PREDICTION-BASED CHARACTERIZATION OF CHAOTIC PHENOMENA IN BIO-SCIENCES

TOHRU OZAKI
The Institute of Statistical Mathematics, Tokyo

PEDRO VALDES-SOSA
Cuban Neuroscience Center, Havana

VALERIE HAGGAN-OZAKI
Sophia University, Tokyo

ZHAOYUN SHI
The Institute of Statistical Mathematics, Tokyo

1. INTRODUCTION

From time immemorial, the problem of predicting the future has captured the imagination of statisticians, mathematicians, engineers and scientists from many disciplines, not to mention humankind in general. The scope and depth of applications of prediction theory are boundless. Leaving aside the crystal ball approach, any prediction method based on scientific principles first needs some kind of model which closely describes and to some extent explains the observed phenomena. When we want to describe the time evolution of dynamic phenomena for prediction purposes, we could decide to use either of two distinct types of dynamic model, i.e. either a deterministic model, or a stochastic model.

In the last decades, there has been a great deal of interesting argument between two schools, on the one side "believers" in the deterministic model and on the other "believers" in the non-deterministic model. Recent trends in prediction studies seem
to be leaning in favour of the deterministic side. Ruelle\textsuperscript{49} says "There is evidence that the time evolution of many physical systems can be described by an equation of the form \[ \dot{Z}_t = \phi_t(Z_t), \quad Z_t \in \mathbb{R}^l \] in a space of small dimension $l$. This is typical of how determinists see the prediction problem. On the other side are non-determinists, such as Akaike, and, in the past, Wiener, who have used statistical argument to try to highlight and deal with uncertainties and unavoidable noise such as observation error and microscopically small external forces stimulating the system.

In the inferential stage, the distinction between the two schools is not as clear as some scientists who believe in determinism might think. There are two ways of thinking about stochastic models, and the way in which they are used. In most inferential circumstances, we assume that the true model is stochastic but unknown, and the observed data is used to seek for the most likely model. The stochastic model which is most likely to have generated the present data is chosen as the "best" model. In many situations, this is equivalent to the model found by the maximum likelihood method, and hence is also equivalent, asymptotically, to the model estimated by least squares. Another way of looking at stochastic models is to think of them as approximations being used because of analysts' imperfect knowledge and ignorance. When fitting such models, it is natural to try make the model-fitting error as small as possible, which leads to a model chosen by the least squares method.

Some chaos scientists approve of the second method of using stochastic models, and denounce the first method\textsuperscript{18,54}. The difference between the two situations, however, although it may be clear in a philosophical sense, is not clear in a mathematical sense. All analysis, whether of the "deterministic" or "non-deterministic" school, have the same aim of trying to reduce the uncertainty of the future as much as possible through the use of a model. The only apparent difference lies in whether or not the analyst believes that the phenomena underlying the observed data arise deterministically. However, since many analysts have ended up using a stochastic model, despite their inherent belief in determinism, the lessons that statisticians have learned over the years about stochastic modelling could still be useful even for them.

The essential point is to propose a prediction model which will be useful for many kinds of prediction problem. Recently, prediction models have been introduced in several research areas, including neural network studies, function approximation, nonlinear time series analysis etc. Interestingly, if we put these prediction models together and compare them within the same framework, suppressing and masking the philosophical ideas behind these models, many similar characteristics may be observed, yielding many fresh, new ideas for the further development of nonlinear prediction models.

In the present paper, we will look at several types of nonlinear prediction model, being developed in different fields, and, in section (2), discuss the similarities and the differences when they are used in prediction. In section (3) we consider a prediction problem for typical nonlinear and non-Gaussian time series data, i.e. the prediction of EEG data in subjects suffering from epilepsy, and discuss the potential and limitations of currently available nonlinear prediction models. In section (4), the stochastic theories behind the different kinds of predictors will be discussed, together with new techniques for model identification, and various ways in which models might be generalized in future will be suggested.

2. RECONSTRUCTING THE DYNAMICS

First, we must note some of the many methods which have been presented for the reconstruction of the model dynamics from observation data, is of course very important for predicting dynamic phenomena. Although, on the face of it, these methods look quite different, they do in fact have many similarities, and may in some ways be thought of as equivalent.
2.1. Nonlinear AR Models

In nonlinear AR (Auto-Regressive) time series modelling, analysts are interested in predicting the future values of a time series \( z_t \) from a set of past observations \( z_{t-1}, z_{t-2}, \ldots \) by using a nonlinear function of the past observations up to a finite time-lag. Here, it is unlikely that the prediction errors will be zero because the true model is unknown. In the inferential stage, analysts are always exposed to the effects of uncertainty and noise, so that exact prediction is impossible. A prediction error \( e_t \) is always included in the model, which is of the form

\[
    z_t = \phi (z_{t-1}, z_{t-2}, \ldots, z_{t-d}) + e_t, \quad \ldots \ (1)
\]

where \( \phi (\cdot) \) is a nonlinear autoregressive function and \( e_t \) is usually regarded as a Gaussian white noise with variance, \( \sigma^2 \). The aim of statistical modelling is to try to find those values of \( \phi (\cdot) \) and \( \sigma^2 \) which maximize the log-likelihood of the model. Since \( e_t \) is assumed to be Gaussian white noise, this is asymptotically equivalent to choosing \( \phi (\cdot) \) which minimizes the sum of squares of the errors \( e_t \), i.e. minimizes

\[
    \sum_{t=d+1}^{N} [z_t - \phi (z_{t-1}, z_{t-2}, \ldots, z_{t-d})]^2
\]

Finding an appropriate function for \( \phi (\cdot) \) out of huge number of possible candidates is very important but obviously extremely difficult. Experience gained in earlier works by the present authors\(^{21,22,34,41} \) showed that representing (1) by a state-dependent AR model of the following form,

\[
    z_t = \phi_1 (z_{t-1}, z_{t-2}, \ldots, z_{t-d}) z_{t-1} + \ldots
\]

\[
    + \phi_d (z_{t-1}, z_{t-2}, \ldots, z_{t-d}) z_{t-d} + e_t
\]

and parameterizing the dynamics of its coefficients, \( \phi_1 (z_{t-1}, \ldots, z_{t-2}, \ldots, z_{t-d}) \), \( \phi_d (z_{t-1}, z_{t-2}, \ldots, z_{t-d}) \) by some simple smooth functions such as \( \exp (-\gamma z_{t-1}^2) \), provided a steady step forward towards the goal of general parametric nonlinear AR models.

2.2. Multi-Layer Neural Network Models

The reconstruction of the dynamic map, or model dynamics, is also one of the aims of neural network studies, where the nonlinear function \( f_d (x_{t-1}, \ldots, x_{t-d}) \) representing the dynamic map is approximated by a special type of nonlinear function constructed from the time series sample \( z_{1T}, \ldots, z_N \). In fact, the multi-layer neural network model and in particular the three-layer neural network model, seems to be the most popular and widely used neural network model for the identification and prediction of the future, since there is a theory which shows that it is sufficient to use a maximum of three layers to solve an arbitrarily complex pattern classification problem\(^{50} \). For illustration, the basic architecture of a two-layer model is shown in Fig. 1.

![Fig. 1. The architecture of a two-layer model.](image)

In the first layer of this figure, we can see that the \( h_1 \) hidden units (processing elements) perform a weighted summation of the inputs which are then transformed by nonlinear "activation" functions, \( \{ \psi_{ji}^{(1)} \}_{i=1,2,\ldots,h_1} \), for example sigmoid (S-shaped) functions of the form
\[ \psi_{ji}(x) = \tanh(-\gamma x) = \frac{1 - e^{-2\gamma x}}{1 + e^{-2\gamma x}} \]

where \( \gamma \) is a parameter controlling the "steepness" (slope) of the function. The relationship between inputs and outputs in the first layer may be expressed by

\[ O_{ji} = \psi_{ji} \left[ \sum_{i=1}^{n_0} w_{ji} x_{i-1} + w_{j0} \right], j = 1, 2, ..., h_1 \]

where the synaptic weights \( w_{ji} \) are adjustable and can take positive, negative or zero values. The same kind of procedure is repeated in second and later layers, using the output of the previous layer as inputs.

We could think of using this kind of model with samples of time series sequences \( z_{t-1}, ..., z_{t-d} \) \((t = N, N-1, ..., d+1)\) as inputs, and one output \( \hat{z}_{t+1} \) at the final layer, where \( \hat{z}_{t+1} \) is the one-step-ahead prediction of \( z_t \) from the time point \( t-1 \), so that \( \hat{z}_{t+1} \) could be written as \( \psi(z_{t-1}, ..., z_{t-d}) \) for some nonlinear function \( \psi \). The prediction of the time series \( z_t \) by \( \psi(z_{t-1}, ..., z_{t-d}) \) is not exact, of course, for the same reasons as given for the nonlinear AR model case. The prediction error is defined as the difference between the actual value \( z_t \) and the value \( \hat{z}_{t+1} = \psi(z_{t-1}, ..., z_{t-d}) \) predicted by the neural network model, i.e.

\[ e_t = z_t - \hat{z}_{t+1} = z_t - \psi(z_{t-1}, ..., z_{t-d}) \]

We naturally expect that the error \( e_t \) is unbiased, i.e. that

\[ E[e_t] = 0 \]

and we try to find the best model by minimizing the sum of squares of the errors, i.e.

\[ \sum_{i=d+1}^{N} [z_i - \psi(z_{i-1}, ..., z_{i-d})]^2 \]

which is asymptotically equivalent to maximizing the log-likelihood of the model

\[ z_t = \psi(z_{t-1}, ..., z_{t-d}) + \epsilon_t \]

if \( \epsilon_t \) is a Gaussian white noise. Although most neural network researchers believe that the model is a deterministic approximation to the dynamics of a deterministic time series, what they are actually doing in reconstructing and approximating the dynamics is mathematically equivalent to the method described in section (2.1), i.e. the statistical identification of stochastic nonlinear time series models from observed empirical data.

