Regularized local linear prediction of chaotic time series

D. Kugiumtzis *, O.C. Lingjærde, N. Christophersen
Department of Informatics, University of Oslo, PO Box 1080, Blindern, N-0316 Oslo, Norway

Received 5 March 1997; received in revised form 23 June 1997; accepted 23 June 1997
Communicated by A.M. Albano

Abstract

Local linear prediction, based on the ordinary least squares (OLS) approach, is one of several methods that have been applied to prediction of chaotic time series. Apart from potential linearization errors, a drawback of this approach is the high variance of the predictions under certain conditions. Here, a different set of so-called linear regularization techniques, originally derived to solve ill-posed regression problems, are compared to OLS for chaotic time series corrupted by additive measurement noise. These methods reduce the variance compared to OLS, but introduce more bias. A main tool of analysis is the singular value decomposition (SVD), and a key to successful regularization is to damp the higher order SVD components. Several of the methods achieve improved prediction compared to OLS for synthetic noise-corrupted data from well-known chaotic systems. Similar results were found for real-world data from the R–R intervals of ECG signals. Good results are also obtained for real sunspot data, compared to published predictions using nonlinear techniques.

Keywords: Time series; Chaos; Local linear prediction; Regularization

1. Introduction

Nonlinear prediction of chaotic time series has been a popular subject over the last decade [1]. The term “chaotic time series” is applied here in a rather broad sense, covering observations with complex behavior derived from underlying nonlinear phenomena of a presumed deterministic nature. In this field, a variety of more or less complex prediction techniques have been used including neural networks, projection pursuit regression and radial basis functions [2,3]. However, many methods give more or less the same quality of fit as shown in the comparative study in [1]. Applying Occam’s razor in this situation, local linear methods represent simple and attractive alternatives, which have not yet been explored to their full potential in time series prediction.

The ordinary least squares (OLS) solution was applied to chaotic time series several years ago [4], but linear regression and other areas provide a rich set of so-called regularization techniques, developed to tackle ill-posed problems where the regression matrix is poorly conditioned. In prediction with noisy data, the OLS solution can have large variance and the regularization methods, designed to be more robust against noise, may provide better results. Well-known regularization techniques include principal components regression, partial least squares, ridge regression [5], and the truncated total least squares [6].

To our knowledge, a thorough assessment of regularization methods applied to local linear prediction has not previously been attempted. A simple form...
of regularization was applied by Sauer [7], but in a rather ad hoc manner. In nonlinear noise reduction, some techniques include procedures similar to principal components regression, but the regularization property is not emphasized [8].

The linear regression models are presented and compared in Section 2 using the singular value decomposition (SVD). These techniques are studied as part of other ongoing work [9]. Quantitative results for the prediction errors are presently only available under simplifying assumptions neglecting linearization errors and not taking effects of measurement noise fully into account. Simulation studies are presented in Section 3 using synthetic data from several well-known chaotic systems and in Section 4 using the R–R intervals of human electrocardiogram (ECG) data. Real sunspot data are considered in Section 5 where the performance of the local linear techniques is compared to previously published results using nonlinear methods.

2. Local linear prediction and regularization

2.1. The local linear model

Suppose a scalar time series $x_i = x(i \tau_s)$, $i = 1, \ldots, N$, is given, where $\tau_s$ is the sampling time (for maps we have $\tau_s = 1$). Using the method of delays (MOD), the data are represented in $m$-dimensional space by the vectors $X_i = [x_i-(m-1)\tau, x_i-(m-2)\tau, \ldots, x_i]^T$, $i = 1+(m-1)\tau, \ldots, N$, where $m$ is the embedding dimension, $\tau$ is the delay time in units of $\tau_s$ and the superscript $T$ stands for the transpose [10,11]. Given a subset $T$ (the training set) of the vectors in Eq. (1) and a target point $\tilde{x}$, not included in $T$, a local linear prediction of $x_{t+T}$ can be achieved as follows. Let $\tilde{x}_t(1), \ldots, \tilde{x}_t(k)$ be the $k$ nearest neighbors in $T$ to $\tilde{x}$, ordered with respect to increasing distance from the target point. The local linear model is based on centered versions of the matrices $\tilde{X} = (\tilde{x}_t(1), \ldots, \tilde{x}_t(k))^T \in \mathbb{R}^{k \times m}$ and $\tilde{y} = (x_t(1)+T, \ldots, x_t(k)+T)^T \in \mathbb{R}^k$. Let $\tilde{x}$ be the column vector of averages of the $m$ columns of $\tilde{X}$, $\bar{y}$ the average of the components of $\tilde{y}$, and

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

$$x_t = \tilde{x}_t - \bar{x}, \quad y_t = x_{t+T} - \bar{y}. $$

where $1$ is a $k \times 1$ vector of ones, $x_t$ is the centered target point and $y_t$ is the centered response. Scaling columns to equal variance is also frequently done. However, in our case the columns in $X$ are already on the same scale, since they originate from the same time series. The following model is assumed

$$y = Xb + \epsilon,$$

where $b$ is a vector of $m$ unknowns and $\epsilon$ is a random error with expectation $E{\epsilon} = 0$ and covariance matrix $\text{Var}{\epsilon} = \sigma^2 I$, where $I$ is the $k \times k$ identity matrix.

