Published in: Neural Computation 10, 731-747 (1998)

Nonlinear Time-Series Prediction with Missing and Noisy Data

Volker Tresp and Reimar Hofmann
Siemens AG, Corporate Technology
Department Information and Communications
81730 Munich, Germany*

Abstract

We derive solutions for the problem of missing and noisy data in nonlinear time-series prediction from a probabilistic point of view. We discuss different approximations to the solutions, in particular approximations which require either stochastic simulation or the substitution of a single estimate for the missing data. We show experimentally that commonly used heuristics can lead to suboptimal solutions. We show how error bars for the predictions can be derived and we show how our results can be applied to K-step prediction. We verify our solutions using two chaotic time series and the sun-spot data set. In particular, we show that for K-step prediction stochastic simulation is superior to simply iterating the predictor.

1 Introduction

Over the past years, neural networks have been applied successfully in numerous applications to nonlinear time-series prediction (Weigend and Gershenfeld, 1994). Common problems in time-series prediction are missing and noisy data. The goal is to obtain optimal predictions even if some measurements are unavailable, are not recorded or are uncertain. For linear systems, efficient algorithms exist for prediction with missing data (Kalman, 1960, Shumway and Stoffer, 1982). In particular, the Kalman filter is based on a state space formulation and achieves optimal predictions with arbitrary patterns of missing data. For nonlinear systems, the *extended* Kalman filter can be employed which is based on a first order series expansion of the nonlinearities. The extended Kalman filter is suboptimal (Bar-Shalom and Li, 1993) and summarizes past data by an estimate of the means and the covariances of the variables involved. The extended Kalman filter fails to

*This work was supported by grant number -01 IN 505 A9- from the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie. E-mail: Volker.Tresp@mchp.siemens.de, Reimar.Hofmann@mchp.siemens.de

give good predictions if the system is not approximated well by a localized linearization, i.e. for highly nonlinear systems, in particular if the inaccuracies in the approximations propagate through several iterations, as in K-step prediction. In this paper we propose stochastic sampling which converges to the optimal solution as the number of samples approaches infinity and can handle arbitrary patterns of noisy and missing data. We demonstrate the benefits of stochastic sampling using three examples.

The related issue of training a time-series model with missing and noisy data will be addressed in a companion paper (Tresp and Hofmann, 1997).

In Section 2 we derive equations for prediction with missing data. As in the case of regression and classification with missing data (Little and Rubin, 1987, Ahmad and Tresp, 1993, Buntine and Weigend, 1991), the solution consists of integrals over the unknown variables weighted by the conditional probability density of the unknown variables given the known variables. In time-series prediction we can use the fact that the unknown data themselves are part of the time series. By unfolding the time-series in time we obtain a Bayesian network (Pearl, 1988, Jensen, 1996) (a probabilistic graph with directed arcs) which allows us to clarify dependencies between the variable to be predicted and the measurements which provide information about that variable. In Section 3 we generalize the results towards noisy measurements. For nonlinear systems, the integrals cannot be solved in closed form and have to be approximated numerically. In Section 4 we propose stochastic sampling which has the advantage that asymptotically (i.e. with the number of samples approaching infinity) we obtain the optimal prediction. As an alternative approximation, we propose that maximum likelihood estimates can be substituted for the missing data. Furthermore, we discuss solutions based on an iterative approximation of the information provided by past data using probability density estimates. In Section 5 we present experimental results demonstrating the superiority of the stochastic sampling approach. In particular, we show that for K-step prediction, stochastic sampling is superior to both simply iterating the system and the extended Kalman filter (the latter two turn out to be identical for K-step prediction). In Section 6 we present conclusions.

2 Prediction with Missing Data

2.1 An Illustrative Example

Consider the situation depicted in Figure 1, top. The time series model is

$y_t = f(y_{t-1}, y_{t-2}) + \epsilon_t$

where ϵ_t is additive i.i.d. noise and $f()$ is a nonlinear function. The goal is to predict y_t based on past measurements. Let's assume that y_{t-2} is missing. A common procedure is to obtain an estimate $\hat{y}_{t-2}$ of the missing value and then substitute that estimate in the predictive model

$\hat{y}_t = f(y_{t-1}, \hat{y}_{t-2})$.

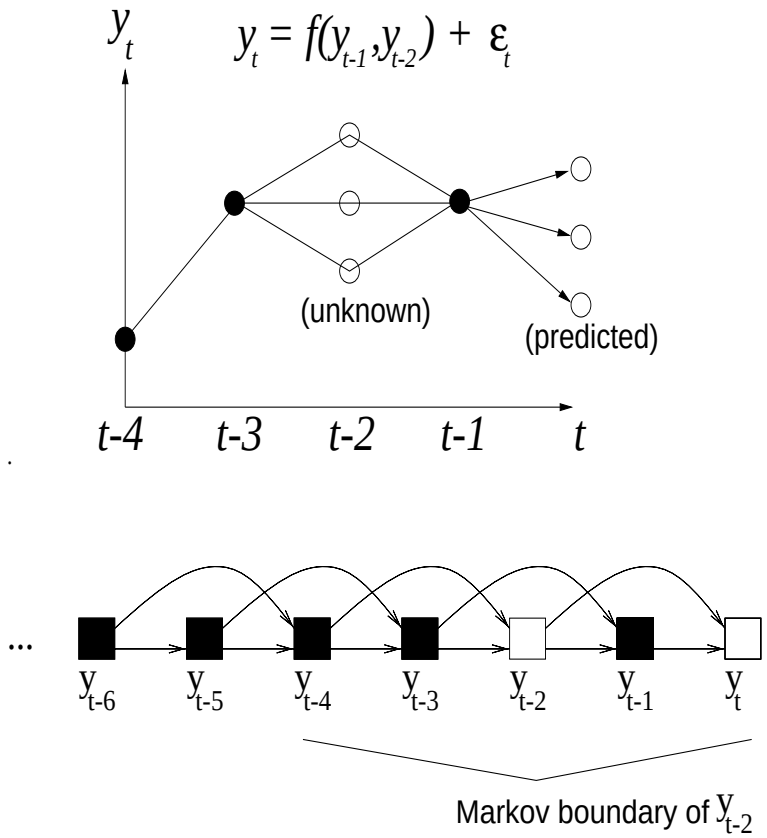

Figure 1: Top: y_{t-2} is missing and the goal is to predict y_t. The estimate $\hat{y}_t$ is dependent on the substituted value for y_{t-2}. Bottom: A time series unfolded in time. White squares indicate unknown variables and black squares indicate measured variables. The arrows indicate that the next realization of the time series can be predicted from only the two most recent values, $y_t = f(y_{t-1}, y_{t-2}) + \epsilon_t$. Here, y_{t-2} is assumed to be missing. The bracket indicates the nodes in the Markov boundary of y_{t-2} (see Section 4.1).