![Fig. 2. Examples of nonlinear activation function.](image-url)
The special feature of the neural network model is its method of identifying the nonlinear function \( \psi \). Here the functional form of each layer is fixed and weights \( w_p \) are estimated for each layer using the sample data. There are several algorithms for the estimation of these coefficients. However, the computational speed of these estimation (or learning) procedures is extremely slow compared with ordinary least squares estimation of nonlinear models parameterized as linear functions of nonlinear functions, such as polynomial AR models. Although we fix the nonlinear functional form for each layer, the choice is rather arbitrary. There are many possible forms for the activation function (some examples are given in [13] and in Fig. 2), and the prediction results will depend on the initial choice of the functional form.

Although the multi-layer neural network models provide us with wide range of tractable nonlinear prediction models, the computational burden of estimating the weights of each layer in the model generally forces researches to move from the general multi-layer neural network model to a family of single layer networks with a general nonlinear function family. The shift is commonly seen in many scientific fields where scientists become more and more interested in a specific dynamic structure for their own problem, as their vision becomes clearer in the light of preliminary analysis with a general model such as a multi-layer neural network model.

2.3. Single-Layer RBF Neural Network Model

An alternative model to the multi-layer neural network model for the reconstruction of the dynamics and for prediction is provided by the single-layer neural network model employing RBFs (Radial Basis Functions). An RBF is a multidimensional function which depends on the distance \( r = \| Z - X \| \) (where \( \| \cdot \| \) denotes a vector norm) between a \( d \)-dimensional input vector \( Z \) and a “centre” \( X \). One of the simplest approaches to the approximation of a nonlinear function is to represent it by a linear combination of fixed nonlinear basis functions \( w_i(Z) \), i.e.

\[
 f(Z) = \sum_{i=1}^{m} c_i w_i(Z).
\]

Typical choices for radial basis functions

\( w(Z) = W(\| Z - X \|) \) are

i) \( W(r) = r : \) piecewise linear approximation,

ii) \( W(r) = r^3 : \) cubic approximation,

iii) \( W(r) = \exp(-r^2/h^2) : \) Gaussian function,

iv) \( W(r) = r^2 \log(r) : \) thin-plate-spline function,

v) \( W(r) = (r^2 + h^2)^{1/2} : \) multiquadratic function and

vi) \( W(r) = (r^2 + h^2)^{-1/2} : \) inverse multiquadratic function,

where \( h \) is a real coefficient called the width or scaling parameter. Among the functions described above the most popular and widely used is the Gaussian function which has a peak at the centre \( X \) and decreases monotonically as the distance from the centre increases.

In general, an RBF network is specified by three sets of parameters: the centres \( X_i \in \mathbb{R}^d \), the width or distance scaling parameters \( h_i \), and the synaptic weights \( c_i \in \mathbb{R} (i = 1, 2, ..., m) \). The performance of an RBF network depends critically on the chosen centres \( X_i \in \mathbb{R}^d (i = 1, 2, ..., m) \). Several heuristic methods for determining the centres have been proposed. Once the cluster centres \( X_i \) have been determined, the distance scaling parameters \( h_i \) are determined from the \( p \) nearest neighbors heuristically by

\[
 h_i = \frac{1}{p} \sum_{j=1}^{p} \left( \| X_i - X_j \|^2 \right)^{1/2}
\]

where \( X_j \) are the \( p \) nearest neighbours of \( X_i \). Having obtained the centres \( X_i \) and the distance scaling parameters \( h_i \), we are able to
adjust the synaptic weights \( c_i \) by minimizing the sum of the squares of prediction errors

\[
e_t = z_t - \left\{ c_0 + \sum_{i=1}^{m} c_i W_i \left( \| Z_{t-1} - X_t \| \right) \right\},
\]

i.e. by minimizing

\[
\sum_{t=d+1}^{N} \sum_{t=d+1}^{N} 2 e_t^2 = \sum_{t=d+1}^{N} \left[ z_t - \left\{ c_0 + \sum_{i=1}^{m} c_i W_i \left( \| Z_{t-1} - X_t \| \right) \right\} \right]^2,
\]

which is equivalent to least squares estimation of the parameters of the statistical linear regression

\[
z_t = c_0 + \sum_{i=1}^{m} c_i W_i \left( \| Z_{t-1} - X_t \| \right) + \epsilon_t,
\]

where \( Z_{t-1} = (z_{t-1}, z_{t-2}, ..., z_{t-d})^T \). The calculation of the coefficients \( \{c_i\} \) may be performed very efficiently, in contrast to the slow back-propagation algorithm used for the multi-layer neural network models. This shows that if the prediction model is formulated on general but efficient basis function regressors, the reconstruction of the predictor can be performed much more efficiently. Under the assumption that the error \( \epsilon_t \) is Gaussian white noise, the above least squares estimates \( c_i (i=1, ..., m) \) are asymptotically equivalent to the estimates obtained by the maximum likelihood method.

If the approximation is to remain valid for points of the embedded trajectory not used in estimating the model, that is for predicting the future of the dynamics, the model should have well-behaved generalization capabilities. This requires that the reconstructed, map \( f(\cdot) \) should be smooth and have good interpolation properties for scattered data points. \( f(\cdot) \) may be written in the form

\[
f(z_{t-1}, ..., z_{t-d}) = c_0 + \sum_{i=1}^{m} c_i w_i (z_{t-1}, ..., z_{t-d})
\]

with appropriate functional forms for the multivariate localized basis functions \( w_i (z_{t-1}, ..., z_{t-d}) \) \( i = 1, ..., m \). Since the class of RBFs has been recognized \( 14, 15 \) to be a class of localized basis functions possessing a high degree of smoothness or regularity, and can very easily incorporate multivariate data, this model should provide a good basis for prediction.

RBF expansions have good interpolation properties in dealing with scattered data points, and are endowed with the "universal approximation" and "best approximation" capabilities of any continuous function. Universal approximation implies the possibility of approximating a function to any required degree of accuracy. The stronger property of best approximation entails that the approximation error surface always has a unique global minimum for any approximation performance measure \( 46, 47 \). As a consequence of these approximation capabilities, the RBF model has been recognized as an alternative to the multi-layer neural network model \( 16 \). But the RBF model has a clear computational advantage over the multi-layer neural network models, deriving from its "linear-in-the-parameter" formulation. The one-layer structure of the RBF model is a feature which can be exploited for parameter estimation, and allows a faster learning scheme in comparison with the back-propagation techniques used for the multi-layer neural network models.

2.4. Conditional Distributions \( p(z_t | z_{t-1}, ..., z_{t-d}) \)

By this stage it is probably quite clear, in all the methods described above, that what is actually being done when we reconstruct the dynamics of
\[ x_t = f_\theta (x_{t-1}, \ldots, x_{t-d}) \]  \hspace{1cm} \text{(2)}

from the observation data \( z_t = (x_t + \epsilon_t) \) is equivalent to approximating the conditional expectation of \( z_t \),

\[ E [z_t \mid z_{t-1}, \ldots, z_{t-d}] \]

under the condition that the data are given up to the previous time point \( t-1 \), i.e. \( z_{t-1}, \ldots, z_{t-d} \) are known. In all the three models discussed, the conditional distribution of \( z_t \) \( p(z_t \mid z_{t-1}, \ldots, z_{t-d}) \) is defined by a Gaussian distribution whose mean is given by

i) \( \phi(z_{t-1}, z_{t-2}, \ldots, z_{t-d}) \) for the nonlinear AR model case,

ii) \( \psi(z_{t-1}, z_{t-2}, \ldots, z_{t-d}) \) for the multi-layer neural network model case and

iii) \( c_0 + \sum_{i=1}^{m} c_i w_i (z_{t-1}, \ldots, z_{t-d}) \)

for the single-layer RBF model case.

In each case the predictor of \( z_t \) is given by the mean of the conditional distribution, and the variance of the prediction error is given by the variance of the conditional distribution. This implies that the above three models, although appearing to be different, all belong to the same family of models, i.e. locally Gaussian models which may have non-Gaussian distributions globally, but have a Gaussian distribution locally in time, where the variance of the local Gaussian distribution is constant over time, and the mean is a function of \( z_{t-1}, \ldots, z_{t-d} \).

By differentiating between the observed data \( z_t \) and the actual true value of the state \( x_t \) of the dynamics, we will see (in section 4) that it is possible to find a more sophisticated parametric method for the specification of the locally Gaussian conditional distribution \( p(z_t \mid z_{t-1}, \ldots, z_{t-d}) \). However, before that, in the next section let us first look at another method which has gained popularity in recent years, which uses kernel estimators to find non-parametric estimates of the conditional distribution and of the conditional mean, in particular.

2.5. Non-Parametric AR Models

We know from the theory of multivariate kernel density estimation\(^{11}\) that the conditional mean

\[ E [z_t \mid z_{t-1}, \ldots, z_{t-d}] = \int z_t p(z_t \mid z_{t-1}, \ldots, z_{t-d}) \, dz_t \]

of \( z_t \) may be approximated using a kernel density estimator,

\[ \hat{z}_t = \frac{1}{(N-d)\nu_d} \sum_{i=d+1}^{N} z_i \prod_{j=1}^{d} K \left( \frac{z_{t-j} - z_{i-j}}{h} \right) \]  \hspace{1cm} \text{(3)}

where \( \bar{v}_d \) is an \( N \)-sample kernel estimate of the \( d \)-variate probability density function of \( Z_{t-1} = (z_{t-1}, \ldots, z_{t-d}) \), \( \nu_d \), i.e.

\[ \bar{v}_d = \frac{1}{(N-d)} \sum_{i=d+1}^{N} z_i \prod_{j=1}^{d} K \left( \frac{z_{t-j} - z_{i-j}}{h} \right) \]

and where \( h = h(d; N) \) is a bandwidth satisfying the standard conditions.

