques for state space representation and filtering, localisation and map smoothing, local linear models can usually attain high levels of predictability. For example, during the last decade local linear models were the winning entries in two prediction competitions [4, 5], in the first of which they competed against a neural network model.

Local modelling in terms of kernel regression estimation is discussed in Chapter 7 (see also [6]). Here, local linear prediction is studied in the frame of nonlinear dynamics and so-called *chaotic time series analysis*, taking up techniques from linear regression.

The local linear prediction model is simply the linear regression model applied locally, in the sense that the fitted hyper-plane is restricted to a small area around the target point in the state space, reconstructed from the scalar data. This approach can be seen as an extension of piecewise linear regression and threshold autoregressive models [7]. Since the first implementation of the idea of local linear prediction on chaotic time series in [8], additional refinements have been suggested with respect to the selection of the points in the vicinity of the target point and the weighting of the points, as well as the parameters for the reconstruction of the state space (see [1] and the references therein).

For the estimation of the regression parameters, the ordinary least squares (OLS) solution has been routinely employed, but recently it was shown that so-called regularisation techniques may modify the OLS solution towards better predictions [9]. Well-known regularisation techniques include principal components regression, partial least squares and ridge regression [10]. In our study we include also the method of total least squares [11], which solves "error-in-variables" problems. A variation of this method was used successfully to model the underlying dynamics in [12], but it is found to perform poorly on prediction tasks. The regularisation techniques were initially developed to tackle ill-posed linear problems where the regression matrix is poorly conditioned. Later it was found that they could also reduce the effect of noise on the parameter estimation. In local linear prediction with noisy data, the OLS solution can have large variance and the regularisation methods, designed to be more robust against noise, may provide better results. For example, a simple version of principal component regression was applied in [13] and in the winning entry of the first prediction competition [14]. This method was also used successfully in [15] for local prediction with small multivariate data sets.

The OLS and regularisation techniques will be presented within the framework of state space local linear prediction of time series in Section 2. In Section 3, the application of the models will be discussed and will be presented.

2 Local prediction

Suppose a scalar time series $x_i = x(i\tau_s)$, $i = 1, \ldots, t$, is given, where τ_s is the sampling time (for discrete systems we have $\tau_s = 1$), and the prediction of the future scalar state x_{t+T} is sought for a prediction time T. The principal idea of local prediction goes back to the analogue method of Lorenz [16], where the prediction of x_{t+T} is estimated from the scalar state being T time step ahead of the past segment of the time series, which is most similar to the segment consisting of the current samples. More than one past similar segment may also be utilised. Many improvements of this idea can be realised once the segments are viewed as points in a pseudo state space.

2.1 State space representation

Using the method of delays (MOD), the scalar data can be represented in $\mathbb{R}^m$ by the vectors $\tilde{\mathbf{x}}_i = [x_i, x_{i-\tau}, \ldots, x_{i-(m-1)\tau}]^{\mathrm{T}}$, $i = 1 + (m-1)\tau, \ldots, t$, where m is the embedding dimension, τ is the delay time in units of τ_s and the superscript T denotes the transpose [17]. In this way, the scalar time series is transformed to a trajectory in the reconstructed state space $\mathbb{R}^m$, and the points of the trajectory are the data segments of time length $\tau_w = (m-1)\tau$. In the setting of deterministic systems, the set of these points, so-called attractor, is meant to preserve the topological properties of the original unknown attractor if $m \geq 2d + 1$, where d is the fractal dimension of the original attractor [18]. However, in practice, this theoretical result cannot be validated and it is assumed to hold approximately, given the limitations of data size and noise.

The idea of analogies, i.e. finding similar segments of scalar data, is directly related to the property of recurrence for the orbits of dynamical systems, which constitutes the main theoretical grounds for attempting local predictions. So, the problem of finding segments similar to the target segment is formulated as that of finding the neighbour points to the point $\tilde{\mathbf{x}}_t$ and using them to predict x_{t+T}. Even if the original dynamics is chaotic, close orbits deviate gradually and some degree of short-term prediction can still be achieved. However, the reconstructed orbits starting from $\tilde{\mathbf{x}}_t$ and its neighbours may not deviate as smoothly as the original orbits, e.g. because the dimension of the reconstructed state space is too low and the orbits are badly projected on it or too high and noise dominates along the redundant directions. In this respect, careful state space reconstruction is of immense importance for local prediction, and this does not simply rely on the selection of m, but rather on the selection of τ_w. For a given τ_w, state space reconstructions for varying m (adjusting τ accordingly, so that $\tau_w = (m-1)\tau$ holds) are qualitatively the same [19]. Certainly, $\tau = 1$ accounts for the most detailed representation of the data segments, as all samples in the segments are considered, but at the cost of a high dimensional state space ($m = \tau_w + 1$). In the next sections, we will discuss how we can retain $\tau = 1$, and combine large τ_w with low dimensions.

2.2 Functional approximation

The simplest form of local prediction is to consider only the most similar segment, or equivalently the nearest neighbour point, $\tilde{\mathbf{x}}_{t_1}$, say, for some time index $t_1 < t–T$, and use the sample x_{t_1+T} to predict x_{t+T}. For small or noisy data, this approach has limited predictive power, but due to its simplicity and computational efficiency it is useful for other purposes, e.g. for discriminating different data sets [20] or for long-term predictions [21]. An improvement is to take the average of the mappings of the k nearest neighbours, $\mathcal{N}_t = \{\tilde{\mathbf{x}}_{t_1}, \ldots, \tilde{\mathbf{x}}_{t_k}\}$. The average may be modified to account for the location of the neighbour points, either by weighting the mappings according to the distance of the respective neighbour points from the target point $\tilde{\mathbf{x}}_t$ [22], or by considering only the $m+1$ neighbour points closest to $\tilde{\mathbf{x}}_t$ under the constraint that they form a simplex including $\tilde{\mathbf{x}}_t$ [23]. Here, we will not consider these geometric or zero order approaches, but we will study predictions provided through the estimation of a local map, i.e. the neighbour points are taken as regressors of their respective future scalar states.