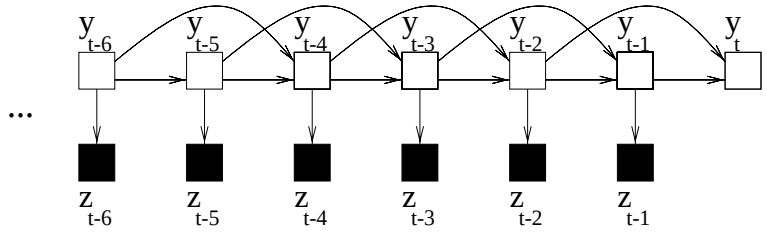

Figure 2: The figure displays the Bayesian network corresponding to the problem of time-series prediction with noisy measurements ($N = 2$). White squares indicate unknown variables and black squares indicate measured variables.

In some applications it might make sense to substitute for the missing value the previous value $\hat{y}_{t-2} = y_{t-3}$ or to substitute the predicted value $\hat{y}_{t-2} = f(y_{t-3}, y_{t-4})$. Both heuristics might often work in practice but note the following two points:

- since in our example y_{t-1} is known, it should improve our estimate of y_{t-2},
- since y_{t-2} is only estimated, it should be possible to achieve better predictions by not just substituting one estimate but several estimates and by then averaging the predictions based on those estimates.

In the following sections we will show that a theoretical analysis confirms these intuitions.

2.2 Theory

Let y_t be the value of the discrete time-series at time t. We assume that the underlying probabilistic model of the time series is of order N and can be described by

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-N}) + \epsilon_t \tag{1}$$

where $f()$ is either known or approximated sufficiently well by a function approximator such as a neural network. ϵ_t is assumed to be additive i.i.d. zero-mean noise with probability density $P_\epsilon(\epsilon)$ and typically represents unmodeled dynamics. The conditional probability density of the predicted value of the time series is then

$$P(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-N}) = P_\epsilon(y_t - f(y_{t-1}, y_{t-2}, \ldots, y_{t-N})). \tag{2}$$

Often, Gaussian noise is assumed such that

$$P(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-N}) = G(y_t; f(y_{t-1}, \ldots, y_{t-N}), \sigma^2) \tag{3}$$

where $G(x; c, \sigma^2)$ is our notation for a normal density evaluated at x with center c and variance σ^2.

It is convenient to unfold the system in time which leads to the system shown in Figure 1, bottom. The realizations of the time series can now be considered random variables or nodes in a Bayesian network, in which directed arcs indicate direct dependencies (Pearl, 1988). The joint probability density in a Bayesian network is the product of all conditional densities and the prior probabilities

$$P(y_1, y_2, \ldots, y_t) = P(y_1, \ldots, y_N) \prod_{l=N+1}^{t} P(y_l|y_{l-1}, \ldots, y_{l-N}) \tag{4}$$

where $P(y_1, \ldots, y_N)$ is the prior probability of the first N values of the time series.

We use the following notation: $Y^u_{t_2,t_1} \subseteq \{y_{t_1}, y_{t_1+1}, \ldots, y_{t_2}\}$ is the set of missing variables from t_1 to t_2, $Y^m_{t_2,t_1} \subseteq \{y_{t_1}, y_{t_1+1}, \ldots, y_{t_2}\}$ is the set of measurements between t_1 and t_2 and $Y_{t_2,t_1} = Y^m_{t_2,t_1} \cup Y^u_{t_2,t_1}$ $(t_1 \leq t_2)$.

The theory of Bayesian networks is helpful to decide, which past measurements provide information about y_t. Let A and B be nodes in a directed acyclic graph D (in our case a Bayesian network). A and B are independent given the evidence entered into the network if they are d-separate. The definition of d-separation is (Pearl, 1988, Jensen, 1996):

DEFINITION(d-separation): *Two variables A and B in a directed acyclic graph are* d-separated *if for all paths between A and B there is an intermediate variable V such that either*
(1) the connection is serial or diverging and the state of V is known
or
(2) the connection is converging and neither V nor any of Vs descendents have received evidence.[1]

In other words, A and B are *d-separated* if every path between both nodes is blocked by either condition (1) or (2). An example of a serial connection is $\rightarrow V \rightarrow$, of a diverging connection is $\leftarrow V \rightarrow$ and of a converging connection is $\rightarrow V \leftarrow$. We now apply the concept of d-separation to time-series prediction. Let y_{t-L} be the most recent case, where N consecutive measurements are known, i. e. $y_{t-L}, y_{t-L-1}, \ldots, y_{t-L-N+1}$ are all known. In this case, y_t is d-separate from measurements previous to $t-L-N+1$ given $y_{t-L}, y_{t-L-1}, \ldots, y_{t-L-N+1}$. Consider Figure 1 (bottom). Here, y_{t-5} is d-separated from y_t by y_{t-3} and y_{t-4} since these nodes block all paths from y_{t-5} to y_t. The same d-separation is true for all measurements previous to y_{t-5}. y_{t-4}, on the other hand is not blocked by y_{t-3} and y_{t-1} since there is the path $y_{t-4} \rightarrow y_{t-2} \rightarrow y_t$ which is not blocked.