Since the most commonly used kernel functions are Gaussian kernel functions, we will usually discuss the present modeling procedure using the Gaussian kernels,

\[ \prod_{j=1}^{d} K \left( \frac{z_{t-j} - z_{i-j}}{h} \right) = \prod_{j=1}^{d} \exp \left\{ - \left( \frac{z_{t-j} - z_{i-j}}{h} \right)^2 \right\} \]

However, the discussion applies to other kernels without any essential change. We may also sometimes use the more compact multivariate functional expression,
where $F_2(Z_{t-1})$ is a $1 \times d$ vector whose elements are functions of the $d$-vector $Z_{t-1}$.

The advantage of the non-parametric method is that it retains maximal flexibility of model choice at the initial stage of data analysis. The attractive asymptotic results related to the method provide analysts with a feeling of psychological security. However, the studies of [56] and [23] indicate that very large sample sizes may be needed to obtain a close correspondence between asymptotic theory and simulations. It is also recognized in analyzing real data that $d$ must be very large for the method to produce sensible estimates, thus requiring massive computation to reach the final result. It will be useful if we could take advantage of the results of the non-parametric analysis for the introduction of a new nonparametric nonlinear models. We will discuss this point again later in section 4.

3. MODELLING THE DYNAMICS OF EEG DATA

3.1. Laying the "Curse of Dimensionality"

When faced with large scale RBF models having many lag orders and many centers, naturally statisticians feel more comfortable to avoid using them, citing the "curse of dimensionality" as a convenient phrase. While many statisticians have been busy building elegant mathematical models of order 1 or 2, or using non-parametric methods for preliminary data analysis, control engineers, bio-scientists, computer scientists etc. have had to try to tackle real-world problems, yielding extremely complicated data, which they have to analyze by any possible means, elegant or not. Neural network models and RBF models, although actually special types of statistical time series models (i.e. locally Gaussian nonlinear autoregressive models), have been the result of such efforts by non-statisticians. The potential of RBF models as a general statistical time series model for the modelling of general nonlinear dynamics should not be overlooked because of "curse of dimensionality". It may be time for statisticians to put more
emphasis on larger scale models for solving real practical problems in many scientific fields such as neuroscience, meteorological and environmental science, world-wide global macro-economics and finance etc.

Statisticians do have a lot of experience to offer in tackling these complicated problems, which often need large scale parametric models. In the late 1960's, Akaike's solution of a large-scale cement rotary kiln control problem with the help of control engineers provided an interesting example for present-day statisticians. Faced with a strong demand for practical efficient control using a parametric model with a large number of state variables and complicated dynamics, Akaike realized the weakness of statistical testing methods for the selection of an appropriate model among so many candidates, and introduced the FPE (Final Prediction Error) criterion, which is asymptotically equivalent to AIC, for selection of the control model. The new control method based on this model identification technique strongly influenced Japanese industries related to control throughout the 1970s and 1980s.

Akaike's work had two by-products which have had profound impact on front-line scientists struggling with the gap between established theory and the new and different kinds of observation data found in many scientific research fields. The two by-products, which have been strongly promoted by Akaike since then, are:

i) AIC, a new paradigm for statistical model identification based on the generalized Boltzmann entropy.

ii) The development of new statistical model families, whether ordered in increasing degree of complexity or not, for solving various new types of inferential problem in collaboration with scientists from different engineering or scientific research fields.

By-product ii) is of course related to the ever-increasing development of the very high-speed computing power of the modern age backed by AIC. Many statisticians, especially conventional theoretical statisticians educated in the Fisherian or Neyman-Pearson paradigm, are still antagonistic towards AIC. However, when the power and capacity of computers has multiplied almost exponentially (at least one of the authors recalling very well that at the time Box and Jenkins famous book was published in 1970, research had to be carried out on a mainframe computer with 64 KB RAM) it is no longer reasonable for analysts to keep on saying, for example, that "fitting a high-order linear ARMA model is a very difficult computational problem." The curse of dimensionality should no longer be a problem, and this has been realized by many applied statisticians and scientists outside the inner circle of statistics, who have wholeheartedly accepted by-product ii) of Akaike's work (for example, see how often the word "AIC" is cited by scientists in academic journals outside statistics every year).
A good benchmark test data for nonlinear time series is provided by modelling EEG data taken during an epileptic seizure. Fig. 3 shows a typical nonlinear time series pattern taken from this kind of data; a very complicated limit cycle pattern, which no conventional parametric nonlinear time series models can reproduce. We would like to see if modern time series techniques can fully characterize this time series.

3.2. Generalized ExpAR Models

First we will see what may be done with the EEG data of Fig. 3 by a parametric approach with nonlinear AR type modelling. The original idea of Ozaki & Oda’s ExpAR modelling\(^{41}\) was to make the characteristic roots of the AR model amplitude-dependent, instead of constant, so that the model can adapt to the temporal instantaneous change of the amplitude-dependent nonlinear rolling characteristics of a container ship. Here the mathematical model for the dynamics of the ship rolling angle \(z(t)\) is given by a continuous time nonlinear differential equation,

\[
\frac{d^2 z}{dt^2} + a \frac{dz}{dt} + bz + \beta z^3 = n,
\]

... (7)

where \(n\) is a random external force driving the ship including the action of sea waves and wind. There are two basic ideas behind ExpAR modelling:

i) **AR-based modelling** : An AR-based model is employed so that the resulting nonlinear AR models may be interpreted on the basis of the corresponding nonlinear differential equation model.

ii) **Instantaneous state-dependent AR coefficients** : The AR coefficients (and hence the roots of the characteristic equation) are made to be instantaneously state-dependent so that the correspondence between the discrete time model and continuous time model is easy to see.

Nonlinear dynamics such as the amplitude-dependent frequency shift phenomena of ship rolling come from the amplitude-dependent nonlinear restoring force in (7). For this reason ExpAR models included only the term \(\exp\left(- \gamma z_{l-1}^2\right)\) in the AR coefficients. Incidentally, some constructive methods, based on the above two ideas, for obtaining a continuous time dynamical system through ExpAR models have been presented, with applications to fairly simple nonlinear problems in [36] and [37].

Recently, many applied time series areas are dealing with more complicated nonlinear dynamics, where it is possible to take advantage of the flexibility of RBF models to cope with more complicated nonlinear dynamics of characteristic roots and develop them in line with the ExpAR modelling procedure. One pioneering step was made by Vesin\(^{59}\) whose RBF based AR model is given by,

\[
z_i = \phi_0(Z_{i-1}, Z_1, Z_2, ..., Z_m) + \sum_{j=1}^{d} \phi_j(Z_{i-1}, Z_1, Z_2, ..., Z_m) z_{i-j} + e_i
\]

where \(\{\phi_k(Z_{i-1}, Z_1, Z_2, ..., Z_m), (k = 0, 1, 2, ..., d)\}\) are RBF functions of the form

\[
\phi_k(Z_{i-1}, Z_1, Z_2, ..., Z_m) = c_{k,0} + \sum_{l=1}^{m} c_{k,l} \prod_{j=1}^{r} \exp\left\{\frac{-(z_{i-j} - z_l)^2}{h_l}\right\}
\]

with properly chosen centres, \(Z_l = (z_{l,1}, z_{l,2}, ..., z_{l,r})'\) and scaling parameters \(h_l\) \((l = 1, 2, ..., m)\). (Actually in [59], \(\phi_0(Z_{i-1}, Z_1, Z_2, ..., Z_m) = 0\). The potential power of the model was confirmed by Shi \textit{et al.}\(^{52}\) for the case of modelling of Mackay-Glass chaos data. For this model, we need to identify the orders \(d, m, r\) from the data as well as the scaling parameters \(h_l\) centers \(Z_l = (z_{l,1}, z_{l,2}, ..., z_{l,r})'\) \((l = 1, 2, ..., m)\) and coefficients \(c_{k,l}\) \((l = 0, 1, 2, ..., m, k = 0, 1, 2, ..., d)\). Although the identification is computationally much faster than that of the general multi-layer network models, a great deal of computation is still required, especially for choosing appropriate centers \(Z_l\) \((l = 1, 2, ..., m)\). Here the key state characterizing the change of characteristic roots is the \(r\)-dimensional vector.
\( z_{t-1} = (z_{t-1}, z_{t-2}, ..., z_{t-r}) \)' instead of the \( z_{t-1} \) used in the ExpAR model.