Assuming that the data are generated from a dynamical system and the state space reconstruction is sufficient, there is a functional dependence of x_{t+T} onto $\tilde{\mathbf{x}}_t$: $x_{t+T} = F^T(\tilde{\mathbf{x}}_t)$. The graph of the reconstructed dynamics for T time steps ahead, F^T, is a smooth surface in the augmented state space $\mathbb{R}^{m+1}$, where the last coordinate is for x_{t+T}. Restricted to the neighbourhood of $\tilde{\mathbf{x}}_t$, the surface may be approximated by a plane, but polynomials of higher degree may also be used to approximate F^T locally. The linearisation of F^T at the centre of mass of the neighbourhood of $\tilde{\mathbf{x}}_t$, $\bar{\mathbf{x}}$, gives

$$x_{t+T} - \bar{y} = \nabla F^T(\bar{\mathbf{x}})(\tilde{\mathbf{x}}_t - \bar{\mathbf{x}}) + \epsilon_t^d, \tag{1}$$

which can be written as

$$y_t = \mathbf{b}^{\mathrm{T}}\mathbf{x}_t + \epsilon_t^d, \tag{2}$$

where $\bar{y}$ is the T-mapping of $\bar{\mathbf{x}}$, $\nabla F^T(\bar{\mathbf{x}}) = \mathbf{b}^{\mathrm{T}}$ is the gradient of F^T at $\bar{\mathbf{x}}$, and $\mathbf{x}_t$ and y_t are the centred differences for $\tilde{\mathbf{x}}_t$ and x_{t+1}, respectively. The error term ϵ_t^d is assumed to have zero mean and variance $(\sigma^d)^2$, and accounts for model errors due to linearisation as well as due to incomplete reconstruction of the dynamics, e.g. due to inappropriate selection of the MOD parameters. For the extension to higher-order approximation, additional terms are included in the Taylor expansion in eq.(1), ending up with a linear model of the same form as in eq.(2) but with augmented $\mathbf{b}$ and $\mathbf{x}_t$, where $\mathbf{x}_t$ includes also higher order terms of component-wise centre differences. A computational advantage of centring the points is the suppression of the constant term in the local linear model.

The model in eq.(2) applies also to the neighbour points. The centred versions of the matrix of the neighbour points $\tilde{X} = (\tilde{\mathbf{x}}_{t_1}, \ldots, \tilde{\mathbf{x}}_{t_k})^{\mathrm{T}} \in \mathbb{R}^{k,m}$ and the vector of their respective mappings $\tilde{\mathbf{y}} = (x_{t_1+T}, \ldots, x_{t_k+T})^{\mathrm{T}} \in \mathbb{R}^k$ are

$$X = \tilde{X} - \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}, \quad \mathbf{y} = \tilde{\mathbf{y}} - \mathbf{1}\bar{y},$$

where $\mathbf{1}$ is a $k \times 1$ vector of ones. The centre point $\bar{\mathbf{x}}$ can be simply taken as the column vector of averages of the m columns of $\tilde{X}$, and the centred mapping $\bar{y}$ as the average of the

components of $\tilde{\mathbf{y}}$. Thus the following linear regression model is derived:

$$\mathbf{y} = X\mathbf{b} + \boldsymbol{\epsilon}^d.$$

For the linear (first order) approximation, $X \in \mathbb{R}^{k,m}$ is the predictor matrix formed by the centred neighbour points, $\mathbf{y} \in \mathbb{R}^k$ is the response vector of the centred mappings of the neighbour points, $\mathbf{b} \in \mathbb{R}^m$ is the vector of regression parameters and $\boldsymbol{\epsilon}^d$ is the vector of model errors with expectation $E\{\boldsymbol{\epsilon}^d\} = \mathbf{0}$ and covariance matrix $\text{Var}\{\boldsymbol{\epsilon}\} = (\sigma^d)^2 I$, where I is the identity matrix. For higher-order approximations, X and $\mathbf{b}$ are expanded to account for higher-order terms, aiming to reduce σ^d. However, the advantage of using more degrees of freedom in the approximation is down-weighted by the shortcoming of having more coefficients to estimate, so it is not certain that high order approximation is better when k is not sufficiently large. In the following, we assume only first order approximation.

2.3 Noise in the data

So far, we have not considered any error sources apart from the model error ϵ^d. When we deal with real data, this error term may include also other uncertainties regarding the hypothesised dynamics (random effects on the system evolution and exogenous factors interacting with the system) and is referred to as dynamical error. Other errors not related to the system evolution may be present as well, such as measurement uncertainty and round-off errors, referred to as measurement noise and denoted by ϵ^m. In regression problems, the predictor matrix X is assumed to be known exactly and the error ϵ^m is associated only with the response $\mathbf{y}$, so that the model becomes

$$\mathbf{y} = X\mathbf{b} + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^d + \boldsymbol{\epsilon}^m$, $\boldsymbol{\epsilon}^m$ has uncorrelated components, each with mean zero and variance $(\sigma^m)^2$. Thus $E\{\boldsymbol{\epsilon}\} = \mathbf{0}$ and $\text{Var}\{\boldsymbol{\epsilon}\} = \sigma^2 I$ with $\sigma^2 = (\sigma^d)^2 + (\sigma^m)^2$.

In principle, the model in eq.(3) is not appropriate when X and $\mathbf{y}$ are derived from the time series because they are both corrupted by the same level of measurement noise ϵ^m. Thus the implicit assumption in eq.(3) is that $\mathbf{y}$ is linearly related to the corrupted X rather than the true X^{true}. A more realistic approach is to model the measurement error in X explicitly. This is referred to as an "errors-in-variables" problem and the total least squares (TLS) solution is called for [24]. The TLS solution is more appropriate than OLS for parameter estimation in the modelling of noisy time series [12], but not for prediction purposes [9]. In the next section, we will discuss the problem of dealing with noisy X, but we will base our analysis on the model in eq.(3) with the assumption that X is fixed.