Usually one has $k > m$ in linear regression. It is convenient to retain this constraint in the simulations, although our theoretical treatment is valid also for $k < m$. If the observed time series is corrupted mainly by white measurement noise and the relationship between nearby points on the reconstructed attractor is well approximated by a linear map, there should be some merit in making predictions based on the model in Eq. (2).

By definition of $X$ and $y$, we should expect elements of the former to be corrupted by noise at the same level as elements of the latter. Thus the implicit assumption in Eq. (2) is that $y$ is linearly related to the corrupted $X$ rather than the true $X^{\text{true}}$. A more realistic model would be

$$y = X^{\text{true}}b + \tilde{\epsilon}, \quad X = X^{\text{true}} + \mathcal{E}. $$

Here $E{\tilde{\epsilon}} = 0$, $\text{Var}{\tilde{\epsilon}} = \tilde{\sigma}^2 I$ and the columns of the $k \times m$ matrix $\mathcal{E}$ are uncorrelated and have the same distribution as $\tilde{\epsilon}$. It is easily seen, however, that the model in Eq. (3) is of the form of Eq. (2) with $\sigma^2 = \tilde{\sigma}^2(1 + \|b\|_2^2)$, where $\| \cdot \|_2$ is the $l_2$-norm. In this case, $X$ and $\epsilon$ become correlated which complicates a detailed theoretical analysis of the linear estimators [12]. To simplify further analysis, we use the model in Eq. (2) with the assumption that $X$ is fixed. However, it is our experience from simulations that moderate
Assume in the following that the SVD of $X \in \mathbb{R}^{k,m}$ is defined as

$$X = U \Sigma V^T, \quad U^T U = I, \quad V^T V = I,$$

where $r \leq \min(k, m)$ is the rank of $X$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ 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 $R(X^T) \subset \mathbb{R}^m$, respectively. The subspace $R(X)$ is often referred to as the signal space and its orthogonal complement, the null space $N(X^T)$, as the noise space [13].

### 2.2. Prediction estimators

Suppose $X \in \mathbb{R}^{k,m}$, $y \in \mathbb{R}^k$ and $x_t \in \mathbb{R}^m$ are given and an estimate for $y_t$ is wanted. The prediction estimator is defined as $\hat{y}_t = x_t^T \hat{b}$, where $\hat{b}$ is an estimate for $b$. We consider estimators for $\hat{b}$ of the following form [14]:

$$\hat{b} = V \Sigma^{-1} A U^T y = \sum_{i=1}^r \lambda_i (u_i^T y) v_i,$$

where $A = \text{diag}(\lambda_1, \ldots, \lambda_r)$. This encompasses most of the well-known regularization estimators. Estimators of this form differ only in their choice of the diagonal elements of $A$, called filter factors. Typically $0 \leq \lambda_i \leq 1$ and each $\lambda_i$ then determines the extent of filtering or shrinking in the corresponding singular direction $u_i$. Following the tradition in functional analysis, the inner products $(u_i^T y)$ are denoted the Fourier coefficients for $y$.

The centered target vector $x_t$ may itself be corrupted by noise, i.e.

$$x_t = x_t^\text{true} + \epsilon_t, \quad E[\epsilon_t] = 0 \text{ and } \text{Var}[\epsilon_t] = \sigma^2 I,$$

where $x_t^\text{true}$ represents the true, but unknown, centered target vector (here the identity matrix $I$ is $m \times m$). Assuming that the model in Eq. (2) holds for the true centered target vector, the true hypothetical centered response is $y_t^\text{true} = (x_t^\text{true})^T b$. Ignoring random fluctuations in $X$ (i.e. assuming that $X$ is fixed), the mean-square error in predicting $y_t^\text{true}$ with $\hat{y}_t$ is

$$\text{MSE}(\hat{y}_t) = E[(y_t^\text{true} - \hat{y}_t)^2] = (E[\hat{y}_t] - y_t^\text{true})^2 + E[\hat{y}_t - E[\hat{y}_t]]^2,$$

where $E$ denotes expectation with respect to errors in both $y$ and the target vector in Eq. (5). The second identity in Eq. (6) decomposes MSE as a squared bias term and a variance term. For estimators $\hat{b}$ of the form in Eq. (4) where $A$ is independent of $y$, we have

$$\text{bias}(\hat{y}_t) = E[\hat{y}_t] - y_t^\text{true} = \sum_{i=1}^r (\lambda_i - 1)(v_i^T x_t^\text{true} v_i^T b)$$

and

$$\text{Var}(\hat{y}_t) = E[\hat{y}_t - E[\hat{y}_t]]^2 = \sigma^2 \sum_{i=1}^r \frac{\lambda_i^2}{\sigma_i^2} (v_i^T x_t^\text{true} v_i^T b)^2 + \sigma^2.$$ 

An immediate observation from Eq. (8) is that reduction of the filter factors $\lambda_i$ decreases the variance of the predictor at the possible cost of introducing extra bias in Eq. (7). The optimal trade-off between bias and variance consists of finding filter factors such that the mean-square error in Eq. (6) is minimized.