Vesin\(^{59}\) pointed out, in his introduction of RBF-AR models, that a possible defect of the ExpAR model is that its nonlinear AR coefficients depend only on \( z_{t-1} \) and he suggested making them dependent on past values, \( z_{t-1}, z_{t-2}, ..., z_{t-r} \). He also showed some interesting successful numerical results applied to real data using this idea\(^{59}\). The results seemed to suggest giving up the idea of the original ExpAR model, i.e. looking for the dynamics of a continuous time model through ExpAR modelling. We think, however, that Vesin’s idea could be implemented with regard to the two original ideas of ExpAR model by using generalized state \( Z \), where \( Z = (z, dz/dt, d^2z/dt^2, ..., d^{r-1}z/dt^{r-1}) \)', so that the model can still maintain its instantaneous dependence on the state vector of the dynamics. If the differential operator is replaced by the difference operator so that, for example, \( dz/dt \) is replaced by \( \Delta z_{t-1} = z_{t-1} - z_{t-2} \), the discrete time state at time point \( t-1 \) is \( Z_{t-1} = (z_{t-1}, \Delta z_{t-1}, \Delta^2 z_{t-1}, ..., \Delta^{t-1} z_{t-1}) \)' which is essentially equivalent to \( (z_{t-1}, z_{t-2}, ..., z_{t-r}) \)' for \( k \) values. Therefore, we could have a Vesin-type RBF-AR model,

\[
\begin{align*}
    z_t = & \phi_0 (Z_{t-1}, Z_1, Z_2, ..., Z_m) + \\
    & \sum_{j=1}^{d} \phi_j (Z_{t-1}, Z_1, Z_2, ..., Z) z_{t-j} + e_t 
\end{align*}
\]

but with state vector and centers, \( Z_{t-1} = (z_{t-1}, \Delta z_{t-1}, \Delta^2 z_{t-1}, ..., \Delta^{t-1} z_{t-1}) \)' and \( Z_l = (z_l, \Delta z_l, \Delta^2 z_l, ..., \Delta^{t-1} z_l) \)' (\( l = 1, 2, ..., m \)) respectively. Then the RBF coefficients would be

\[
\phi_k (Z_{t-1}, Z_1, Z_2, ..., Z_m) = c_{k,0} + \sum_{l=1}^{m} c_{k,l} \prod_{j=0}^{r-1} \exp \left\{ -\frac{(\Delta^j z_{t-1} - \Delta^j Z_l)^2}{h_l} \right\},
\]

where \( (k = 0, 1, 2, ..., d) \). Fig. 4 shows the plot of time series \( z_t, \Delta z_t, \Delta^2 z_t, \Delta^3 z_t \) for the epilepsy data from \( t = 30 \) to \( 60 \) where we can see detail of each \( \Delta^k z_t \) \( (k = 0, ..., r) \). Obviously \( \Delta^k z_t \)'s \( (k = 1, 2, 3) \) near the spike (around \( t = 30 \) to \( 60 \)) carry much more information than in other periods. However, a clear fracture is that as \( k \) increases, the corresponding plots become flatter, and are less informative. Ozaki et al.\(^{45}\) used these plots as a basis for fitting an RBF-AR to the EEG data with \( r = 2 \).

![Fig. 4.](image)

Ozaki et al.\(^{45}\) fitted the model to the first 500 data points (which we call training data set) and obtained a model giving reasonably small prediction errors: residual variance is \( \hat{\sigma}^2 = 1.4670 \times 10^3 \), with \( d = 8 \), \( m = 4 \) and \( r = 2 \). Fig. 5 shows the errors after fitting the estimated model. The estimated model was then
applied to the last 140 data points (which we call testing data set) in order to calculate the one-step-ahead prediction errors, which are shown in Fig. 6. The figures show that the prediction performance is good both for the testing data set and for the training data set.

\[ z_i = \phi_0 \left( Z_{i-1}, Z_1, Z_2, ..., Z_m, u_{i-1} \right) \]
\[ + \sum_{k=1}^{d} \phi_k \left( Z_{i-1}, Z_1, Z_2, ..., Z_m, u_{i-1} \right) z_{i-k} + e_i, \quad \text{(9)} \]

where the impulses \( u_{i-1} = (u_{i-1}, u_{i-2}, ..., u_{i-d}) \) and the coefficients \( \phi_k \left( Z_{i-1}, Z_1, Z_2, ..., Z_m, u_{i-1} \right) \) \((k = 0, 1, 2, ..., d)\) are

\[ \phi_k \left( Z_{i-1}, Z_1, Z_2, ..., Z_m, u_{i-1} \right) = \]
\[ c_{k,0} + \sum_{l=1}^{m} \sum_{j=0}^{r-1} c_{k,l} \prod_{j=0}^{r-1} \exp \left\{ -\frac{(N z_{i-1} - N z_j)^2}{h_l} \right\} \]
\[ + \sum_{j=1}^{d} b_{k,j} u_{i-j}. \]
When the occurrence times of impulses $u_1, u_2, \ldots, u_N$ are specified, the co-efficients, $c_{k,l}$ and $b_{k,l}$ ($k = 0, 1, 2, \ldots, d$), ($l = 1, 2, \ldots, m$), ($j = 1, 2, \ldots, s$) are estimated by the ordinary least squares method. Identifying the occurrence of the impulses is a rather cumbersome job, but not intractable. The further study of the specification of $u_i$ in the form $u_i = h(x_{i-\delta})$ is very important and challenging work which will greatly help the identification of feedback systems in biomedical sciences in future.

### 3.3. Non-Parametric AR Modelling

Recently the non-parametric AR approach has been recognized to be useful, not only in theoretical studies, but also in many examples of real time series analysis (see Valdes et al.58). To understand the advantages and disadvantages of the parametric and non-parametric modelling approaches, we will examine contrasting results of non-parametric AR modelling approach applied to the same epilepsy EEG data. We employed here for the analysis of the epilepsy EEG data a Nadaraya-Watson type Kernel estimate, where the choice of the bandwidth $h$ is one of the crucial tasks for the success of the modelling. We used a commonly used technique for this, i.e. the cross validation method. The basic idea of the cross-validation methods is to estimate a model using a portion of the training data first, then to use the remaining samples to evaluate the prediction risk for the same model. This approach has been proved to be asymptotically optimal in non-parametric modelling26. For the selection of the bandwidth $h$ in the present study, we take the leave-one-out cross-validation rule $CV(h)$ defined by

$$CV(h) = \frac{1}{N-p} \sum_{i=p+1}^{N} \left( x_i - \hat{f}_i(X_{i-1}) \right)^2 K_i \left( \frac{X_{i-1} - X_i}{h} \right)$$

$$K_i \left( \frac{X_{i-1} - X_i}{h} \right) = \prod_{j=1}^{p} \exp \left\{ -\frac{(x_{i-j} - x_{i-j})^2}{h^2} \right\}$$

where \( \hat{f}_i(X_{i-1}) \) is the leave-one-out estimator defined by

$$\hat{f}_i(X_{i-1}) = \frac{1}{(N-p-1)v_h} \sum_{i=p+1}^{N} x_i K_i \left( \frac{X_{i-1} - X_i}{h} \right)$$

$$v_h = \frac{1}{(N-p-1)} \sum_{i=p+1}^{N} K_i \left( \frac{X_{i-1} - X_i}{h} \right)$$

The optimum bandwidth $h$ is the one that minimizes $CV(h)$. The remarkable point about non-parametric AR modelling in nonlinear time series analysis is its capacity for reconstructing very complex nonlinear oscillatory patterns. Fig. 7 shows the noise-free
simulation of the non-parametric model obtained from the training data set of the epilepsy EEG data. As the initial values for the simulation we used an arbitrarily chosen vector, \( X_0 = (100, 100, ..., 100) \). The simulated series exhibits the characteristic pattern of the spike and wave data very clearly. This kind of reconstructability is rarely seen in conventional parametric AR models. So far, the only kind of parametric AR model which can reconstruct the spike and wave patterns successfully is the generalized ExpAR model with multiplicative impulses, described in section (3.2).

The above example seems to suggest the superiority of the non-parametric AR modelling approach over the parametric approach in nonlinear time series modelling. However, we should not jump to this conclusion before we check the prediction performance. The prediction errors of the estimated non-parametric AR model applied to the test data set of the 140 data points are shown in Fig. 8, where we see much larger prediction errors than those in Fig. 7. This means although the non-parametric AR models are superior in reconstructing complex nonlinear patterns in time series they are not necessarily so good for prediction. Another difficulty is that we do not have any standard method for the quantitative assessment of the goodness of a model, where both the prediction performance and the reconstructability of the dynamics of the series are desirable properties. It is clear that, at present, this kind of research is at a very early stage and more experience of elaborate data analysis is needed before these kinds of problems can be satisfactorily addressed.

4. STOCHASTIC THEORIES RELATED TO THE IDENTIFICATION OF MODEL DYNAMICS

4.1. The Fantasy and Reality of Prediction Errors

The reconstruction of dynamics from time series with the expectation of obtaining better prediction than from stochastic models has been one of the aims of "chaos" studies. According to [53] "the definition of chaos proposed at a prestigious international conference on chaos held by the Royal Society in London in 1986 is Stochastic behaviour occurring in a deterministic system." Some statisticians thought that the subject brought a revolution in modelling and data analysis in statistics. Indeed it would truly be revolutionary if all the noise in stochastic models were replaceable by deterministic chaos. Everyone would naturally prefer to have a prediction model with less uncertainty. All statisticians would agree that deterministic prediction, if possible, is preferable to stochastic prediction. Time series analysis would be extremely happy to have time series models with zero system noise variance if possible, so that the innovations \( \nu_t \)

\[ \nu_t = z_t - E [z_t | z_{t-1}, z_{t-2}, \ldots] \]
defined by the model become zero. The point is whether such
deterministic prediction can really work and whether prediction
can really be zero in practice. Although some time series
analysts may still be reluctant to accept leaving behind the idea of
determinism (see how many time series analysts had conceded an
overly rosy view of chaos in the discussion paper by [7]) there are
several reasons why we say "determinism is lost". If we
have a chaos model to explain the behaviour of a time series \( x_t \), i.e.
a model of form

\[
x_t = f_\theta (x_{t-1}, \ldots, x_{t-d_0})
\]

then

i) The true embedding dimension, \( d_0 \), is unknown:
We need to start modelling by guessing a dimension \( d \) since the true
embedding dimension \( d_0 \) is unknown.