Following the discussion in the previous paragraph, y_t is independent of measurements earlier than $y_{t-L-N+1}$ given $y_{t-L}, y_{t-L-1}, \ldots, y_{t-L-N+1}$. This means that we have to condition y_t only on measurements $Y^m_{t-1,t-L-N+1}$ and we obtain for the expected value of the next realization of the time series

$$E(y_t|Y^m_{t-1,1}) = \int y_t P(y_t|Y^m_{t-1,t-L-N+1}) dy_t \tag{5}$$

$$= \int f(y_{t-1}, \ldots, y_{t-k}, \ldots, y_{t-N})\ P(Y^u_{t-1,t-N}|Y^m_{t-1,t-L-N+1})\ dY^u_{t-1,t-N}$$

$$= \int f(y_{t-1}, \ldots, y_{t-k}, \ldots, y_{t-N})\ P(Y^u_{t-1,t-L+1}|Y^m_{t-1,t-L-N+1})\ dY^u_{t-1,t-L+1}$$

where (assuming $t-L \geq N$)

$$P(Y^u_{t-1,t-L+1}|Y^m_{t-1,t-L-N+1}) = \frac{1}{const} \times \prod_{l=t-L+1}^{t-1} P(y_l|y_{l-1}, \ldots, y_{l-N})$$

and $const = P(Y^m_{t-1,t-L+1}|Y^m_{t-L,t-L-N+1})$ is a normalization constant independent of the unknown variables.

[1] In our case this means that neither V nor any of Vs descendents are known.

3 Prediction with Noisy Measurements

Let again $y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-N}) + \epsilon_t$ but now we assume that we have no access to y_t directly. Instead, we measure $z_t = y_t + \delta_t$ where δ_t is independent zero-mean noise (Figure 2) with probability density $P_\delta(\delta)$. Let $Z_{t-1,1} = \{z_1 \ldots z_{t-1}\}$ and $Y_{t,1} = \{y_1 \ldots y_t\}$. The joint probability density is

$$P(Y_{t,1}, Z_{t-1,1}) = P(y_1, \ldots, y_N) \prod_{l=N+1}^{t} P(y_l|y_{l-1}, \ldots, y_{l-N}) \prod_{l=1}^{t-1} P(z_l|y_l) \tag{6}$$

with $P(z_l|y_l) = P_\delta(z_l - y_l)$. The corresponding Bayesian network is shown in Figure 2. Note, that for each known variable z_{t-k} there is a path to y_t which is not blocked by any of the other known variables and which has no converging arrows, i.e. the path $z_{t-k} \leftarrow y_{t-k} \rightarrow y_{t-k+1} \rightarrow \ldots \rightarrow y_t$. This means that y_t is dependent on all past measurements.

The expression for the expected value of the next instance of the time series (prediction) is then

$$E(y_t|Z_{t-1,1}) = \int f(y_{t-1}, \ldots, y_{t-N})\ P(Y_{t-1,t-N}|Z_{t-1,1})\ dY_{t-1,t-N} \tag{7}$$

$$= \int f(y_{t-1}, \ldots, y_{t-N})\ P(Y_{t-1,1}|Z_{t-1,1})\ dY_{t-1,1}$$

where $P(Y_{t-1,1}|Z_{t-1,1}) = 1/const \times P(Y_{t-1,1}, Z_{t-1,1})$ which is obtained from Equation 6. $const = P(Z_{t-1,1})$ is a normalization constant independent of $Y_{t-1,1}$. Note, that the case of noisy measurements includes the case of missing data. In particular, if we allow the measurement noise to be time-dependent (which does not introduce any additional complexity) we can use $\sigma_\delta^2(t) = 0$ for certain measurements and $\sigma_\delta^2(t) = \infty$ for unknown data.

4 Approximations to the Theoretical Solutions

In general, if $f()$ is a nonlinear function the equations (5) and (7) we obtained for prediction cannot be solved analytically and must be approximated numerically. First, we propose an approximation based on stochastic simulation which provides the optimal prediction when the number of samples approaches infinity. As a second approximation, we discuss an approach where the most likely values are substituted for the missing data. The latter approach tends to be computationally less expensive but provides biased predictions. Finally, we discuss the extended Kalman filter which can be used on-line and is based on a first order series expansion of the nonlinearities.

4.1 Stochastic Simulation

We will discuss a solution based on stochastic simulation. Note that all solutions have the general form $\int h(U, M)P(U|M)dU$ where U is a set of unknown variables and M is a set of known variables. An integral of this form can be solved by drawing random samples of the unknown variables following $P(U|M)$. Let $U^1, \ldots, U^S$ denote these samples. Then we can approximate

$$\int h(U, M)P(U|M)dU \approx \frac{1}{S}\sum_{s=1}^{S} h(U^s, M).$$

The problem now reduces to sampling from $P(U|M)$. Let's first assume that only one variable is missing. Then the problem reduces to sampling from a one-variate distribution which can be done using *sampling-importance-resampling* or other sampling techniques (Bernardo and Smith, 1994).

If more than one realization is missing the situation becomes more complicated. The reason is that the unknown variables are in general dependent and we have to draw from the joint probability distribution of all unknowns. A general solution to this problem is Markov Chain Monte Carlo sampling, with the Metropolis-Hastings algorithm and Gibbs sampling being the two most important representatives. We will briefly describe the latter.

In Gibbs sampling we initialize the unknown variables either randomly or better with reasonable initial values. Then we select one of the unknown variables $u_i \in U$ and pick a sample from the one-dimensional conditional density $P(u_i|MB(i))$ and set u_i to that value. $MB(i)$ is the Markov boundary of u_i.[2] Then we select another unknown variable u_j, pick a sample from $P(u_j|MB(j))$ and set u_j to that value. We repeat the procedure for another unknown variable and so on. In this way, repeated samples of all unknowns are drawn. Discard the first samples since they strongly depend on which initial values were chosen. Then, for strictly positive distributions, samples are produced with the correct distribution, that is for $s \to \infty$, U^s tends in distribution to a joint random vector whose joint density is $P(U|M)$ (Bernardo and Smith, 1994). Gibbs sampling reduces the problem of drawing a sample from the joint density of all unknowns to sequentially drawing samples from the univariate densities of each unknown conditioned on the variables in its Markov boundary.