The bias and variance terms in Eqs. (7) and (8) do not take into account random fluctuations in $X$. Typically $\hat{y}_t$ will depend on $X$ in a complicated manner, making derivations of bias and variance expressions considerably more involved. With multinormal errors in $X$, an exact formula for MSE of the OLS estimator and asymptotic formulas for most of the other predictors considered below are given e.g. in [15]. In our application, linearization errors will also contribute to the prediction uncertainties, but these cannot be quantified in practice.

### 2.3. Ordinary least squares regression

The ordinary least squares (OLS) regression estimate $b_{\text{OLS}}$ is the solution with least $l_2$-norm of the
quadratic minimization problem
\[ \min \| y - Xb \|_2^2 \]

and is easily shown to be
\[ \hat{b}_{OLS} = X^+ y = V \Sigma^{-1} U^T y = \sum_{i=1}^{r} \frac{1}{\sigma_i} (u_i^T y) v_i, \]

(9)

where \( X^+ \) is the Moore–Penrose generalized inverse of \( X \). Note that only the projection of \( y \) into \( R(X) \) (spanned by the columns of \( U \)) contributes to \( \hat{b}_{OLS} \). The filter factors for the OLS estimator are simply \( \lambda_i = 1, i = 1, \ldots, r \) (compare Eq. (9) and Eq. (4)). Without errors in \( X \), it follows immediately from Eq. (7) that the OLS estimator is unbiased. Including such errors, OLS is no longer unbiased and yields estimates where the degree of shrinkage generally increases with the noise level [12].

2.4. Regularization techniques

From Eq. (8) we observe that the prediction variance can be large due to several factors, one being small singular values of the design matrix \( X \). This motivates consideration of regularized estimators that trade bias for variance. Regularization can be achieved by downweighting some or all of the singular directions by choosing filter factors less than 1 in Eq. (4), as is seen from the variance expression in Eq. (8).

One class of estimators (the so-called subspace methods) makes use of a projection \( P \) of \( y \) into a subspace of \( R(X) \). The corresponding estimator for \( b \) is then defined as
\[ \hat{b} = X^+ Py. \]

One such method is principal components regression (PCR), which uses for \( P \) a projection \( P_{PCR} \) into the subspace spanned by the first \( q < r \) singular directions, i.e. the first \( q \) columns in \( U \). For this method, \( \lambda_i = 1, i = 1, \ldots, q \), while \( \lambda_i = 0, i = q + 1, \ldots, r \). The underlying qualitative assumption behind this method is that the projections of \( y \) onto the last \( r - q \) columns of \( U \) are below the noise level and therefore give little or no information about the true \( y \) and \( b \). In other words, \( X \) and \( y \) are assumed compatible, in the sense that \( y \) should have its major component in the subspace spanned by the first \( q \) (and relative to \( X \)) most representative columns of \( U \).

Partial least squares regression (PLS) [16,17] uses instead a projection into the subspace spanned by the Krylov vectors
\[ XX^T y, (XX^T)^2 y, \ldots, (XX^T)^q y, \]

(10)

where the vectors in Eq. (10) are assumed to be linearly independent. Letting \( K \) denote the \( k \times q \) matrix with the vectors in Eq. (10) as columns, the PLS estimator for \( b \) can be written as
\[ \hat{b}_{PLS} = X^+ K (K^T K)^{-1} K^T y. \]

Qualitatively, the PLS estimator shrinks the OLS estimate (in the sense that \( \| \hat{b}_{PLS} \|_2 \leq \| \hat{b}_{OLS} \|_2 \) by taking into account not only the size of the singular values, but also the size of the Fourier coefficients [18]. The expressions for the PLS filter factors are rather involved and can be found in [9].

Another estimator is the ridge regression (RR) estimator [19], which is defined as the solution of the penalized minimization problem
\[ \min \| y - Xb \|_2^2 + \mu \| b \|_2^2, \]

(11)

where \( \mu \) is an additional parameter which could be estimated from the data. The solution of Eq. (11) is easily shown to be
\[ \hat{b}_{RR} = (X^T X + \mu I)^{-1} X^T y. \]

(12)

The inverse in Eq. (12) always exists provided \( \mu > 0 \). The filter factors for the RR estimator are given by \( \lambda_i = \sigma_i^2 / (\sigma_i^2 + \mu) \), showing that the ridge regression estimator shrinks the OLS estimator in every direction when \( \mu > 0 \).