It should be stressed that though the noise variance may be small compared to the data variance (high signal to noise ratio), the error term $\boldsymbol{\epsilon}$ in eq.(3) may be comparatively large because the model is local. For example, a relatively small measurement noise level of 5% of the data magnitude gives an error term in eq.(3) with higher variance than the variance of the predictor and response data if the radius for the neighbourhood is 5% of the standard deviation of the data, which is not an extreme choice for the neighbourhood size. Therefore, techniques robust to noise are sought in the estimation of the local linear prediction model.

2.4 Estimation of the linear prediction map

In this section, we focus on the estimation of the parameter vector $\mathbf{b}$ in order to make inference statements about a future scalar response y_t at a prespecified point $\mathbf{x}_t$ (see eq.(2)), given X and $\mathbf{y}$.

In our analysis, we use the singular value decomposition (SVD) of $X \in \mathbb{R}^{k,m}$ defined as

$$X = U\Sigma V^{\mathrm{T}}, \quad U^{\mathrm{T}}U = I, \quad V^{\mathrm{T}}V = I, \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r),$$

where $r \leq \min(k, m)$ is the rank of X and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ are the non-zero singular values of X. The columns of U and V span the r-dimensional range spaces $R(X) \subset \mathbb{R}^k$ and $R(X^{\mathrm{T}}) \subset \mathbb{R}^m$, respectively.

The prediction estimator of y_t is defined as $\hat{y}_t = \mathbf{x}_t^{\mathrm{T}}\hat{\mathbf{b}}$, where $\hat{\mathbf{b}}$ is an estimate for $\mathbf{b}$. From eq.(3), an estimate $\hat{\mathbf{b}}$ can be found from the pseudo-inverse of X. We consider a general approximation to the pseudo-inverse of X involving a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$. The estimators for $\mathbf{b}$ are then expressed in the following form using the SVD of X [25]:

$$\hat{\mathbf{b}} = V\Sigma^{-1}\Lambda U^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{r} \frac{\lambda_i}{\sigma_i}(\mathbf{u}_i^{\mathrm{T}}\mathbf{y})\mathbf{v}_i, \tag{4}$$

where the inner products $(\mathbf{u}_i^{\mathrm{T}}\mathbf{y})$ are usually referred to as the Fourier coefficients for $\mathbf{y}$. Estimators of this form differ only in their choice of the diagonal elements of Λ, called *filter factors*. Each λ_i weights the corresponding singular direction $\mathbf{u}_i$ and determines the extent of shrinking, when $0 \leq \lambda_i \leq 1$, or stretching, when $\lambda_i > 1$, along $\mathbf{u}_i$. The *ordinary least squares* (OLS) and most of the well-known regularisation estimators can be expressed in terms of the filter factors as follows [25, 26].

The filter factors for the OLS estimator are simply

$$\lambda_i = 1, \quad i = 1, \ldots, r,$$

i.e. all directions of $R(X)$ spanned by the columns of U contribute equally to $\hat{\mathbf{b}}_{\mathrm{OLS}}$.

The method of *principal components regression* (PCR) uses a subspace of $R(X)$ spanned by the first $q < r$ singular directions and the filter factors are

$$\lambda_i = 1, \quad i = 1, \ldots, q \quad \text{and} \quad \lambda_i = 0, \quad i = q+1, \ldots, r.$$

showing that only the first q terms are used in eq.(4) to determine $\hat{\mathbf{b}}_{\mathrm{PCR}}$, i.e. the effective dimension for the estimation problem falls from m to q.

The *Partial least squares regression* (PLS) estimator shrinks the OLS solution similarly to PCR using a shrinkage parameter q, but takes into account not only the size of the singular values, as does PCR, but also the size of the Fourier coefficients [27]. The filter factors are not the best descriptors of the shrinkage in the case of PLS, as a subspace other than the one spanned by the principal vectors $\mathbf{u}_i$ is utilised. In [26], the relevant expressions for the PLS filter factors are given as

$$\lambda_i = 1 - \prod_{j=1}^{q}(1 - \frac{\sigma_i^2}{\theta_j}), \quad i = 1, 2, \ldots, m,$$

where $\theta_1 \geq \theta_2 \geq \ldots, \geq \theta_m$ are the Ritz values (for the definition of Ritz values see e.g. [28]). Notice that λ_i, for $i = q+1, \ldots, m$, are not set to zero and that some filter factors can even be larger than one, but in an ordered manner so that $\|\hat{\mathbf{b}}_{\text{PLS}}\|_2 \leq \|\hat{\mathbf{b}}_{\text{OLS}}\|_2$ always holds [26], as it holds in general for every regularisation of OLS ($\|\mathbf{x}\|_2$ is the Euclidean norm of vector $\mathbf{x}$).

The *ridge regression* (RR) estimator is defined simply by adding a constant μ to the diagonal elements of the matrix $X^{\text{T}}X$ [29]. The filter factors for the RR estimator are given by

$$\lambda_i = \frac{\sigma_i^2}{\sigma_i^2 + \mu}, \quad i = 1, 2, \ldots, r.$$

Thus the ridge regression estimator $\hat{\mathbf{b}}_{\text{RR}}$ shrinks the OLS estimator $\hat{\mathbf{b}}_{\text{OLS}}$ in every direction when $\mu > 0$ by an amount depending on μ and on the corresponding singular values σ_i.

For problems with noisy X and $\mathbf{y}$, the *total least squares* (TLS) solution, is supposed to be more suitable. Actually, more useful is the *truncated total least squares* (TTLS) estimator for a given truncation parameter $q \leq r$ [11]. It is difficult to express the estimator $\hat{\mathbf{b}}_{\text{TTLS}}$ in the form of eq.(4). The exact expressions for the filter factors are complicated and can be found in [25]. In the TLS literature prediction is not discussed, probably because TLS does not perform well in prediction problems. This is not surprising considering that $\lambda_i > 1$ for $i = 1, \ldots, q$, i.e. TTLS does the opposite of regularisation in the first q directions.