ii) The form of the true nonlinear function \( f_\theta (\cdot) \) is unknown.
We need to start modelling with our first guess of the nonlinear
autoregressive function, \( \phi(x_{t-1}, z_{t-2}, \ldots, z_{t-d}) \).

iii) Observation errors are always present:
Observed time series are not free from observation errors and the
observed variables \( z_t, z_{t-1}, z_{t-2}, \ldots \) are not exactly equal to the state
variables, \( x_t, x_{t-1}, x_{t-2}, \ldots \). Even if we happened to fix upon the true
embedding dimension \( d_0 \) and the true nonlinear function
\( f_\theta (x_{t-1}, x_{t-2}, \ldots, x_{t-d_0}) \), it is not likely that the exact deterministic
relation \( z_t = f_\theta (z_{t-1}, \ldots, z_{t-d_0}) \) would hold between the
contaminated observed variables, \( z_t, z_{t-1}, \ldots, z_{t-d_0} \).

iv) Computational stochastic chaos:
Even if the true model were really deterministic, it is impossible to
use exactly the same theoretical model at the computational stage,
where all models become approximate. This means the evolution of
the deterministic model is always exposed to some kind of

PREDICTION-BASED CHARACTERIZATION OF CHAOTIC PHENOMENA

computational perturbation error, and the models actually treated
thus become stochastic models driven by these microscopically
small errors. We could call this type of stochastic model
computational stochastic chaos.

Since prediction is not deterministic and prediction errors
are unavoidable, we are forced to consider the stochastic
distribution of errors or the conditional predictive distributions,
whether we like it or not. Determinism is lost. A similar conclusion
has been reached from the determinists' side, where it has been
shown that small system noise and observation noise are not
distinguishable in practice. If the prediction cannot be
deterministic, our second concern should be unbiased prediction
and smallest prediction error variance which has all along been
one of the main objectives of conventional time series analysis. In
effect, this means there is nothing left to distinguish non-stochastic
chaos studies from conventional stochastic nonlinear time series
analysis.

4.2. Nonlinear Predictors using the Least Squares
Criterion \( E[z_t - Z_t \mid x_{t-1}]^2 \)

The reason why the sum of squares of errors is employed as a
criterion for model estimation is because implicitly it is assumed
that the error distribution is close to being Gaussian. Thus the idea
behind the procedure becomes the same, as pointed out in section
(2), as the maximum likelihood method for nonlinear AR
modelling with Gaussian white noise when \( f(z_{t-1}, \ldots, z_{t-d}) \) is
specified parametrically with parameter \( \theta \), i.e.

\[
z_t = f(z_{t-1}, \ldots, z_{t-d} \mid \theta) + n_t,
\]

where the nonlinear AR part specifies the conditional mean,
and the noise \( n_t \) specifies the conditional variance of the predictive
distribution \( p(z_t \mid z_{t-1}, \ldots, z_{t-d}) \).

It is known, however, that even if the process is
non-Gaussian, the best predictor \( z_{t\mid1:t-1} \) in the sense of
\( E[z_t - z_{t\mid1:t-1}]^2 \) is still given by the conditional mean,
\[ z_{t \mid t-1} = E [z_t \mid z_{t-1}, t_{t-2}, \ldots] . \]

This is not trivial, although it may seem so to time series analysts (see for example [25] for the proof). We know that for Gaussian processes the conditional mean \( z_{t \mid t-1} \) is given by a linear predictor. However, in the non-Gaussian case, there is no reason why a nonlinear predictor might not give a smaller least squares than the best linear predictor. Numerical experience with non-Gaussian processes tends to confirm the superiority (in the least squares sense) of nonlinear predictors over the best linear predictor.

The conditional mean and conditional variance may be written mathematically using the conditional density function as

\[
z_{t \mid t-1} = E [z_t \mid z_{t-1}, z_{t-2}, \ldots] = \int \xi_t p(\xi_t \mid z_{t-1}, \ldots, z_{t-d}) \, d\xi_t
\]

and

\[
\sigma^2_{t \mid t-1} = E[z_t - z_{t \mid t-1}]^2
= \int (\xi_t - z_{t \mid t-1})^2 p(\xi_t \mid z_{t-1}, \ldots, z_{t-d}) \, d\xi_t
\]

Masani and Wiener's mathematical work\(^{32}\) shows that \( E[z_t \mid z_{t-1}, z_{t-2}, \ldots] \) may be approximated more and more closely with increasing numbers of higher order moments for a non-Gaussian process. Wiener\(^{61}\) suggested using a linear combination of nonlinear functions of past observations for the prediction of non-Gaussian processes. In general, it is necessary to use an infinite number of moments to characterize a non-Gaussian process. Hence, an infinitely large number of polynomial terms will also generally be needed to give the best predictor. However, in practice we need to use truncated finite parameter models, where computational explosion problems turned out to be serious. For these reasons, Wiener's general approach with high order moments and high order polynomial models died away in the 1960's.

**PREDICTION-BASED CHARACTERIZATION OF CHAOTIC PHENOMENA** 95

Nonlinear AR modelling in the 1970's, such as ExpAR modelling, aimed for parsimonious parameterization and for stability. In nonlinear AR modelling, the following two problems are critical for model validity in practice:

i) efficient parameterization; and

ii) computational stability.

The number of possible nonlinear terms is huge in nonlinear AR modelling. Since there is no standard way to set an order for model complexity, modelling must proceed on a "case-by-case" basis. In simulating the estimated nonlinear AR model, computational stability is a well-known burdensome problem. Estimated nonlinear AR models are often explosive in simulations, casting doubt on their validity in practical applications. Nonlinear AR models such as ExpAR models\(^{32,35,64}\) or the nonlinear threshold AR model\(^{34,35}\) tried to tame computational instability by considering the dynamics of the instantaneous characteristic roots yielding coefficients of the AR models which were amplitude dependent. In the 1980's, neural network modelling use of the sigmoid function was an ingenious way to perform the same function of taming the explosiveness of the discrete-time nonlinear dynamics. Neural network modelling avoids the problem of overparameterization by shifting the stance from statistical parameter modelling to an algorithmic modelling problem.

The nonlinear prediction problem may be treated in a more general framework by explicitly separating the state variable and the observed variable as is done with the linear state space model for linear predictors. The problem becomes one of nonlinear filtering and prediction with a nonlinear state space model. With a general nonlinear state space model, several nonlinear Kalman filters (discrete time) have been introduced\(^{30,51}\). Here the state space is

\[ x_t = f(x_{t-1}) + g(x_{t-1}) n_t \]

and

\[ z_t = h(x_t) + \varepsilon_t \]
where the state dynamics and the observation equation are both nonlinear.

For example, from a nonlinear AR type predictor such as

$$y_t = \phi(y_{t-1}, \ldots, y_{t-d} | \theta) + u_t$$

and

$$z_t = y_t + \epsilon_t,$$

we obtain the multivariate nonlinear state space representation.

$$x_t = f(x_{t-1} | \theta) + n_t \quad \ldots \quad (11)$$

and

$$z_t = Hx_t + \epsilon_t,$$

where

$$x_t = (y_n, \ldots, y_{t-d+1})',$$

$$n_t = (u_n, 0, 0, \ldots, 0)',$$

$$f(x_{t-1} | \theta) = \{ \phi(y_{t-1}, \ldots, y_{t-d} | \theta), y_{t-1}, \ldots, y_{t-d+1} \}',$$

and

$$H = (1, 0, \ldots, 0).$$

Here the conditional mean of the state is the optimal estimate in terms of minimizing the squares of the errors whatever the conditional distribution $p(x_t | z_t)$.

Although the nonlinear state space model is more general than the nonlinear AR type model, the estimation of parameters is much more cumbersome. In nonlinear state-space modelling, it is necessary to carry out numerical nonlinear optimization for estimation while for nonlinear AR type modelling we can estimate parameters simply by solving a linear equation. As in the case of the nonlinear AR type model, computational stability is a big problem in practical applications. It is necessary to check in advance whether the state dynamic model

$$x_t = f(x_{t-1}) + g(x_{t-1}) n_t$$

really defines a stationary Markov chain process. When the observation noise is large, the state estimation error is also large, and the filtered state estimate $x_{t-1} | t-1$ jumps rapidly to a previously unexpected region for $x_{t-1}$ making the prediction $f(x_{t-1} | t-1)$ unrealistically large or small, and causing a computational explosion in the filtering calculation. Thus the proper parameterization of the dynamic part $x_t = f(x_{t-1})$ of the state equation and the evaluation of the stationarity of the Markov chain process defined by the state equation is also an important problem in nonlinear filtering.

Available algorithms for dealing with nonlinear state space models are rather limited because of this computational explosion problem. The most well-known algorithm is the extended Kalman filter where the first Taylor expansions $\frac{\partial f(x)}{\partial x}$ and $\frac{\partial h(x)}{\partial x}$ of the nonlinear functions $f(x_{t-1})$ and $h(x_t)$ are used. Based on the idea of truncating the Taylor expansion at a higher order, there are some more sophisticated algorithms using second order terms such as $\frac{\partial^2 f(x)}{\partial x^2}$ and $\frac{\partial^2 h(x)}{\partial x^2}$ (see [51]). We will see in section (4.5) that an approximate method based on a different idea (preserving all the differentials of $f(x)$ and $h(x)$) brings a simple and much more consistent algorithm for the nonlinear filtering problem.