Truncated total least squares (TTLS) [20] is designed to improve the estimates for \( b \) when \( X \) is also contaminated by errors and is therefore a natural candidate for prediction of noisy time series. Note that \( \hat{b}_{OLS} \) solves \( Xb = \tilde{y} \) where \( \tilde{y} \) is the smallest possible perturbation of \( y \) (in the sense that \( \| y - \tilde{y} \|_2 \) is minimum) such that \( \tilde{y} \in R(X) \). Perturbing \( y \) may be seen as an attempt to correct for the noise present in the
observation vector, since \( \hat{y} \) is the projection of \( y \) into the signal space \( R(X) \) [6]. When \( X \) too is contaminated by noise, a more natural approach might be to determine \( b \) from

\[
\hat{X} b = \hat{y},
\]

(13)

where \( [\hat{X} \hat{y}] \) is the smallest perturbation of \( [X \ y] \) such that Eq. (13) becomes solvable, i.e. \( [\hat{X} \hat{y}] \) is a solution of

\[
\text{minimize} \quad \| [X \ y] - [\hat{X} \hat{y}] \|_F
\]

such that

\[
\hat{y} \in R(\hat{X}),
\]

(14)

where \( \| \cdot \|_F \) is the Frobenius norm. Any \( b \) satisfying \( \hat{X} b = \hat{y} \) is called a total least squares (TLS) solution.

The TTLS estimator is found by solving Eq. (14) under the constraint that \( \text{rank}(\hat{X}) = q \), for a given \( q \leq r \). The particular choice \( q = r \) yields the TLS estimator. Let \( [X \ y] \) have the SVD

\[
[X \ y] = U \Sigma \bar{V}^T, \quad \Sigma = \text{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_{r+1}),
\]

(15)

where the singular values \( \hat{\sigma}_i \) are ordered decreasingly \( (\hat{\sigma}_{r+1} > 0 \text{ when } y \notin R(X)) \). Suppose \( \bar{V} \) in Eq. (15) is partitioned as

\[
\bar{V} = \begin{bmatrix}
V_{11} & V_{12} \\
V_{21} & V_{22}
\end{bmatrix},
\]

where \( V_{11} \in \mathbb{R}^{m,q}, \ V_{12} \in \mathbb{R}^{m,r-q+1}, \ V_{21} \in \mathbb{R}^{1,q}, \) and \( V_{22} \in \mathbb{R}^{1,r-q+1} \). Then the TTLS solution exists provided \( V_{22} \neq 0 \), and is given by [6]

\[
\hat{b}_{\text{TLS}} = -V_{12} V_{21}^T.
\]

The filter factors \( \lambda_i, i = 1, \ldots, r \), for TTLS satisfy [14]

\[
1 \leq \lambda_i \leq 1 + \frac{\hat{\sigma}_{q+1}^2}{\sigma_i^2} + O\left(\frac{\hat{\sigma}_{q+1}^4}{\sigma_i^4}\right)
\]

for \( i \leq q \),

\[
0 \leq \lambda_i \leq \|V_{22}\|_2^2 \frac{\sigma_i^2}{\sigma_q^2} \left(1 + O\left(\frac{\sigma_i^2}{\sigma_q^2}\right)\right)
\]

for \( q + 1 \leq i \leq r \).

The first \( q \) filter factors are always larger than 1, while the remaining ones are approximately proportional to \( \sigma_i^2 \) provided that \( \|V_{22}\|_2 \) is not too small. The filter factors above 1 will tend to correct for the shrinkage implicit in the OLS estimate when \( X \) is noise-corrupted.

2.5. Selection of regularization parameters

In selecting the parameter \( q \) in PCR, PLS or TTLS, or the parameter \( \mu \) in RR, we aim at the best possible trade-off between bias and variance. Several model selection criteria exist for linear estimators [21]. One of the most popular is cross-validation (CV), but we found that in the implementation on chaotic time series CV often overestimated the regularization parameters.

A simpler approach is to use the same value for the regularization parameter in all local predictions for a given time series. In chaotic time series, 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 [22].

When the complexity of the underlying system varies locally, it may be more appropriate to let the regularization parameter vary with the target point. If there is a clear gap in the singular value spectrum (this is seldom the case in applications of the type considered in this paper), \( q \) can be chosen as the natural cut-off level in the spectrum. For PCR and TTLS, \( q \) is frequently determined by 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 \( X \).