Using the expression for the general estimate $\hat{\mathbf{b}}$ in eq.(4), the uncertainty of the prediction estimate $\hat{y}_t$ can be estimated in terms of the mean squared error, bias and variance. The expressions for these quantities when assuming that the centred target vector $\mathbf{x}_t$ is corrupted by noise are given in [9]. In particular, the prediction variance can be large due to several factors, one being small singular values of the design matrix X. The variance can be decreased by reducing the filter factors λ_i at the cost of introducing extra bias. The optimal trade-off between bias and variance consists of finding filter factors such that the mean squared error is minimised.

2.5 Selection of regularisation parameters

The performance of the regularised estimates is heavily dependent on the selection of the shrinkage parameter (q for PCR, PLS or TTLS and μ for RR), the aim being to obtain best possible trade-off between prediction bias and variance. Several model selection techniques exist for linear estimators, the most popular being cross-validation (CV) and generalised cross-validation (GCV), the latter found to be superior to other selection procedures, e.g. see [30]. Cross-validatory procedures make use of a risk measure or measure of fit, and the effect of the measures on the estimation of the shrinkage parameter has been investigated in particular in the case of RR [31]. Cross-validation is also the usual practice when choosing q for PLS. For PCR and TTLS, there are simpler ways to choose q making use of the singular spectrum alone, either by finding a threshold value that represents the noise variance, or by requiring that the included singular values account for at least a specified proportion of the total data variation in X. For RR, there are a number of data dependent choices for the estimation of μ based on the length of $\hat{\mathbf{b}}_{\text{RR}}$ and some measure of the residuals [10].

In local prediction, the problem of the selection of the regularisation parameter is more involved because the parameter has to be selected or estimated from the data for each target point. A simpler approach is to use the same value for the regularisation parameter in all local predictions for a given time series. Under the assumption of the presence of nonlinear dynamics, a reasonable choice for a fixed q is the topological dimension of the underlying attractor, if this can somehow be estimated, e.g. from a dimension estimation method [1]. However, the local curvature of the graph of F^T may vary substantially and then it may be more appropriate to let the regularisation parameter vary with the target point.

In applications with chaotic time series, it was found in [9] that CV often overestimates the regularisation parameters of the local models. Regarding the singular spectrum based choices for the q of PCR and TTLS, there is seldom a gap in the spectrum that could indicate a clear cut-off level. Also, using noise variance to judge the cut-off level gives q close to m. The proportion of the total variance does not constitute a robust criterion either because the estimated q increases with m. When the local model is given by the RR solution, a simple choice of μ from the residual variance estimate $\hat{s}^2$ was suggested in [9]. The larger the residuals are, the larger the prediction variance becomes and hence the stronger the regularisation should be to deal with this. Overall, the selection of "best" regularisation parameters is an open problem and "trial and error" seems to be the choice of practice.

We end this Section illustrating the shrinkage properties of the different estimators by means of the filter factors. In Fig. 1, the results are shown for a single local prediction using 2000 noise-free data from the chaotic Ikeda map [32]. The residual standard error $\hat{s}$

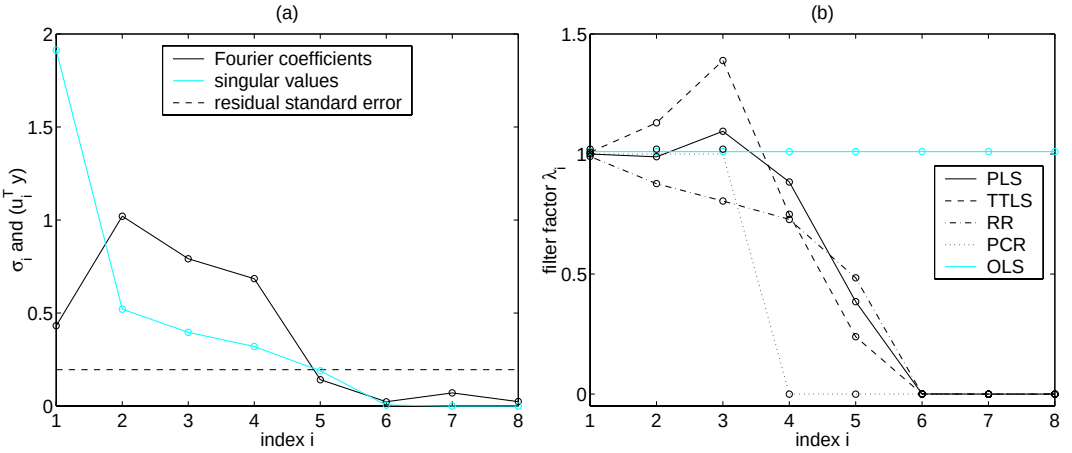

Figure 1: (a) The singular values σ_i, the magnitude of the Fourier coefficients $|\mathbf{u}_i^{\mathrm{T}}\mathbf{y}|$, and the level of the residual standard error $\hat{s}$ for a single local prediction using data from the Ikeda map ($k=15$, $m=8$). (b) The filter factors of the estimators as shown in the legend. For RR, $\mu=\hat{s}^2$ and for PCR, PLS and TTLS, $q=3$.

accounts only for the model error because the data are noise-free and it is small compared to the principal singular values σ_i and the magnitude of the Fourier coefficients $|\mathbf{u}_i^{\mathrm{T}}\mathbf{y}|$. The last three σ_i and $|\mathbf{u}_i^{\mathrm{T}}\mathbf{y}|$ are close to zero and the corresponding filter factors are zero also for all regularisations, so that regularisation is justified for this example. However, PCR

with $q=3$ appears to be too conservative, filtering out directions $\mathbf{u}_i$, $i=4,5$, which explain some variation in X and are correlated with $\mathbf{y}$. For both PLS and TTLS, some of the filter factors λ_i are well above one. The common practice with noisy data is that λ_q for TTLS gets substantially larger than one and then the solution becomes unstable.

3 Implementation of Local Prediction Estimators on Time Series

In the standard global linear or polynomial prediction of time series, the free parameters are the order m of the regression model (or equivalently the embedding dimension m), possibly the delay time τ, and the regularisation parameter (q or μ) if shrinkage of the OLS estimator is wanted. In local prediction, an additional free parameter determines the local region. This can be a distance length, so that all points within this distance from the target point $\tilde{\mathbf{x}}_t$ are included in the model, or a number k, so that only the closest k points are considered. This parameter is usually fixed for all target points, but it can also be optimised through cross-validation.