In the case of missing data, we have to generate samples from all missing data $Y^u_{t-1,t-L+1}$. In the case of noisy measurements we even have to sample from all $Y_{t-1,1}$. In practice, one would restrict the sampling to a reasonably chosen time window in the past.

[2]We only have to condition on the nodes in the Markov boundary since, by definition of the Markov boundary, under the assumption that all nodes in the Markov boundary are known, the node u_i is d-separated from the remaining variables in a Bayesian network. The Markov boundary of a node consists of its direct parents, its direct successors and all direct parents of its direct successors (Pearl, 1988) (as example, see Figure 1).

For independent samples, the variance of an estimated mean is equal to σ_s^2/S where σ_s^2 is the variance of an individual sample. Unfortunately samples generated by Gibbs sampling and other Markov Chain Monte Carlo sampling techniques are typically highly correlated such that —depending on the particular problem— a large number of samples might be required for a good estimate. This is particularly true if regions of high probability are separated by regions of low probability such that the transition between regions has low probability. Another disadvantage is that for each new prediction we have to perform a separate sampling process. Neal (1993) discusses hybrid Monte-Carlo methods and other advanced sampling techniques which try to overcome some of the difficulties associated with dependent samples.

Sampling is simple if only samples of *future* values are required as in K-step prediction (for details, see Section 5.1). The reason is that we can sample forward in time by simply simulating the system. Note that in this procedure, independent samples are generated.

Note, that the idea of generating multiple samples from the unknown variables and averaging the responses using those samples confirms the intuition formulated in Section 2.1 and is known as multiple imputation in statistical approaches to regression and classification with missing data (Little and Rubin, 1987).

Note, that the samples can also be used to estimate variances and covariances from which error bars can easily be derived. As examples, if $\{y_t^s\}_{s=1}^S$ are samples generated from y_t, the standard deviation of y_t can be estimated as

$$stdev(y_t) \approx \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(y_t^s - \hat{y}_t)^2}$$

and the standard deviation of the estimated $\hat{y}_t = 1/S\sum_{s=1}^S y_t^s$ can be estimated as

$$stdev(\hat{y}_t) \approx \sqrt{\frac{1}{S(S-1)}\sum_{s=1}^{S}(y_t^s - \hat{y}_t)^2}.$$

4.2 Maximum-Likelihood Substitution

The approach consists of substituting the most likely values

$$Y_{t-1,1}^{ml} = \arg\max_{Y_{t-1,1}^u} P(Y_{t-1,1})$$

for the missing variables. Then, we estimate

$$\hat{y}_t = f(Y_{t-1,t-N}^{ml}, Y_{t-1,t-N}^{m}). \tag{8}$$

Considering, as example, the case with one missing variable y_{t-k} and assuming Gaussian noise

$$y_{t-k}^{ml} = \arg\min_{y_{t-k}} \sum_{l=t-k}^{t-1}(y_l - f(y_{l-1}, y_{l-2}, \ldots, y_{l-N}))^2 \tag{9}$$

we simply find the substitution which minimizes the sum of the squared errors. As another interesting case consider noisy measurements and Gaussian noise distributions

$$Y^{ml}_{t-1,1} = \arg\min_{Y^u_{t-1,1}} [-\log P(Y_{N,1}) + \frac{1}{2\sigma^2_\epsilon} \sum_{l=N+1}^{t-1} (y_l - f(y_{l-1}, y_{l-2}, \ldots, y_{l-N}))^2 + \frac{1}{2\sigma^2_\delta} \sum_{l=1}^{t-1} (y_l - z_l)^2]$$

where σ^2_ϵ and σ^2_δ are the variances of the two noise sources (Section 3). Note, that this is a multidimensional optimization problem. Also note that, for highly nonlinear systems, Equation 8 can be a crude estimate of the expected value and the prediction based on a maximum likelihood estimate of the unknowns can therefore be highly biased.

4.3 Solutions Based on Iterative Density Estimation and the Extended Kalman Filters

We consider the case of prediction with noisy measurements. Note that a solution based on stochastic simulation of Equation 7 (noisy measurements) means that we have to sample from the space of all unknown variables $y_1, \ldots, y_t$. This becomes intractable for large t. To summarize the information about past measurements more efficiently, we can use that

$$P(Y_{t-1,t-N}|Z_{t-1,1}) = \tag{10}$$

$$\frac{P(z_{t-1}|y_{t-1}) \int P(Y_{t-2,t-N-1}|Z_{t-2,1}) P(y_{t-1}|Y_{t-2,t-N-1}) dy_{t-N-1}}{\int P(z_{t-1}|y_{t-1}) P(Y_{t-2,t-N-1}|Z_{t-2,1}) P(y_{t-1}|Y_{t-2,t-N-1}) dY_{t-1,t-N-1}}.$$

This equation can be derived from the Chapman-Kolmogorov equation and by applying Bayes' rule (Lewis, 1986). The update equation implies that we can summarize all information provided by the past measurements by approximating $P(Y_{t-1,t-N}|Z_{t-1,1})$ and use Equation 10 to update the estimates on-line as time progresses and more measurements become available.

If the system is linear and the noise is normally distributed, Equation 10 can be solved analytically and the probability densities can be represented by a multi-dimensional normal distribution. This is the well-known Kalman filter.

In general the integral in Equation 10 must be solved numerically and an appropriate representation for the conditional density has to be found. Neural network techniques for approximating joint and conditional densities exist (Neuneier *et al.*, 1994, Bishop, 1994).

In Lewis (1986) it is shown that for continuous time systems the time update leads to the Fokker-Planck equation which can only be solved in a few simple cases. The problem can be simplified by only requiring to find the iterative estimates of the mean and the covariance. Unfortunately, this approach leads to computationally intractable solutions (Lewis, 1986). The update equations become tractable by using a first order series expansion of the nonlinearities (Lewis, 1986, Bar-Shalom and Li, 1993) which leads to the *extended* Kalman filter. The extended Kalman filter can be used for both discrete

and continuous time systems and summarizes past data by an estimate of the mean and the covariance of the variables involved and is suboptimal in the sense that even with a perfect model, due to the linearization of the system, it does not provide optimal predictions (Lewis, 1986, Bar-Shalom and Li, 1993). The Kalman filter is an iterative algorithm and has the great advantage that it can be used on-line. The Kalman filter has been used for training neural networks and for neural control (Singhal and Wu, 1989, Kadirkamanathan and Niranjan, 1991, Puskorius and Feldkamp, 1994).