It was found that the former method gave \( q \) close to \( m \) and the latter gave values of \( q \) increasing with \( m \). A different approach would be to use RR, and estimate \( \mu \) directly from the data involved in each local least squares problem. In an automatic determination of the amount of shrinkage, the modeling and measurement errors (residuals) should be taken into account. The residual variance can be estimated by the projection of \( y \) into the noise space, i.e. \( N(X^T) \). A basis for \( N(X^T) \) is conveniently given by the last \( k - m \) column vectors of \( \bar{U} \in \mathbb{R}^{k,k} \) provided by the expanded SVD of \( X \) \( (X = \bar{U} \bar{\Sigma} \bar{V}, \bar{\Sigma} \in \mathbb{R}^{k,r}) \) [23].
The estimate for the residual variance is

$$\hat{s}^2 = \frac{1}{k-m} \sum_{i=m+1}^{k} (u_i^T y)^2.$$  \hspace{1cm} (17)

The larger the residuals are, the larger the prediction variance becomes and hence the stronger the regularization should be. To adjust the regularization in this way, we propose to use filter factors determined by a sigmoid type of function regulated by $\hat{s}^2$, e.g.

$$\lambda_i = f_\beta(z_i) = 1/(1 + \exp(-\beta z_i)).$$

$$z_i = \ln(\sigma_i^2/\hat{s}^2).$$  \hspace{1cm} (18)

where the parameter $\beta$ determines the slope of the sigmoid, i.e. the length of the intermediate phase. For $\sigma_i^2 \gg \hat{s}^2$, $\lambda_i \approx 1$ while $\sigma_i^2 \ll \hat{s}^2$ gives $\lambda_i \approx 0$ as desired. Note that for $\beta = 1$, one obtains the RR solution with $\mu = \hat{s}^2$. The advantage of this method for our purposes is that the regularization is determined automatically from the data in a less subjective manner (from Eqs. (17) and (18)). This approach will be used with $\beta = 1$ when applying RR.

The filter factors for the regularization techniques we study are shown for an example from a noisy chaotic time series in Fig. 1. Here, the residual standard error $\hat{s}$ is small compared to the singular values $\sigma_i$ and the magnitude of the Fourier coefficients $|u_i^T y|$ (see Fig. 1(a)). For both PLS and TTLS, some of the filter factors $\lambda_i$ are well above one.\(^1\) For TTLS, $\lambda_q$ in particular may be very large, in which case the solution becomes unstable (see Eq. (16)).

### 3. Local linear regression applied to chaotic data

For noise-free data embedded in a state space of sufficiently large dimension $m$, e.g. $m \approx 2[d] + 1$, the attractor often has locally little variance in some directions, in which case the data matrix $X$ becomes ill-conditioned. (Here $d$ is the fractal dimension of the chaotic attractor and $[d]$ is the smallest integer larger than $d$.) For example, when $X$ is derived from a time series generated by the chaotic Henon map [24] ($d \approx 1.21$) with $k = 15$ and $m = 5$, the condition number $\sigma_1/\sigma_3$ is typically of the order of $10^{-4}$. The reason for this is that locally the chaotic attractor is mainly confined to some subspace of $\mathbb{R}^m$. However, small variations outside this subspace may still contain valuable information. As long as the condition number is not

\(^1\)For PLS and TTLS the filter factors were computed with the Matlab routines lsqr and $\text{tl_fac}$ in the regutools toolbox [14].
Fig. 2. Prediction for different $m$ with OLS and PCR ($q = 2$) for 2000 samples generated from the Henon map. The test set is comprised of the last 500 points. The one-step prediction error is measured with NRMSE (see text) for increasing embedding dimension $m$. The four curves correspond to OLS and PCR with $k = 8$ and $k = 15$, as shown in the legend. In (a) the data are noise-free and in (b) normal white noise is added to the noise-free data; the standard deviation of the noise is 5% of the standard deviation of the data.

extreme so that numerical problems are encountered, the OLS solution could still be the best.

The situation becomes different when observation noise is introduced. Although this will tend to make $X$ better conditioned [25] (the condition number is of the order 1 when 5% white noise is added to the Henon data), the prediction capability of OLS deteriorates.

3.1. Illustrating regularization

To illustrate the potential of regularization, Fig. 2 shows results for both noise-free and noisy data from the Henon map. One-step prediction with OLS and PCR is applied for neighborhoods of different sizes ($k = 8$ and $k = 15$) over a range of embedding dimensions ($m = 2, \ldots , 7$). The value $q = 2$ of the regularization parameter corresponds to rather strong regularization and matches the topological dimension of the Henon attractor.

The $T$-step prediction error is measured with the normalized root mean square error

\[
\text{NRMSE} = \sqrt{\frac{1/(N - T - n)}{(N - T - n)}} \sum_{t=n+1}^{N-T} (x_{t+T} - \bar{x}_{t+T})^2
\]

where $\bar{x}$ is the sample mean of the data and $x_{t+T}$, $t = n + 1, \ldots , N - T$ are the response values of the test set ($T = 1$ in Fig. 2). A value of NRMSE near 1 means that the prediction is as good as the mean value prediction. Note the difference in scale of the NRMSE in Figs. 2(a) and (b).