For noise-free data from a low-dimensional system embedded in a state space of sufficiently large dimension m, typically the points will have locally little variance in some directions, in which case the data matrix X becomes ill-conditioned. The reason for this is that locally the attractor of the system is mainly confined to some subspace of $\mathbb{R}^m$. However, small variations outside this subspace may still contain valuable information, so that regularisation will actually worsen the prediction unless the condition number of X is so large that numerical problems are encountered. Regularisation can be useful for noise-free data when $m \geq k$, as for this situation the OLS estimate becomes unstable. On the other hand, the solutions with PCR and PLS are stable because the actual dimension of the regression problem is as small as q even though the dimensionality of the state space is large. For data from continuous systems, this property is appreciated when one wants to use $\tau=1$ and a large m to include all the samples within the time window length τ_w in the point representation.

When measurement noise is present, X tends to be better conditioned. However, the prediction capability of OLS deteriorates because the part of the OLS solution that relates the directions masked with noise to the future state does not really contain any useful information. For the regularisation, the problem is how to identify this part of the solution and filter it out and, as mentioned in Section 2, there is no obvious universal strategy to do this one and for all target predictions.

To validate the predictive power of a model, the available data set is simply split into two parts, one for fitting and one for testing. Cross-validatory procedures can be used here as well. Usually the normalised root mean squared error (NRMSE), computed on all points in the test set, is used to measure the quality of prediction. A value of NRMSE at 1 means that the prediction is as good as the mean value prediction whereas NRMSE at 0 accounts for perfect prediction.

3.1 An example with simulated data

We illustrate how the performance of the linear models (OLS and regularisations) changes with noise using a simple chaotic low-dimensional system, the Henon map [33]. The NRMSE for one time step ahead predictions of the noise-free and noisy Henon data with the prediction estimators presented in Section 2 are shown in Fig. 2 for a range of m values and two choices of k. The regularisation parameter $q=2$ is chosen to match the topological

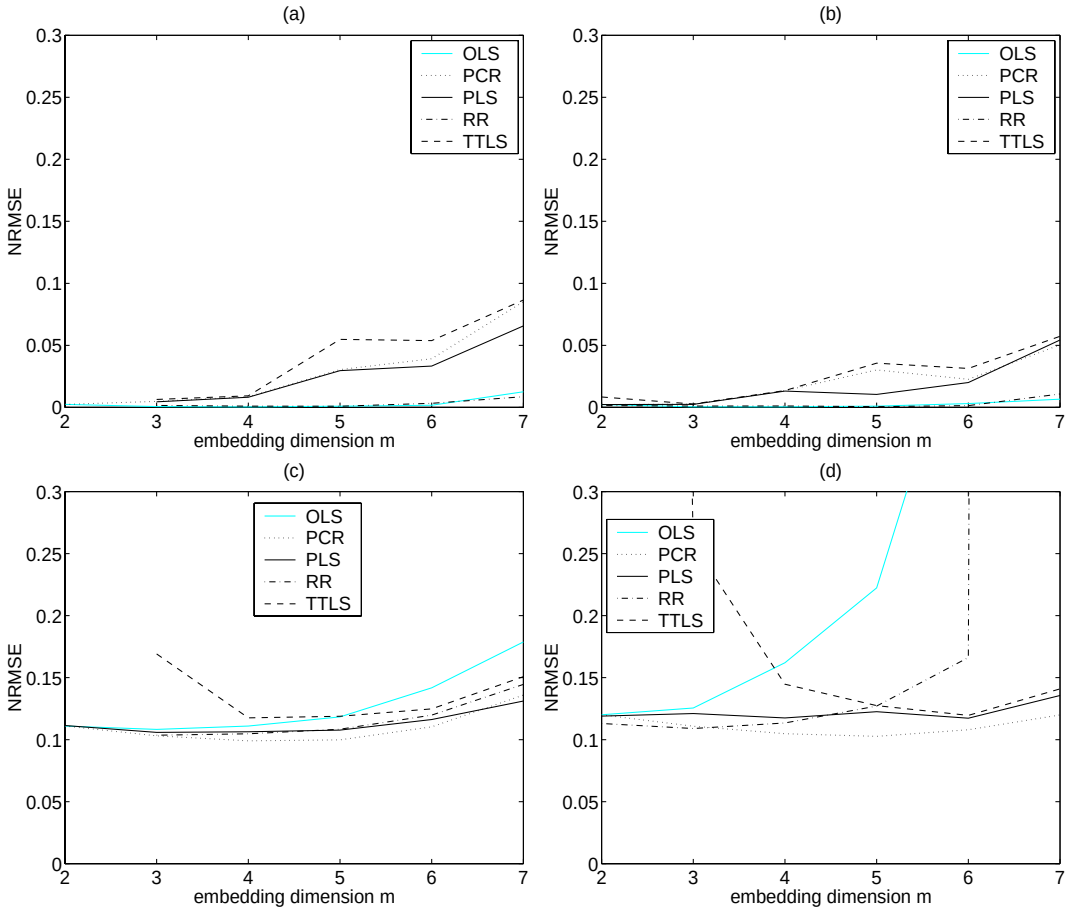

Figure 2: Prediction for different m with OLS and regularised estimates for the Henon data. The first 1500 samples are used to fit the one time step ahead models and the NRMSE is computed on the next 500 data points. (a) Noise-free data, $k=15$. (b) Noise-free data, $k=8$. (c) Data corrupted with 5% normal white measurement noise, $k=15$. (d) Data corrupted with 5% normal white measurement noise, $k=8$. The regularisation parameters are $q=2$ for PCR, PLS and TTLS and $\mu=\hat{s}^2$ for RR.

dimension of the Henon attractor. When the data are noise-free OLS predicts better than PCR, PLS and TTLS, as the regularisations filter out valuable predictive information. However, RR does not seem to shrink the OLS estimator as the estimate $\mu=\hat{s}^2$ is close to zero

because the residual variance is very small. The results do not change significantly when the number of neighbours k drops from 15 to 8, indicating that the linear approximation for the Henon data is good over a range of sizes of the local regions. For noisy data, OLS solutions are more unstable and deteriorate as m increases. In particular, when m approaches k, as in the case of $k=8$ in Fig. 2d, the error gets very large. The PCR predictor is consistently the best for all k and m values, followed closely by the PLS predictor, while TTLS performs worst. The RR prediction fails for m close to k, probably because the residual variance is underestimated and thus the solution is not sufficiently shrunk along the noisy redundant directions.