Kernel estimator functions can also be useful in estimating nonlinear state space models. In particular, when the AR part is specified by non-parametric kernel functions $K(t)$, the conditional mean and the conditional variance are specified non-parametrically by means of the kernel functions. This point is discussed again in section (4.5).

In the study of nonlinear predictors, either in neural network modelling or non-parametric prediction, there is little explanation of why $E[z]^2$ is used as a criterion for a good predictor. Most nonlinear predictors, such as the multi-layer neural network, or RBF function approximations, use $E[z(t) - z(t) | t - 1]^2$ for the objective function. However, from the previous discussion, it is
clear that the use of the quadratic criterion \( E[\hat{\epsilon}^2] \) is closely connected to the concept of conditional densities, and the conditional mean as the best predictor. Whether \( E[\hat{\epsilon}^2] \) is an appropriate criterion error to be used, for example, when we have a multimodally distributed prediction error seems to be beyond the concern of most people. Another point is that some time series analysts also assess \( m \)-step ahead prediction error using the \( E[\hat{\epsilon}^2] \) criterion. However, when system dynamics are nonlinear, the several-step ahead predictive distribution could be asymmetric or multi-modal even though the one-step ahead prediction error is exactly Gaussian.

Some Bayesian methods (for example, see [51]) consider a maximum a posteriori filter which takes the mode of the conditional density \( p(x_t \mid z_t) \) as an estimate of the state. For the \( E[\hat{\epsilon}^2] \) criterion, a Bayes’ modal predictor is not necessarily the best, since it does not coincide with the conditional mean unless the conditional density is unimodal and symmetric. If the criterion \( E[\hat{\epsilon}^2] \) is not used, however, the conditional mean no longer plays a central role. This is discussed further in section (4.3). For a non-\( E[\hat{\epsilon}^2] \)-criterion, a modal criterion may make more sense than the conditional mean. Therefore, a non-\( E[\hat{\epsilon}^2] \) criterion may be more appropriate, for example, for the assessment of \( m \)-step ahead predictors. In later sections, we will discuss the crucial choice of \( E[\hat{\epsilon}^2] \) or non-\( E[\hat{\epsilon}^2] \) criterion in more detail.

4.3. The Bayesian Approach: A Non \( E[\hat{\epsilon}^2] \) Criterion

Whether we need a non-\( E[\hat{\epsilon}^2] \) criterion rather than a \( E[\hat{\epsilon}^2] \) criterion is not a superfluous question but is rather a delicate question which leads us to crucial choices in data analysis.

It is possible to perform prediction and filtering without any consideration of the prediction error \( \hat{\epsilon}_t = E[z_{t+1} \mid z_t] \). Here the conditional distribution \( p(x_t \mid z_t) \) does not need to be unimodal nor symmetric but could be any distribution. There are several works in this area where the least squares criterion, and hence the conditional mean, no longer plays a special role\(^{16,20,46,51}\). A key role in these methods is played by the log-likelihood function. Since Bayes’ theorem also plays an important role in their methods, their approach is often called the Bayesian approach.

To see how the prediction problem can be treated by the log-likelihood function criterion instead of by the \( E[\hat{\epsilon}^2] \) criterion, we start with a very general state space model without the assumptions of linearity and Gaussianity:

\[
x_t = f(x_{t-1}) + g(x_{t-1})\epsilon_t;
\]

and

\[
z_t = h(x_t) + \epsilon_t.
\]

Here we assume that the data \( z_t \) is a nonlinear observation of the state \( x_t \) with i.i.d. observation errors \( \epsilon_t \) which do not necessarily have the Gaussian distribution. We also assume that the prediction of the state by \( f(x_{t-1}) \) has prediction error \( g(x_{t-1})\epsilon_t \), which could be non-Gaussian and could be asymmetric or even multi-modal. As mentioned in section (4.2), the nonlinear AR type models treated in sections (2) and (3) may be regarded as special cases of the present model, (see model (11)).

The log-likelihood for the general state space model may be written in terms of conditional predictive densities as

\[
\log p(z_1, \ldots, z_n \mid \theta) = \sum \log p(z_t \mid z_{t-1}, \ldots, z_1, \theta).
\]

The conditional density, \( p(z_k \mid z_{k-1}, \ldots, z_1, \theta) \) may be expressed using the observation density \( p(z_k \mid x_k, \theta) \) and state prediction density \( p(x_k \mid z_{k-1}, \ldots, z_1, \theta) \) thus

\[
p(z_k \mid z_{k-1}, \ldots, z_1, \theta) = \int p(z_k \mid x_k, \theta) p(x_k \mid z_{k-1}, \ldots, z_1, \theta) dx_k
\]

The observation density \( p(z_k \mid x_k, \theta) \) is obtained if we specify the model \( z_t = h(x_t) + \epsilon_t \). The state prediction density \( p(x_k \mid z_{k-1}, \ldots, z_1, \theta) \) is obtained from \( p(x_k \mid x_{k-1}, \theta) \) and the filtered state density \( p(x_{k-1} \mid z_{k-1}, \ldots, z_1, \theta) \) by

\[
p(x_k \mid z_{k-1}, \ldots, z_1, \theta) = \int p(x_k \mid x_{k-1}, \theta) p(x_{k-1} \mid z_{k-1}, \ldots, z_1, \theta) dx_{k-1} \quad \ldots (12)
\]
The state prediction density comes from the state dynamic model \( x_t = f(x_{t-1}) + g(x_{t-1}) u_t \). The filtered density \( p(x_{k-1} | z_{k-1}, \ldots, z_1, \theta) \) in this formula can be obtained using Bayes' formula, from the observation density \( p(z_{k-1} | x_{k-1}, \theta) \) at the previous time point \( k - 1 \) and the state predictive density \( p(x_{k-1} | z_{k-2}, \ldots, z_1, \theta) \) at \( k - 1 \) from

\[
p(x_{k-1} | z_{k-1}, \ldots, z_1, \theta) = \frac{p(z_{k-1} | x_{k-1}, \theta) p(x_{k-1} | z_{k-2}, \ldots, z_1, \theta)}{\int p(z_{k-1} | \xi_{k-1}, \theta) p(\xi_{k-1} | z_{k-2}, \ldots, z_1, \theta) d\xi_{k-1}}
\]

The recursive relations of the related predictive densities, observation densities and filtered densities form a non-Gaussian generalization of the Kalman filter\(^{24}\). If we employ a numerical integration method for the above recursive formula, starting with an appropriate initial density distribution \( p(x_0) \) for the initial state \( x_0 \), we can, in principle, calculate all the conditional densities recursively starting from some prior density for the initial state,\(^{27}\) yielding a log-likelihood value for a given state space model.

### 4.4. The Innovation Approach and Nonlinear Filtering

Let us return to the question raised at the beginning of the previous section,

"Who would need a non-\( E[|^2 \) criterion?"

Basically, those —

i) who, for some reason, must use only linear models with asymmetric and/or multi-modal \( p(z_t | z_{t-1}) \); and

ii) who fear that the nonlinear predictor being used may lead to asymmetric and/or multi-modally distributed errors.

Those analysts cannot live comfortably in the looking-glass world inhabited by confident determinists. For determinists confident about the modelling on the basis of exact one-step ahead prediction, prediction models with asymmetric one-step ahead errors and/or multi-modal one-step ahead errors are completely out of bounds. In the everyday world, analysts anticipate being faced with errors of all kinds. At the same time, analysts in this mundane world should take care not to lose motivation for diagnosing and improving the model to reduce the errors to be as close to zero as possible. Using a general criterion which works for any kind of predictor, and a filtering method which works in the most general of frameworks, one might not be able to see much need or incentive for checking the distribution of the prediction error \( v_t = z_t - z_{t|t-1} \) or revising and improving the fitted model.

The innovation approach falls between the two schools; the school of the \( E[|^2 \) criterion and the school of the non-\( E[|^2 \) criterion. The characteristics of the innovation approach are that,

i) the prediction error \( v_t = z_t - z_{t|t-1} \) and \( E[z_t - z_{t|t-1}]^2 \) are central features,

ii) it seeks a model whose prediction error is close to being Gaussian white noise; and

iii) it seeks a model whose variance is as small as possible.

Let us see how the innovation approach drastically simplifies the computational burden of the Bayesian approach described in the previous section, while the logical base-line provided by the log-likelihood function is maintained. In this respect we could say that the innovation approach is compatible with both the \( E[|^2 \) criterion and the likelihood criterion for choosing the predictor.

We start from the following general nonlinear dynamic state space model,

\[
x_t = f(x_{t-1}) + B(x_{t-1}) n_t
\]

and

\[
z_t = C x_t + e_t.
\]

The case where the observation equation is nonlneiar will be discussed later. Here we can expand \( f(x_{t-1}) \) giving
\[ f(x_{t-1}) = \left[ f'(0)x + \sum_{k=1}^{\infty} \frac{1}{(k+1)!} f^{k+1}(0)x^{k+1} \right]_{x=x_{t-1}} + f(0), \]

where

\[ A(x_{t-1}) = \left[ f'(0)x + \sum_{k=1}^{\infty} \frac{1}{(k+1)!} f^{k+1}(0)x^{k+1} \right]_{x=x_{t-1}} \]

Thus the general nonlinear state space model may be written as

\[ x_t = A(x_{t-1})x_{t-1} + B(x_{t-1})n_t + f(0) \]

and

\[ z_t = Cx_t + \varepsilon_t. \]

Although the representation of \( A(x) \) is not unique for multivariate cases, the difference in filtering performance between different representations is negligible, while there are distinct advantages in using this expression, as we shall see later.