For noise-free data, OLS predicts better than PCR and equally well for the two values of $k$, 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. For $m$ close to $k$, the error can be very large. For example, with $k = 8$ and $m = 6$, the NRMSE is 0.5, while for $m = 7 = k - 1$ the error is 20. The NRMSE for OLS shown in Fig. 2(b) is less for $k = 15$ than for $k = 8$, since in the former case $m$ is small compared to $k$. It has been observed that for larger values of $m$, close to $k = 15$, the NRMSE increases rapidly for OLS. The PCR solutions are consistently better and more stable over the range of $k$ and $m$ values. Note that the best predictions are obtained for $m \approx 4$–5, which is consistent with the criterion $m \geq 2d+1$ [26]. When $d$ or at least $\lfloor d \rfloor$ is known, one could imagine that PCR with $q = \lfloor d \rfloor$ would be reasonable. However, as we will see below, this is not always correct, so the selection of $q$ is basically an open problem.
3.2. Prediction with different regularization methods

Here a more systematic comparison of the regularization methods will be performed, the issue being how to select appropriate techniques with suitable parameter values. First, one-step and direct multi-step prediction are considered. Direct and iterative prediction are then compared and an estimate of the prediction variance in Eq. (8) is applied to a time series.

3.2.1. One-step and direct multi-step prediction

We consider first the typical situation observed with most of the processes studied (e.g. Henon, Lorenz [27] and Mackey-Glass with delay time 30 and 100 units [28]). In Fig. 3, the results are shown for data generated from the Lorenz system and corrupted with 5% noise. The regularization parameters for PCR, PLS and TLS are all set to $q = 2$, because the Lorenz attractor is well approximated locally by a two-dimensional plane. For RR, we use $\mu = s^2$ (see Eq. (17)).
For one-step prediction (Fig. 3(a)), all regularization methods are slightly better than OLS. For multi-step prediction (Figs. 3(b) and (c)), TTLS is inappropriate, while the other regularization methods perform similarly and always better than OLS. Moreover, their advantage over OLS becomes more pronounced with larger values of $m$.

The Ikeda map \cite{29} illustrates a more complicated situation. The topological dimension $[d]$ of this attractor is 2, but 4 has been suggested as the minimum $m$ for successful state space reconstruction \cite{30}. This implies that the choice $q = [d]$ may not be suitable here, which is also confirmed by our simulations. Results are shown in Fig. 4. A milder regularization with $q = 3$ is used for PCR, PLS, and TTLS, while for RR $\mu = \sigma^2$ as before. Considering one-step prediction in Fig. 4(a), TTLS performs rather poorly for all embedding dimensions, while the other regularization methods behave similarly to OLS. For direct multi-step prediction with $m = 4$,
Fig. 5. Iterative multi-step prediction with different regularization techniques for data generated from the Lorenz system and corrupted with 5% noise ($N = 2000$, $n = 1500$ and $k = 15$). In (a) the NRMSE as a function of $T$ is shown for $m = 4$, in (b) for $m = 7$ and in (c) for $m = 10$. OLS, PCR, PLS and RR are shown according to the legend. For PCR and PLS $q = 2$ and for RR $\mu = \hat{s}^2$.

which represents the best choice of $m$, the results are rather similar as expected, except for TTLS (see Fig. 4(b)). Increasing the embedding dimension to $m = 7$, the regularized solutions are clearly preferable (see Fig. 4(c)).

All results in Figs. 3 and 4 for PCR, PLS and TTLS are with fixed regularization parameters. CV has also been tried and it turns out that the variation of the parameter for a given regularization method is large. No improvements were obtained over the results in Figs. 3 and 4 using CV.

3.2.2. Iterative multi-step prediction

Discrepancies between OLS and the regularization techniques are small for $T = 1$ and grow with $T$ when direct multi-step prediction is employed. When iterative multi-step prediction is used instead, the discrepancies seem to grow relatively slower. In Fig. 5, iterative multi-step predictions with different regularization techniques are shown for the Lorenz data for three different values of $m$. The iterative $T$-step prediction involves $T$ successive one-step predictions, where at each step the previously predicted values are used.
as coordinates for the new target point. Comparing the iterative predictions for $m = 4$ and $m = 7$ in Figs. 5(a) and (b) with the corresponding direct predictions in Figs. 3(b) and (c), respectively, there is less improvement with regularization using the iterative scheme. However, as $m$ increases further, the differences become more pronounced (see Fig. 5(c)). Here, PCR and PLS clearly perform better than RR. Note that with larger $m$, the predictability of the regularized models farther into the future is enhanced, whereas for OLS the opposite is true (see for example Figs. 5(b) and (c)). This is observed in general with data from continuous systems.

In local prediction with OLS, the iterative scheme has been reported to be superior to the direct scheme [31,32]. However, our study using different systems shows that this is not always the case (e.g. compare the results in Figs. 3(b) and 5(a) for $T = 10$).