TTLS does not predict well whenever the data are noisy even though it was initially designed to deal with noise. PCR and PLS perform equivalently well in general. This can be explained by the fact that the response $\mathbf{y}$ and the predictor X are both formed from the same data, so that the correlation between X and $\mathbf{y}$ (on which PLS is based) can be explained in some extent by the correlations within the columns of X (on which PCR is based). The success of the prediction with PCR and PLS relies heavily on the proper selection of the regularisation parameter q. When little is known about the dimension of the underlying system it is safer to make a conservative choice of q to avoid unstable solutions. This choice is doomed to failure for noise-free data, as shown in Fig. 2a and Fig. 2b, but this situation is rather unrealistic in practice as noise is always present. The RR estimate when applied with $\mu=\hat{s}^2$ is classified, with regard to regularisation, in between the OLS estimate and the PLS or PCR estimates (assuming q is small). More elaborate results on chaotic time series including chaotic flows, direct and iterative multi-step predictions and confidence intervals are reported in [9].

3.2 An example with financial data

Chaos theory has been appealing to economists and nonlinear prediction of financial data has become a hot area of research and practice [34, 35]. Local linear prediction has been used in a number of applications and notably for the prediction of exchange rates with reported success [36, 37, 38, 39]. The objective in these works is to find prediction models that perform better than random walk.

The monthly exchange rates of british pound to US dollar (GBP/USD) are used here to illustrate the predictive power of some of the methods discussed in Section 2. The time series of the first differences of the monthly exchange rates is shown in Fig. 3. The data set appears stationary and has very weak correlations if any (the autocorrelation function drops to 0.35 at the first lag and then oscillates around zero).

We used the first 20 years (235 samples) to find the neighbour points for the local models OLS, RR with $\mu = \hat{s}^2$, and PCR with $q = 1$ and $q = 2$, and the rest 10 years (120 samples) to compute the NRMSE of one step predictions. In pursuit for the best parameter setup, we considered three free parameters monitored as follows: the delay time, $\tau=1,2,3,4,5,10,15,20,25,30$ (in accordance with the use of delays in [39]), the embedding dimension, $m=1,\ldots,20$, and the number of neighbours, $k=1,\ldots,20,25:5:110$. For all but very few combinations of τ, m and k (basically for very small m) PCR was by far superior to RR, and RR was slightly better than OLS. As an example, in Fig. 4, the results

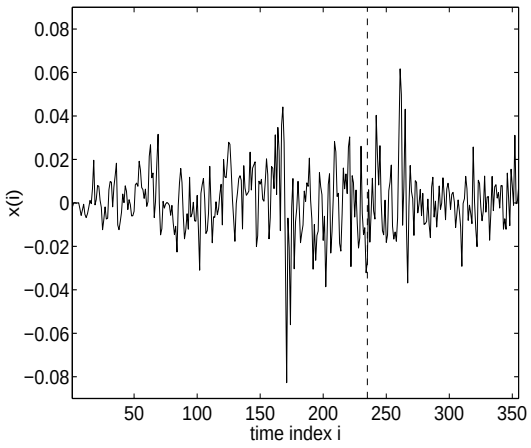

Figure 3: The first differences of the monthly exchange rates GBP/USD from January 1971 to August 2000; in the abscissa is simply the running index of the data. The vertical line at July 1990 distinguishes the learning set from the test set (at time index 235).

for the different m and k are shown for $\tau = 2$. Note that OLS and RR models are not computed for $m \geq k$ and for k close to m the prediction is huge, whereas the PCR predictions deteriorate for k or m values close to q. Moreover, PCR predictions are better than mean value prediction, i.e. $0.93 \leq \text{NRMSE} < 1$, for a long range of k and m values. On the other hand, OLS and RR predictions gave at the best NRMSE $\simeq 1$ only for $m = 1$ and large k (in Fig. 4a and Fig. 4b, respectively, this corresponds to the levelling of the graph of NRMSE to 1 for $m = 1$).

Actually, better predictions were obtained for larger τ for all models and the overall best prediction result was NRMSE$= 0.894$, found with PCR and $q = 1$ for $\tau = 20$, $m = 4$ and $k = 13$. For comparison, the results are shown for the whole range of m values ($\tau = 20$, $k = 13$) in Fig. 5a and for the whole range of k values ($\tau = 20$, $m = 4$) in Fig. 5b. The OLS and RR perform very similarly and much worse than PCR for almost the whole range of m in Fig. 5a and for small k in Fig. 5b. Note that in Fig. 5a, NRMSE increases for larger m because the time window length, $\tau_w = (m - 1)\tau$, spanned by the reconstructed points, approaches the size of the learning set, i.e. the data base of past points from which the 13 neighbours are to be found. Overall, the four models attained best predictions for $\tau = 20$ with OLS and RR giving the same NRMSE$= 0.943$ for $m = 2$ and $k = 70$.