5 Experiments

5.1 K-step Prediction

K-step prediction can be considered a special case of prediction with missing data: y_t must be predicted with $y_{t-1}, \ldots, y_{t-K+1}$ missing. In this case, stochastic simulation is very simple: generate a sample y^s_{t-K+1} of the first missing value using the distribution $P(y_{t-K+1}|y_{t-K}, \ldots, y_{t-K-N+1})$. Using that sample and the previous measurements, generate a sample of y_{t-K+2} following $P(y_{t-K+2}|y^s_{t-K+1}, \ldots, y_{t-K-N+2})$ and so on until a sample of each unknown is produced. Repeat this procedure S times and approximate

$$E(y_t|Y_{t-K,1}) \approx \frac{1}{S}\sum_{s=1}^{S} f(y^s_{t-1}, y^s_{t-2}, \ldots, y^s_{t-N})$$

where we have assumed that $K > N$. If $K \leq N$ substitute measured values for y_{t-k} for $k \geq K$. Note, that in this procedure samples are simply generated by simulating the system including the noise model.

5.1.1 Logistic Map

In the first experiment, we used the noisy logistic map $y_t = 4q_{t-1}(1 - q_{t-1}) + \epsilon_t$ with $0 \leq q_{t-1} < 1$ and where

$$q_t = \begin{cases} y_t & \text{if } 0 \leq y_t < 1 \\ y_t - 1 & \text{if } y_t \geq 1 \\ y_t + 1 & \text{if } y_t < 0 \end{cases}$$

where ϵ_t is uncorrelated Gaussian noise with a variance of $\sigma^2 = 0.01$.[3]

Figure 3 (left) shows a realization of the time series and the predictions based on stochastic simulation and a simple iteration of the map. Figure 3 (right) shows the mean squared error as a function of K averaged over 2000 realizations. Shown are the iterated

[3]Note, that here and in the following experiments q_t is only introduced for notational convenience to differentiate the cases when additive noise results in a value of the time series for which the iteration is not defined. q_t is therefore not a "real" hidden variable.

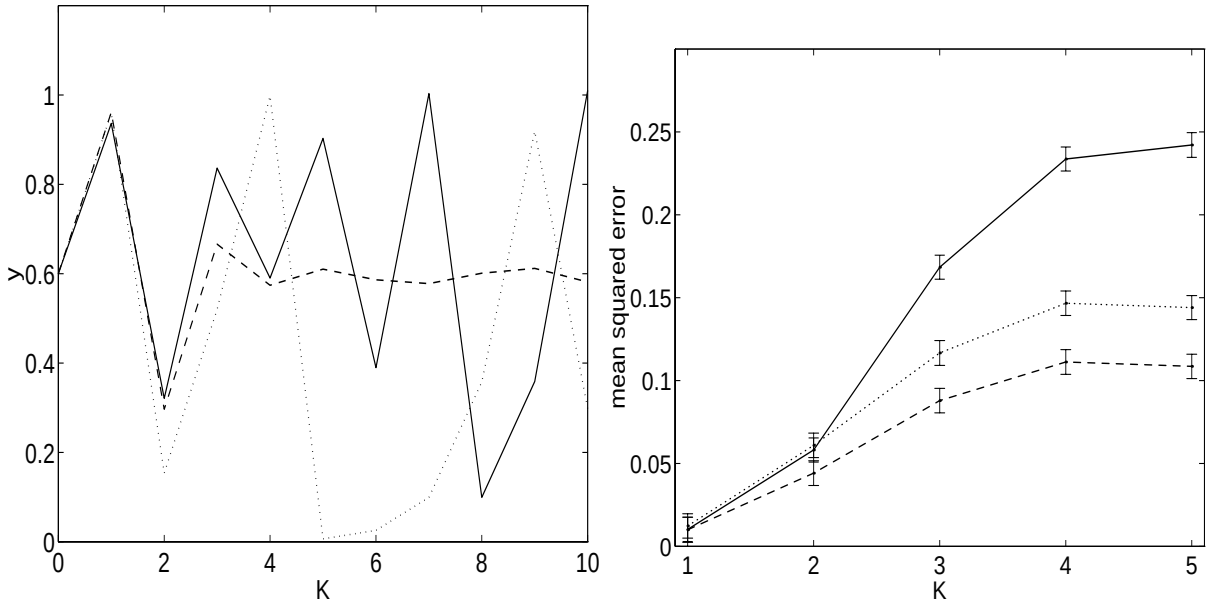

Figure 3: Left: The noisy logistic map (continuous), the K-step prediction using stochastic simulation (dashed) and the K-step prediction by simply iterating the logistic map (dotted). The prediction based on stochastic simulation converges for large K towards the mean of the time-series (which is the optimal solution, since chaotic time series become quickly unpredictable for large K.). Right: The mean squared error as a function of K in $K-$step prediction. The iterated solution (continuous) and the approximation based on stochastic simulation with 3 (dotted) and 20 samples (dashed) are shown. Note, that for $K = 1$ (one-step prediction) the iterated system gives the optimal prediction. For $K > 1$ the accuracy of the prediction of the iterated solution quickly deteriorates. The error bars ($\pm$ one standard deviation) are derived from 2000 independent runs.

system (continuous line) and solutions following the stochastic sampling approach (dotted and dashed). As expected, for $K = 1$ the iterated solution is optimal, but for $K > 1$, stochastic simulation even with only few samples is far superior. This indicates that for highly nonlinear stochastic time series simply iterating the model K-times as it is usually done in $K-$step prediction is suboptimal if $K > 1$. Note, that the $K-$step prediction of the extended Kalman filter, which is based on a local linearization of the nonlinearities, is identical to the iterated system (and therefore is suboptimal as well).