### 3.2.3. Prediction confidence bands

The theoretical expression for the prediction variance was given in Eq. (8). This expression may not be very reliable because the uncertainty in the predictor matrix $X$ and the linearization errors are not considered, but at least it can illustrate differences in the prediction variance of the models. Estimated prediction variance is shown in Fig. 6 for the Lorenz data using OLS and PCR ($q = 2$). In the computation of $\text{Var}[\hat{x}_{T+T}]$ from Eq. (8) (where $\text{Var}[\hat{x}_{T+T}] = \text{Var}[\hat{\gamma}_T + \hat{\gamma}] = \text{Var}[\hat{\gamma}_T])$, $\mathbf{x}_t$ is replaced by the noisy $\mathbf{x}_t$, $\mathbf{b}$ by the OLS and PCR estimate and $\sigma$ by 5% of the standard deviation of the data. The prediction estimates $\hat{x}_{T+T}$ for $T = 1$ and $T = 3$ are shown together with the confidence interval defined as $\hat{x}_{T+T} \pm 2\sqrt{\text{Var}[\hat{x}_{T+T}]}$. Regularization decreases the variance significantly. However, the variances of the estimates both for OLS and PCR grow large with the prediction time $T$, especially for regions on the attractor where the predictions are bad. Note that the prediction variance may be estimated a priori, i.e. before the prediction is made, in order to anticipate the quality of the forthcoming prediction.

### 4. Prediction of the R–R intervals of ECG data

Here, we demonstrate the improved performance of the regularized methods over OLS for the R–R intervals of human ECG data. The R–R intervals measure the length of the cardiac cycles from the interval between two successive ventricular beats [33]. In some papers, it has been suggested that heart rhythms may
In conclusion, we report that predictions with a global linear autoregressive (AR) model were almost as good as PCR and significantly better than OLS for larger $m$ ($m > 4$). In the prediction of normal heart rhythms in [37], it was found using $m = 7$ that multi-step predictions with AR were slightly worse than those of the local simplex method [38], which agrees with our findings using the regularization techniques. We can therefore conclude that OLS is the least appropriate local linear model for predicting heart rhythms and the regularization of OLS is demanded for this type of data.

We found similar qualitative results as those shown in Fig. 8 for other real world data such as the southern oscillation index (SOI) record [39] and the blowfly data [40].

5. Prediction of sunspot data

As a preliminary application of these methods to the prediction of real data, we chose the annual mean sunspot data (Wolf numbers), a benchmark data set in time series prediction. This time series is from 1700 to the present. The nature of the underlying process is still under investigation [41–43] and many different prediction models have been applied to these data. Threshold autoregressive (TAR) models with $m = 12$, corresponding to a “year cycle” of 11–12 years, were found to give good predictions [21,44]. Later studies with other nonlinear models, such as neural networks (NN) [45], multivariate adaptive regression splines (MARS) [46], and local linear prediction with OLS [47,48], reported equally good or better results. We adopt here the prediction error measure used in most of the previous studies, referred to as average relative variance (ARV), which is the square of NRMSE.

For comparison purposes, we use the prediction setup in [45]. There, as in many previous studies, the series was split in two intervals, 1700–1920 for training and 1921–1955 for testing. Iterative multi-step prediction was employed. We chose $k = 30$ in order to be consistent with [49], where this was shown to be the optimal $k$ for OLS. This is a rather large value (comprising 15% of the training set), but our results were
not very sensitive to the choice of $k$. Most regularization techniques gave an ARV for $T = 1$ around 0.1 with best results for small $m$. For example, PCR with $q = 3$ and $m = 4$ gave ARV = 0.078, to be compared with ARV = 0.097 for TAR and ARV = 0.086 for the NN in [45].

For one-step prediction, a small $m$ (from 3 to 5) gives best predictions (regardless of the method applied), while for multi-step prediction a larger $m$ (up to 7–8) gives best results, and for $m > 8$ both one and multi-step predictions become worse. Simulations with different methods and various training and test sets showed that PCR with $m = 7$ and $q = 3$ gave generally good multi-step predictions. The results for OLS ($m = 7$) and PCR ($q = 3, m = 7$) as well as for TAR and the weight-eliminated NN with $m = 12$ used in [45] are shown in Fig. 9(a).² For OLS, the choice

²The ARV values for the NN model were copied directly from Fig. 5 in [45].
$m = 7$ was the best for the chosen training and test set. The PCR model was not optimized for this particular setup. As shown in Fig. 9(a), PCR is at least as good as NN (slightly worse than NN for $T < 9$ and slightly better for $T > 9$), certainly better than OLS and superior to TAR for all but $T = 1$.

The consistently good predictions with $q = 3$ support the hypothesis of a low-dimensional system of $\lceil d \rceil = 3$, which is close to the estimated dimension $\lceil d \rceil = 4$ in [50]. This also seems intuitively correct, as the oscillatory behavior of the sunspot series is reminiscent of well-studied systems of dimension $2 \leq d \leq 3$, e.g. the 2- and 3-torus and the Lorenz and Rössler systems [51] etc.