The results above are very optimistic as they suggest that the prediction with local linear models, and particularly using strong regularisation such as PCR and $q = 1$, is better than the mean value prediction by up to 11% (or 20% if NMSE is considered instead, as used in other reports, see [39]). Note that the persistent prediction estimate, $\hat{x}_{i+T} = x_i$, which can be used to estimate the random walk prediction, gives for this test set NRMSE$= 1.2$. However, the significance of NRMSE< 1 obtained for this test set may be attributed solely to some large samples of rate differences in the first part of the test set. Indeed, the prediction results were generally worse when only the last 90 or 60 samples were used as test set. Moreover, we could not explain why the best predictions were obtained for $\tau = 20$ and $m = 4$, which

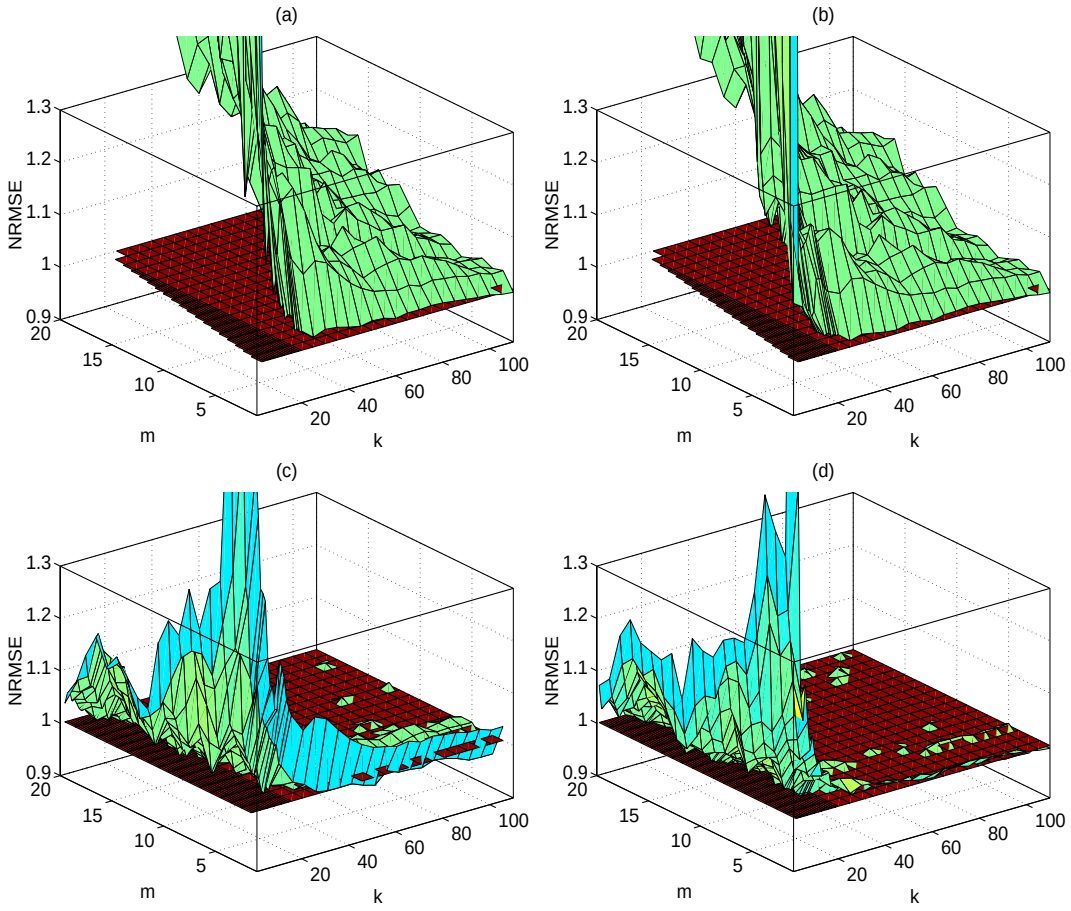

Figure 4: The graph of the NRMSE given as a function of k and m for $\tau = 2$ and for the predictions with OLS in (a), RR ($\mu = \hat{s}^2$) in (b), PCR ($q = 2$) in (c) and PCR ($q = 1$) in (d). The horizontal plane at 1.0 denotes the mean value prediction level, and hides the graph whenever it is larger.

gives a time window τ_w of about 5 years.

4 Discussion

In this chapter, the state space local linear prediction of time series has been discussed with emphasis on regularisation of the standard solution provided by OLS. All the regularisation methods attempt to reduce the variance of the OLS solution, while keeping the bias small. Regularisation certainly improves the prediction of noisy data compared to OLS, with the notable exception of TTLS. Although TTLS is designed to obtain improved parameter estimates when the predictor matrix X is error-corrupted, it turns out to be inappropriate for prediction purposes.

PCR and PLS often give the best local linear predictions, with only marginal differences.

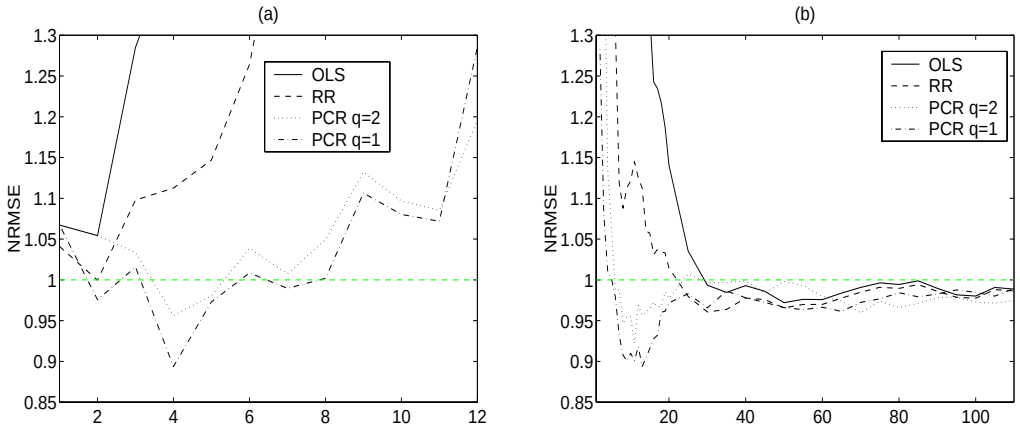

Figure 5: The NRMSE as a function of m ($\tau = 20$, $k = 13$) in (a) as a function of k ($\tau = 20$, $m = 4$) in (b) for the four models as denoted in the legends. The horizontal line denotes the mean value prediction level.

As PCR is by far simpler to implement it turns out to be the method of choice. The PCR solution is derived directly from the SVD of X. As for RR, it performs better than OLS on noisy data but generally worse than PCR and PLS, at least when the regularisation parameter of RR is set to the residual variance.