5.1.2 Sun-Spot Data

The second experiment uses the sun-spot data which are records of yearly sun-spot activities from the year 1700 to 1979. First, a multi-layer perceptron was trained to predict

the sun-spot activity based on the 12 previous years of sun-spot activity. The neural network had 12 inputs and one hidden layer with 8 hidden units. Following other authors we trained on data from 1700-1920. We used a weight decay parameter of 0.2.[4]

After training, the mean squared error on the training set is 51.6, on test set number one (data from 1921 to 1955) the mean squared error is 161.5 and on test set number two (data from 1956 to 1979) is 682.0. We assumed normally distributed additive noise with a variance equal to the average error on the whole data set $\sigma^2 = 124$. Figure 4 shows the sun-spot data (dots) from $T = 1738$ to $T = 1987$. In the experiment we perform K-step prediction starting from $T = 1738$ (i.e. $T = 1738$ corresponds to one-step prediction and $T = 1987$ corresponds to 250-step prediction). The top part of the figure displays the prediction of the iterated system and the second plot shows the prediction by stochastic simulation using 1000 samples. The third plot shows one simulated run (including the noise model). Note, that since the latter includes the simulated noise it is more noisy than the iterated system but note also, that in "character" the more noisy time-series is more similar to the true time-series (dots). Unlike the prediction based on the iterated system the prediction based on stochastic simulation converges towards a constant for large K and gives the correct estimate in predicting the mean if K is large.

Figure 5 shows the mean squared prediction error as a function of K. We see that for $K >> 1$ stochastic simulation is clearly superior. Recall, that for $K = 1$ the iterated prediction is optimal.

5.2 Prediction with Missing Data

In this experiment we used the Henon map[5] $y_t = 1 - aq_{t-1}^2 + bq_{t-2} + \epsilon_t$ with $a = 1.4, b = 0.3$ and where

$$q_t = \begin{cases} y_t & \text{if } -1.26 \leq y_t < 1.26 \\ y_t - 1.26 & \text{if } y_t \geq 1.26 \\ y_t + 1.26 & \text{if } y_t < 1.26 \end{cases}$$

and where ϵ_t is uncorrelated Gaussian noise with a variance of $\sigma^2 = 0.1$. The goal is to predict y_t with different patterns of $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}$ missing and y_{t-5}, y_{t-6} known. We used stochastic simulation (here, Gibbs sampling) of Equation 5 for prediction. Figure 6 shows the results.

Apparent is the considerable reduction in error for the solution based on stochastic simulation compared to the heuristic solution.

[4]Readers unfamiliar with weight decay or the multi-layer perceptron, please consult Bishop (1994).

[5]A variation of this experiment was already presented by Tresp and Hofmann (1995).

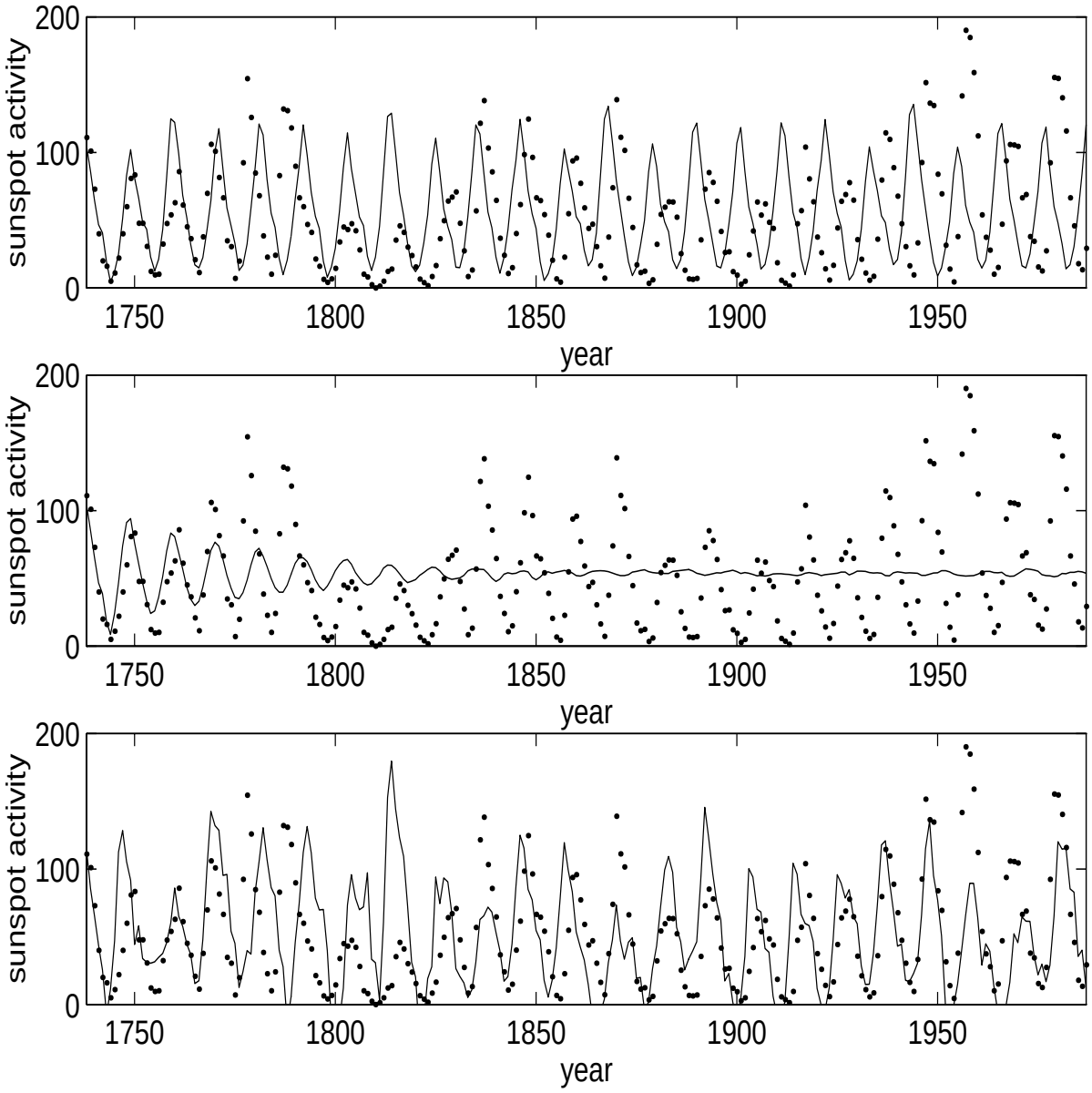

Figure 4: Shown are the sun-spot data from $T = 1738$ to $T = 1987$ (dots). The continuous lines show the K-step predicted value (K increasing with T) based on three different methods. The plot on top shows the iterated system, the plot in the middle shows the prediction based on stochastic simulation using $S = 1000$ samples and the plot on bottom shows one run of the stochastic simulation.