We used PCR with $q = 3$ and $m = 7$ to predict the two forthcoming sunspot cycles based on annual data up to 1995. The point estimates are shown in Fig. 9(b) with bars representing the estimated confidence band for the one-step prediction (conditioned on the last predicted values) $\hat{x}_{t+1} \pm 2\sqrt{\text{Var}(\hat{x}_{t+1})}$ (see Eq. (8)), where $\sigma$ was estimated here with $\hat{s}$. Note that the two cycles are reproduced and that the predictions of the ascent period have large variance, as discussed by other authors [43].

6. Discussion

Regularization certainly improves the prediction of noisy data compared to OLS, with the notable exception of TTLS. Although TTLS is designed to obtain improved estimates for $b$ when $X$ is error-corrupted, it turned out to be unsuitable for local linear prediction. This is not surprising considering the fact that the first $q$ filter factors always are larger than 1, i.e. TTLS does the opposite of regularization in these directions. Differences among the other techniques are small and not systematic. The discrepancies may be explained by the local curvature of the attractor around the target point and by the location of the neighbor points and the way these are contaminated by noise. For example, varying local curvature may require different dimensions of the local state space, and then PCR with fixed $q$ would not give the best results.

We have suggested an automatic scheme (see Eqs. (17) and (18)) that changes the amount of regularization depending on the singular values and the residuals for the local least squares problem. This approach consistently gives fairly good predictions and does not require any parameter determination. On the other hand, for most systems we tested, PCR
and PLS with $q = \lceil d \rceil$ gave the best predictions with only marginal differences. However, in a real application, if $q$ is not indicated by a dimension estimation method, different values of $q$ must be tried when predicting with PCR and PLS. Cross validation does not seem to improve the procedures for choosing $q$ here. Each target point poses a separate problem where a different parameter may be the most appropriate.

When $m$ is close to $k$, OLS deteriorates but the regularized methods do not seem to be affected. This is an important advantage of regularization 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. due to a large dimension of the underlying attractor). Actually, the condition $m > k$ is allowed and it suffices that $q < \min(k, m)$ to assure numerically stable regularized solutions. This means that even with restricted $k$ values (e.g. with short time series), multi-step predictions that may require large $m$ can still be obtained.

For data generated by continuous systems, it seems that a state space reconstruction with large $m$ enables better multi-step predictions. For example, we found that for the noisy Lorenz data, reconstruction with a time window length $\tau_w = (m - 1)\tau$ close to the mean orbital period $\tau_p$ (corresponding to $m = 10$ in Fig. 5(c)) gave better predictions for large $T$ than with a smaller $\tau_w$. The choice $\tau_w \approx \tau_p$ has previously been found appropriate for reconstruction in the estimation of the correlation dimension [52]. However, for the sunspot data a smaller $\tau_w$ ($\tau_w = 6$ years corresponding to $m = 7$) compared to $\tau_p = 11$ years was found to be more appropriate, but this may be due to the limited number of cycles observed. The parameter $\tau$ is usually overlooked in work on prediction, probably because the studied time series are sampled coarsely. When the sampling is dense, $\tau$ should be large enough to avoid excessively large values of $m$ compared to $k$.

This implementation of regularization in local linear prediction gave promising results and motivation for further investigations, such as search for optimal neighborhoods [53] or weighting of the neighbors [54]. It would be interesting to examine the long-term predictability with regularized methods, i.e. their ability to reproduce the original dynamics. Moreover, regularized methods may be implemented in other areas where local linear models are used, such as estimation of Lyapunov exponents, noise reduction, and control. For such purposes, it has been noted that models taking into account the errors-in-variables problem are the most appropriate [55,56]. However, our analysis showed that the TTLS model, which is especially designed to solve the errors-in-variables problem, is the worst prediction model. Further analysis is needed to clarify this point.

7. Conclusion

Several local linear methods designed for ill-posed regression problems have been tested in the prediction of noisy chaotic time series. The regularization techniques have been described within a uniform framework using the SVD. All the regularization methods attempt to reduce the variance of the OLS solution, while keeping the bias small. For most methods, this trade-off between bias and variance is done by damping the higher order SVD components.

Simulations with well-known chaotic systems showed that regularization is beneficial when the data are corrupted with noise. In such cases, all methods except TTLS yield improved predictions compared to OLS. Best results require careful selection of the regularization parameter at each target point. However, PCR and PLS with a properly chosen fixed parameter $q$ often turned out to be the methods of choice. For RR, we suggest to estimate the shrinking parameter $\mu$ from the variance of the residual, which is certainly different for each local least squares problem. In this way, the regularization effect varies with the target point.

The regularized techniques were implemented on real-world data, such as the R–R intervals of human ECG data and the sunspot data, and gave improved predictions over OLS. Particularly for the sunspot data, they gave good predictions compared to nonlinear techniques believed to be the most powerful for these data. The results with synthetic and real data showed that there is no universally best regularization method and testing and comparison are necessary for each new application.
Acknowledgements

The first author thanks the Norwegian Research Council (NFR) for partial support of this work. The authors thank A. Tsonis for providing the SOI data.

References