The innovations approach also utilizes the following three basic recursive relations, c.f. the Bayesian approach, between the conditional densities,

\[ p(z_k \mid z_{k-1}, \ldots, z_1, \theta) = \int p(z_k \mid x_k, \theta)p(x_k \mid z_{k-1}, \ldots, z_1, \theta) \, dx_k \]

\[ p(x_k \mid z_{k-1}, \ldots, z_1, \theta) = \int p(x_k \mid x_{k-1}, \theta)p(x_{k-1} \mid z_{k-1}, \ldots, z_1, \theta) \, dx_{k-1} \]

\[ p(x_{k-1} \mid z_{k-1}, \ldots, z_1, \theta) \, dx_{k-1} = \frac{p(z_k \mid x_{k-1}, \theta)p(x_{k-1} \mid z_{k-2}, \ldots, z_1, \theta)}{\int p(z_{k-1} \mid \xi_{k-1}, \theta)p(\xi_{k-1} \mid z_{k-2}, \ldots, z_1, \theta) \, d\xi_{k-1}} \]

The difference between the innovations approach and the Bayesian approach is three-fold,

i) the gaussian likelihood is calculated explicitly from the innovations;

ii) \( p(x_k \mid x_{k-1}, \theta) \) is given by a Gaussian distribution; and

iii) \( p(z_{k-1} \mid x_{k-1}, \theta) \) is given by a Gaussian distribution, all of which simplifies the recursive calculation of the conditional densities drastically, resulting in computations as simple and straightforward as in linear Kalman filtering. Since the innovation is given by

\[ v_k = z_t - Cx_{t|t-1} \]

\[ = z_k - \int Cx_k p(x_k \mid z_{k-1}, \ldots, z_1, \theta) \, dx_k \]

\[ = z_k - \int z_k p(z_k \mid z_{k-1}, \ldots, z_1, \theta) \, dz_k \]

the Jacobian of the transformation from \( z_1, \ldots, z_N \) to \( v_1, \ldots, v_N \) is a unit matrix. Then we have

\[ \log p(z_k \mid z_{k-1}, \ldots, z_1, \theta) = \log p(v_k \mid z_{k-1}, \ldots, z_1, \theta), \]

and the log-likelihood function is given by

\[ \log p(z_1, \ldots, z_N \mid \theta) = \sum \log p(v_k \mid z_{k-1}, \ldots, z_1, \theta). \]

Since we assume that system noise \( n_t \) and observation noise \( \varepsilon_t \) are Gaussian, we can use the following Gaussian likelihood,

\[ (-2) \log p(z_1, \ldots, z_N \mid \theta) \]

\[ = \Sigma \left[ \log \sigma_{k|k-1}^2 + \frac{(z_k - Cx_{k|k-1})^2}{\sigma_{k|k-1}^2} \right] + N \log 2\pi, \]

Even though the original process is not Gaussian we can expect the prediction error \( v_k \) to be almost, if not exactly, Gaussian so that it is still appropriate to use the Gaussian likelihood. In this case, the estimates are called Gaussian estimates and are known to share the same attractive asymptotic properties as maximum likelihood estimates. Here, maximizing the log-likelihood function and
minimizing the sum of squares of the prediction errors is asymptotically equivalent. The innovation variance, \( \sigma^2_{k | k - 1} \), is defined by

\[
\sigma^2_{k | k - 1} = \int (z_k - C\hat{x}_{k | k - 1})^2 p(z_k | z_{k - 1}, \ldots, z_1, \theta) dz_k.
\]

We note that, from the previous discussion, all the conditional distributions are composed, locally in time, of two basic distributions, the observation density \( p(z_k | x_k, \theta) \) which we assume Gaussian, and the state prediction density \( p(x_k | x_{k - 1}, \theta) \) and the state prediction density \( p(x_k | x_{k - 1}, \theta) \) which is given from the nonlinear Kalman filtering procedure.

We will need to find estimates \( \hat{x}_{k | k - 1} \) and \( \sigma^2_{k | k - 1} \) denoted by \( \hat{x}_{k | k - 1} \) and \( \sigma^2_{k | k - 1} \) and hence find the estimated (-2) log-likelihood from

\[
\Sigma \left[ \log \sigma^2_{k | k - 1} + \frac{(z_k - C\hat{x}_{k | k - 1})^2}{\sigma^2_{k | k - 1}} \right] + n \log 2\pi \quad \text{(13)}
\]

For the linear case, where we have

\[ x_k = Ax_{k - 1} + Bw_k, \]

we find that, when we take expectations, we obtain \( x_{k | k - 1} = A\hat{x}_{k - 1 | k - 1} \) which gives us estimates \( \hat{x}_{k | k - 1} \) found from

\[ \hat{x}_{k | k - 1} = A\hat{x}_{k - 1 | k - 1}. \]

For the nonlinear case, i.e. when \( x_k = A(x_{k - 1})x_{k - 1} + B(x_{k - 1}) \sigma_n n_k + f(0) \), there is no exact formula for

\[ x_{k | k - 1} = \int x_k p(x_k | z_{k - 1}, \ldots, z_1) dx_k. \]

However, \( A(x_{k - 1})x_{k - 1} + f(0) \) could be taken as a reasonable approximation.

\[
\text{PREDICTION-BASED CHARACTERIZATION OF CHAOTIC PHENOMENA 105}
\]

Here \( x_{k - 1 | k - 1} \) is the conditional mean of \( x_{k - 1} \) given \( z_{k - 1}, \ldots, z_1 \), i.e.

\[
x_{k - 1 | k - 1} = \int x_{k - 1} p(x_{k - 1} | z_{k - 1}, \ldots, z_1, \theta) dx_{k - 1}
\]

which has been estimated at time \( k - 1 \) by \( \hat{x}_{k - 1 | k - 1} \) from the previous recursion. Hence

\[
\hat{x}_{k | k - 1} = A(x_{k - 1})\hat{x}_{k - 1 | k - 1} + f(0)
\]

Then the estimated innovation at the time point \( k \) is given by

\[
\hat{v}_k = z_k - C(A(x_{k - 1})\hat{x}_{k - 1 | k - 1} + f(0)),
\]

The actual value of the estimates, \( \hat{x}_{k | k - 1}, \hat{x}_{k + 1 | k} \) are calculated simply using the following recursive nonlinear Kalman filtering scheme,

\[
\hat{x}_{k + 1 | k} = A(\hat{x}_{k | k})\hat{x}_{k + 1 | k} + f(0), \quad \hat{x}_{k | k} = \hat{x}_{k | k - 1} + K_k \hat{v}_k,
\]

\[
K_k = P_k C^T(CP_k C^T + \sigma^2_e)^{-1},
\]

\[
P_k = A(\hat{x}_{k | k}) V_{k - 1} A(\hat{x}_{k | k})' + B(\hat{x}_{k | k}) \Sigma_w B(\hat{x}_{k | k})',
\]

\[
V_k = P_k - K_k C P_k
\]

and

\[
\sigma^2_{k | k - 1} = CP_k C^T + \sigma^2_e
\]

with initial values, \( x_0, V_0 \) and a system parameter vector \( \theta \) which includes parameters characterizing the system dynamics in addition to parameters characterizing the observation error variance \( \sigma^2_e \) and the system noise variance \( \Sigma_w \). For the derivation of the scheme we refer to Ozaki et al.\textsuperscript{44,}

The difference between the well-known extended Kalman filter and the present filter is in the transition matrix \( A(\hat{x}_{k - 1 | k - 1}) \). In the extended Kalman filter the transition of the state is approximated by
which does not equal \( f(x_{t-1}) \), while the state transition of the present scheme is exact since it holds that

\[
\frac{\partial f(x)}{\partial x} \bigg|_{x=x_{t-1}}
\]

The present scheme is more consistent that the extended Kalman filter in this sense.

Another problem is that the observation equation may also be nonlinear so that

\[
z_t = h(x_t) + \epsilon_t
\]

for general \( h(\cdot) \). A general solution to this needs various kinds of approximation techniques. The most commonly used and computationally efficient scheme is the extended Kalman filter scheme, where \( h(x_t) \) is approximated by \( \partial h(x_t)/\partial x_t \). For a more accurate and efficient method than the extended Kalman filter, we refer to Ozaki and Thomson\(^{42,43}\), where the new scheme together with the conventional extended Kalman filter are applied to the seasonal adjustment problem through nonlinear dynamic systems.

4.5. The Conditional Mean and the Skeleton

With the \( E[\cdot]^2 \)-criterion, the conditional mean plays an essential role, not only in the quantitative characterization of the predictors, but also in specifying the qualitative characterization of the process. The qualitative nature of the time series model has sometimes been discussed using the deterministic part of the identified linear or nonlinear AR models. The recent fashion is to call this part of the model its "skeleton". We would like to point out in this section that the conditional mean has always been essential in specifying the qualitative nature of any time series, linear or nonlinear.

In linear, AR model cases,

\[
x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_d x_{t-d} + n_t
\]

the conditional mean is specified by the linear difference equation,

\[
x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_d x_{t-d}
\]

\[
\cdots \quad (15)
\]

In nonlinear cases, for example for the ExpAR model where,

\[
x_t = (\phi_1 + \theta_1 \exp(-\lambda_{t-1}^2)) x_{t-1} + \ldots + \phi_d + \theta_d \exp(-\lambda_{t-1}^2) x_{t-d} + n_t
\]

the conditional mean is given by the nonlinear difference equation,

\[
x_t = (\phi_1 + \theta_1 \exp(-\lambda_{t-1}^2)) x_{t-1} + \ldots + \phi_d + \theta_d \exp(-\lambda_{t-1}^2) x_{t-d}
\]

\[
\cdots \quad (16)
\]

The skeletons of (15) and (16) have been used to characterize the qualitative nature of the linear and nonlinear time series. For example, the number of peaks in the spectrum of the AR process is specified by the number of complex conjugate pairs, \( \lambda_i, \bar{\lambda}_i \), of complex roots of the characteristic equation,

\[
\Lambda^d - \phi_1 \Lambda^{d-1} + \phi_2 \Lambda^{d-2} + \ldots + \phi_d = 0
\]

The stability of the process is also characterized by the absolute values of the characteristic roots. If all \( |\lambda_i| < 1 \) the process is stable and otherwise the process is explosive.