Best results require careful selection of the regularisation parameter at each target point and further investigation of this is needed. For example, for data giving rise to varying local curvatures, different dimensions of the local state space may be required, and then PCR with fixed q would not give the best results. Each target point poses a separate problem where a different parameter may be the most appropriate. However, cross-validation does not seem to improve the estimation of q.

When m is close to k, OLS deteriorates but the regularised methods do not seem to be affected. This is an important advantage of regularisation because $m \approx k$ may sometimes be desired, when k has to be small (e.g. due to few available data) or when m has to be large (e.g. to resolve completely the attractor). Moreover, the condition $m > k$ is allowed and it suffices that $q < \min(k, m)$ to ensure numerically stable regularised solutions.

Regularisation in local linear prediction is found to be successful whenever the time series is noisy. If incorporated together with other sophisticated approaches, such as search for optimal neighbourhoods or proper weighting of neighbour points, or even for multivariate data analysis, it may turn into a powerful tool for the prediction of apparently random time series, such as the financial data.

Acknowledgements

The author thanks Mike Titterington for his valuable comments on the manuscript.

References

[1] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, 1997.

[2] C. Diks. *Nonlinear Time Series Analysis: Methods and Applications*. World Scientific, 2000.

[3] B. Lillekjendlie, D. Kugiumtzis, and N. Christophersen. Chaotic time series part II: System identification and prediction. *Modeling, Identification and Control*, 15(4):225 – 243, 1994.

[4] A. S. Weigend and N. A. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley Publishing Company, 1994.

[5] J. A. K. Suykens and J. Vandewalle. *Nonlinear Modeling: Advanced Black-Box Techniques*. Kluwer Academic Publishers, 1998.

[6] A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus Illustrations*. Clarendon Press, Oxford, 1997.

[7] H. Tong. *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, New York, 1990.

[8] J. D. Farmer and J. J. Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59:845 – 848, 1987.

[9] D. Kugiumtzis, O. C. Lingjærde, and N. Christophersen. Regularized local linear prediction of chaotic time series. *Physica D*, 112:344 – 360, 1998.

[10] P. J. Brown, editor. *Measurement, Regression, and Calibration*. Oxford University Press, 1993.

[11] G. H. Golub and C. F. Van Loan. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17(6):883 – 893, 1980.

[12] L. Jaeger and H. Kantz. Unbiased reconstruction of the dynamics underlying a noisy chaotic time series. *Chaos*, 6:440, 1996.

[13] D. Yu, W. Lu, and R. G. Harrison. Phase-space prediction of chaotic time series. *Dynamics and Stability of Systems*, 13(3):219 – 236, 1998.

[14] T. Sauer. Time series prediction by using delay coordinate embedding. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, pages 175 – 193. Addison-Wesley Publishing Company, Reading, MA, 1994.

[15] Y-L. Xie and J. H. Kalivas. Local prediction models by principal component regression. *Analytica Chimica Acta*, 348:29 – 38, 1997.

[16] E. N. Lorenz. Atmospheric predictability as revealed by naturally occuring analogies. *Jouranl of Atmospheric Science*, 26:636, 1969.

[17] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Physical Review Letters*, 45:712, 1980.

[18] F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L. S. Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, Lecture Notes in Mathematics 898, pages 366 – 381. Springer, Berlin, 1981.

[19] D. Kugiumtzis. State space reconstruction parameters in the analysis of chaotic time series - the role of the time window length. *Physica D*, 95:13 – 28, 1996.

[20] M. B. Kennel, R. Brown, and H. D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45:3403 – 3411, 1992.

[21] F. Paparella, A. Provenzale, L. A. Smith, C. Taricco, and R. Vio. Local random analogue prediction of nonlinear processes. *Physics Letters A*, 235:233 – 240, 1997.

[22] G. Sugihara. Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions Royal Society London A*, 348:477 – 495, 1994.

[23] G. Sugihara and R. M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344:734 – 741, 1990.

[24] W. A. Fuller, editor. *Measurement Error Models*. Wiley, New York, 1987.

[25] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, Philadelphia, 1998. Monographs on Mathematical Modeling and Computation.

[26] O. C. Lingjærde and N. Christophersen. Shrinkage structure of partial least squares. *Scandinavinan Journal of Statistics*. in press.

[27] C. Goutis. Partial least squares algorithm yields shrinkage estimators. *The Annals of Statistics*, 24(2):816 – 824, 1996.

[28] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Halsted Press, Manchester, 1992.

[29] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55 – 109, 1970.

[30] M. A. Lukas. Comparisons of parameter choice methods for regularization with discrete noisy data. *Inverse Problems*, 14:161 – 184, 1998.

[31] P. Hall and D. M. Titterington. Common structure of techniques for choosing smoothing parameters in regression problems. *Journal of the Royal Statistical Society. Series B*, 49(2):184 – 198, 1987.

[32] K. Ikeda. Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system. *Optics Communications*, 30:257, 1979.

[33] M. Hénon. A two-dimensional map with a strange attractor. *Communications in Mathematical Physics*, 50:69 – 77, 1976.

[34] B. LeBaron. Technical trading rule profitability and foreign exchange intervention. *Journal of International Economics*, 49:125 – 143, 1999.

[35] J. D. Farmer. Physicists attempt to scale the ivory towers of finance. *Computing in Science and Engineering*, 1(6):26 – 39, 1999.

[36] F. X. Diebold and J. A. Nason. Nonparametric exchange rate prediction? *Journal of International Economics*, 28:315 – 332, 1990.

[37] O. Bajo-Rubio, F. Fernandez-Rodriguez, and S. Sosvilla-Rivero. Chaotic behavior in exchange-rate series: First results for the peseta-United States dollar case. *Economics Letters*, 39:207 – 211, 1992.

[38] F. Lisi and A. Medio. Is a random walk the best exchange rate predictor? *International Journal of Forecasting*, 13:255 – 267, 1997.

[39] F. Lisi and R. A. Schiavo. A comparison between neural networks and chaotic models for exchange rate prediction. *Computational Statistics and Data Analysis*, 30(1):87 – 102, 1999.