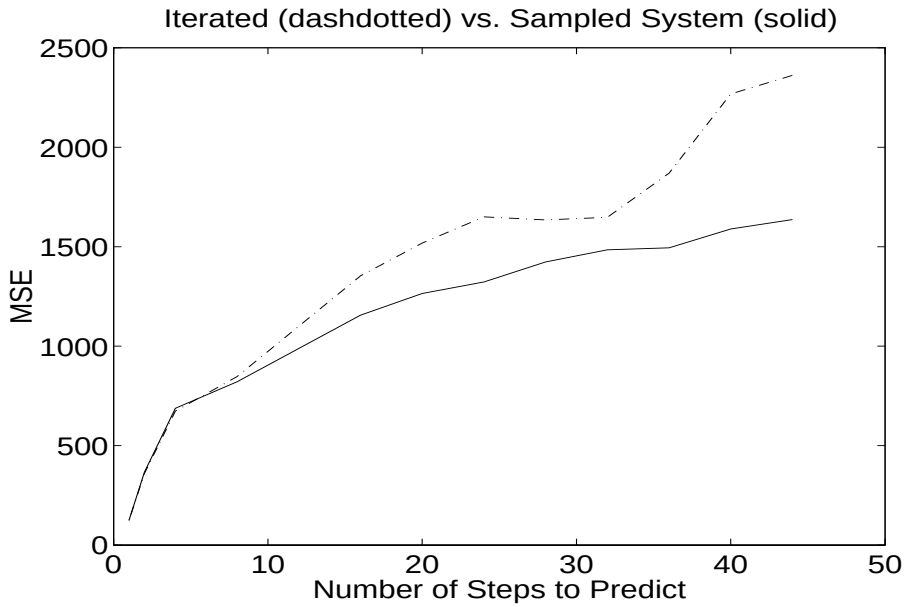

Figure 5: Shown is the mean squared error for K-step prediction for the iterated system (dash-dotted) and the prediction based on stochastic simulation (continuous) for the sunspot data. It is apparent, that for $K >> 1$, the prediction based on stochastic simulation is superior. Shown are averages over all possible experiments where in each experiment the prediction was started from a different point in time. For 1-step prediction we used 250 different starting times possible which means we averaged over 250 experiments and for 50-step prediction, we used 200 possible starting times and consequently, we could average over 200 experiments.

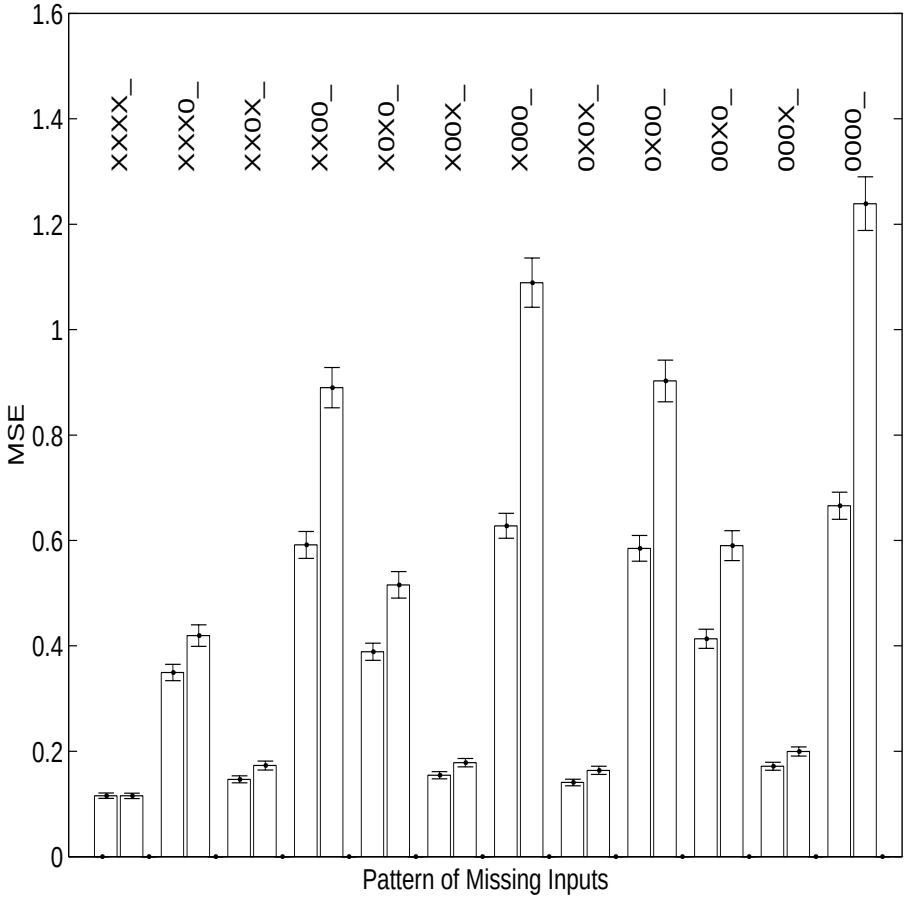

Figure 6: Time series prediction with missing data. The patterns of the missing data are indicated using "X" for known, "0" for unknown values, and "_" for the value to be predicted. As example, $XXOO\_$ indicates that y_{t-4} and y_{t-3} are known and that y_{t-1} and y_{t-2} are missing. y_{t-5} and y_{t-6} are always known. The goal is to predict y_t using either stochastic sampling (left bars) or a heuristic where predicted values are substituted for the missing data (right bars). The height of the bars indicates the squared prediction error averaged over 1000 experiments. The error bars show $\pm$ their standard deviation. For stochastic sampling, we used 200 samples for each prediction. It can be seen that except for one-step prediction ($XXXX\_$) the stochastic sampling solution is significantly better than the heuristic.

6 Conclusions

We have shown how the problem of missing and noisy data can be approached in a principled way in time-series prediction. By unfolding the time-series in time we could apply ideas and methods from the theory of Bayesian networks. We proposed approximations based on stochastic simulations. Experimental results using the logistic map, the Henon map and the sun-spot data confirmed that stochastic sampling leads to excellent predictions which are clearly superior to simple heuristical approaches. The main drawback of stochastic sampling is that generated samples are often highly correlated and a large number of samples might be required to obtain good approximations. For the problem of noisy measurements, the solution would require to generate samples from the joint probability space of all past realizations of the time-series which is clearly unfeasible. In practice, one would only sample from realizations of the time series up to a reasonable chosen time window into the past, which —as a draw back— would lead to suboptimal solutions even with a large number of samples. In this paper we focussed on univariate time-series prediction. The results can easily be extended to multi-variate times series (see the appendix).

Appendix

Multivariate Nonlinear Time-Series

The results can easily be generalized to general nonlinear multivariate models. It is convenient to switch to a state space representation where now $y_t \in \Re^{D_y}$ is a D_y-dimensional state space vector containing all relevant states of all time-series involved. Typically, y_t will be the present and past realizations of all time-series involved, up to a time window in the past. The nonlinear states space model is

$$y_t = f(y_{t-1}) + \epsilon_t$$

where ϵ_t is a D_y-dimensional vector of possibly correlated noise and with probability density P_ϵ. We assume that we have access to a D_z-dimensional measurement vector $z_t \in \Re^{D_z}$ with

$$z_t = g(y_t) + \delta_t$$

where δ_t is a D_z-dimensional vector of possibly correlated noise and with probability density P_δ. Recall from the discussion in Section 3 that the problem of missing data can be considered a special case of noisy data. The joint density of the time-series up to time t (not including z_t, since we consider predictions) is

$$P(Y_{t,1}, Z_{t-1,1}) = P(y_1) \prod_{l=2}^{t} P_\epsilon(y_l - f(y_{l-1})) \prod_{l=1}^{t-1} P_\delta(z_l - g(y_l)). \tag{11}$$

with $Z_{t-1,1} = \{z_1 \ldots z_{t-1}\}$ and $Y_{t,1} = \{y_1 \ldots y_t\}$. Now

$$E(y_t|Z_{t-1,1}) = \int f(y_{t-1})\ P(Y_{t-1,1}|Z_{t-1,1})\ dY_{t-1,1}$$

where $P(Y_{t-1,1}|Z_{t-1,1}) = P(Y_{t-1,1}, Z_{t-1,1})/P(Z_{t-1,1})$ is obtained from Equation 11.

References

[1] Ahmad, S. and Tresp, V. (1993). Some Solutions to the Missing Feature Problem in Vision. In S. J. Hanson, J. D. Cowan and C. L. Giles, (Eds.), *Neural Information Processing Systems 5.* San Mateo, CA: Morgan Kaufmann, pp. 393-440.

[2] Bar-Shalom, Y. and Li, X.-R. (1993). *Estimation and Tracking.* Artech House, Boston.

[3] Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory.* Wiley & Sons.

[4] Bishop, C. M. (1994). *Neural Networks for Pattern Recognition.* Oxford University Press.

[5] Buntine, W. L. and Weigend, A. S. (1991). Bayesian Back-Propagation. *Complex systems,* Vol. 5, pp. 605-643.

[6] Jensen, F. V. (1996). *An Introduction to Bayesian Networks.* Springer, New York.

[7] Kadirkamanathan, V. and Niranjan, M. (1991). Nonlinear Adaptive Filtering in Nonstationary Environments. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. 2177-2180.

[8] Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME J. of Basic Eng.,* 8, pp. 35-45.

[9] Lewis, F. L. (1986). *Optimal Estimation with an Introduction to Stochastic Control Theory.* John Wiley, New York.

[10] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

[11] Neal, R. M. (1993). *Probabilistic Inference using Markov Chain Monte Carlo Methods.* Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

[12] Neuneier, R., Hergert, F., Finnoff, W. and Ormoneit, D. (1994). Estimation of Conditional Densities: A Comparison of Neural Network Approaches. *Proc. of ICANN 94,* Sorrento, pp. 689-692.

[13] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, San Mateo. 1988.

[14] Puskorius, G. V. and Feldkamp, L. A (1994). Neurocontrol of Nonlinear Dynamical Systems with Kalman Filter Trained Recurrent Networks. *IEEE Transactions on Neural Networks,* 5:2, pp. 279-297.

[15] Singhal, S. and Wu, L. (1989). Training Multi-layer Perceptrons with the Extended Kalman Algorithm. In: Touretzky, D. S., ed., *Advances in Neural Information Processing Systems 1,* Morgan Kaufman, pp. 133-140.

[16] Shumway, R. H. and Stoffer, D. S. (1982). An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm. *Journal of Time Series Analysis 3,* pp. 253-264.

[17] Tresp, V. and Hofmann, R. (1995). Missing and Noisy Data in Nonlinear Time-Series Prediction. in F. Girosi, J. Makhoul, E. Manolakos und E. Wilson (Hrsg.), *Neural Networks for Signal Processing 5,* IEEE Signal Processing Society, New York, NY, IEEE catalog number: 95TH8094, pp. 1-10.

[18] Tresp, V. and Hofmann, R. (1997). Missing and Noisy Data in Nonlinear Time-Series Modeling. Manuscript in preparation.

[19] Weigend, A. S. and Gershenfeld, N., eds., (1994). *Time-Series Prediction.* Addison-Wesley, 1994.