For ExpAR models, the singular points (i.e. fixed points) and limit cycles are all specified by the skeleton (16). The stability conditions for singular points or limit cycles are also fully specified by the skeleton. The result and techniques used to find the qualitative properties of the process defined by nonlinear difference equations are essentially the same as the results and the techniques used for the qualitative characterization of the continuous time nonlinear differential equations by Poincaré and others (see [34]). The conditions for the existence of a strange
attractor for ExpAR could be obtained using the same techniques discussed in [34]. Here the emphasis should be on elucidating which conditions give us no stable singular points and no stable limit cycles at the same time as the process remains absolutely stable and non-explosive for any $z_{t-1} \in R$. It must be remembered, however, that the existence of a strange attractor is crucial for the existence of "random" behaviour only when the process is purely deterministic. When the process is stochastic from the start, however small the system noise, it makes no difference whether the identified model has a strange attractor, at least for statisticians.

The fact that the conditional mean determines the qualitative nature of the time series implies that any method for estimating the conditional mean, whether non-parametric kernel estimation, neural network modelling, or the innovations approach, not only specifies the predictors but also specifies the qualitative nature of the time series. In the non-parametric AR modelling case, the kernel estimate for the conditional mean is, for example in the Nadaraya-Watson case, given by pull

$$M(z_{t-1}, z_{t-2}, \ldots, z_{t-d}) = \frac{1}{N-d} \sum_{i=d+1}^{N} z_i \prod_{j=1}^{d} K\left(\frac{z_{t-j} - z_{t-j}}{h}\right)$$

$$+ \left\{ \frac{1}{N-d} \sum_{i=d+1}^{N} \prod_{j=1}^{d} K\left(\frac{z_{t-j} - z_{t-j}}{h}\right) \right\}$$

$$V(z_{t-1}, z_{t-2}, \ldots, z_{t-d}) = \frac{1}{N-d} \sum_{i=d+1}^{N} z_i^2 \prod_{j=1}^{d} K\left(\frac{z_{t-j} - z_{t-j}}{h}\right)$$

$$+ \left\{ \frac{1}{N-d} \sum_{i=d+1}^{N} \prod_{j=1}^{d} K\left(\frac{z_{t-j} - z_{t-j}}{h}\right) \right\}$$

$$- M^2 (z_{t-1}, z_{t-2}, \ldots, z_{t-d}).$$

PREDICTION-BASED CHARACTERIZATION OF CHAOTIC PHENOMENA

If we assume that the observation error is negligibly small, these conditional mean and conditional variance estimates imply that, in terms of the $E[\cdot]^2$ criterion for the internal $[t-1, t]$, the best predictor is given by the following nonlinear prediction model,

$$z_t = (1, 0, \ldots, 0)x_t$$

... (17)

and

$$x_t = A(x_{t-1})x_{t-1} + B(x_{t-1})n_t.$$

Here $A(x_{t-1})$ is a function of the state $x_{t-1} = (z_{t-1}, \ldots, z_{t-d})'$ defined from the non-parametric estimate $\hat{M}(z_{t-1}, z_{t-2}, \ldots, z_{t-d})$ of the conditional mean, and $B(x_{t-1})$ is a function of the state $x_{t-1}$ defined from the non-parametric estimate $\hat{V}(z_{t-1}, z_{t-2}, \ldots, z_{t-d})$ of the conditional variance in the following way,

$$\hat{M}(z_{t-1}, z_{t-2}, \ldots, z_{t-d}) = A(x_{t-1})x_{t-1} + \hat{M}(0)$$

$$= \hat{M}(x_{t-1})$$

and

$$\hat{V}(z_{t-1}, z_{t-2}, \ldots, z_{t-d}) = B(x_{t-1}) \Sigma B(x_{t-1})'$$

$$= \hat{V}(x_{t-1}).$$

Since

$$\hat{M}(x_{t-1}) = \left[ \hat{M}(0)x + \sum_{k=1}^{\infty} \frac{1}{(k+1)!} \hat{M}^{k+1}(0) x^{k+1} \right]_{x=x_{t-1}} + \hat{M}(0)$$

$$= A(x_{t-1})x_{t-1} + \hat{M}(0),$$

we have

$$A(x) = \hat{M}(0) + \sum_{k=1}^{\infty} \frac{1}{(k+1)!} \hat{M}^{k+1}(0) x^k.$$
The expression for \( A(x_{t-1}) \) is not unique for the case where we have a multivariate state dimension case. However, the fact that there are different possible representations does not prevent any of these representations from being much superior to the unique first order approximation in specifying the model characteristics. We also can tell at least which is more desirable when two candidates are presented explicitly. Obviously the one which gives the best value of the model fitting criterion is the best choice for the statisticians.

The eigenvalues \( \lambda_1 (x_{t-1}), \ldots, \lambda_d (x_{t-1}) \) of the transition matrix \( A(x_{t-1}) \) of the state space model (17) characterize the qualitative nature of the process, in the same way as the characteristic roots \( \lambda_1 (z_{t-1}), \ldots, \lambda_d (z_{t-1}) \) of the instantaneous characteristic equation

\[
\Lambda^d - \left\{ \phi_1 + \theta_1 \exp(-\gamma \chi_{t-1}^2) \right\} \Lambda^{d-1} - \\
- \left\{ \phi_{d-1} + \theta_{d-1} \exp(-\gamma \chi_{t-1}^2) \right\} \Lambda \\
- \left\{ \phi_d + \theta_d \exp(-\gamma \chi_{t-1}^2) \right\} = 0
\]

specified the qualitative nature of the ExpAR processes. In non-parametric AR modelling, analysts often have difficulty in finding out what is actually happening in terms of the dynamic nature of the process even though the conditional mean estimates are informative and useful. However, the above discussion shows that the results obtained by the non-parametric methods can also be useful in suggesting the dynamics of the local (in time) eigenvalues of the state, which could then be used for introducing a new more sophisticated parametric model having the same capabilities as those implied by the dynamics found from the non-parametric results.

4.6. Quantitative Assessment of Skeletons

When analysts discuss the success of the fitted model in nonlinear time series modelling, they often simulate the estimated model without noise, i.e. the "skeleton", and discuss whether the simulated time series is similar to the original time series. It may be encouraging for data analysts if they see a similar pattern in the simulation of the deterministic part, but statisticians need to be alive to the fact that models should be accepted without any quantitative assessment. In nonlinear time series modelling, some time series analysts seem to have been so much influenced by determinists that they seem to have stopped paying attention to the variance of prediction errors or the log-likelihood value of the model.

Looking at the skeleton of a fitted model makes sense only in understanding the "qualitative stochastic" nature of the process, but we must be careful when we use it for "quantitative" discussion. Analysts should bear in mind that a model makes sense in statistical argument only when it is handled in total, i.e. when the autoregressive part is combined with the error term. When we use the skeleton to explain the phenomena underlying the observed data, we are essentially interpreting the observed data by means of the following state space model,

\[
x_t = \phi(x_{t-1}, x_{t-2}, \ldots, x_{t-d}) \\
z_t = x_t + \epsilon_t,
\]

where the system noise \( \sigma^2 \) = 0. Then the adequacy of the model may be explicitly measured by the log-likelihood function \( l(\hat{\theta}, \sigma_n, \hat{\sigma}_e; x_0 | z_1, z_2, \ldots, z_N) \) for the model (18) using the techniques given previously in section (4.4). Here for the skeleton, \( \hat{\theta} \) and \( \hat{\sigma}_0 \) are fixed to the parameter values estimated through the stochastic model. The value \( \sigma^2 \) is the only parameter we cannot obtain from the skeleton, and we need to give it a certain value in order to calculate the log-likelihood. Of course we can maximize the log-likelihood with respect to \( \sigma^2 \) and the value \( \hat{\sigma}_e^2 \) which maximizes the log-likelihood is the maximum likelihood estimate of the observation noise variance for the skeleton.
The log-likelihood value of the skeleton must certainly be very poor because it interprets all the deviation of the data points from the deterministic prediction based on the initial state value as observation error. This means that, when it comes to actual nonlinear filtering, the deterministic state model is used for state prediction, and any resulting prediction error $\hat{e}_k = z_k - C\hat{x}_{k \mid k-1}$ is consistently ignored when making the state estimate, so that our estimation of $\hat{x}_k$ is no better at time $k$ than it was at time $k-1$, i.e.

$$\hat{x}_{k \mid k} = \hat{x}_{k \mid k-1}$$

rather than

$$\hat{x}_{k \mid k} = \hat{x}_{k \mid k-1} + K_k \hat{e}_k$$

which we would have in the stochastic case. Here $K_k$ is the Kalman gain calculated by

$$K_k = P_k C' \left( CP_k C' + \sigma_e^2 \right)^{-1}.$$

$V_k$ is the state filtered estimation error variance, $V_k = E[(x_k - \hat{x}_{k \mid k}) (x_k - \hat{x}_{k \mid k})']$ and $P_k$ is the prediction error variance $P_k = E[(x_k - \hat{x}_{k \mid k-1})^2]$, which would evolve in time following,

$$P_k = A(\hat{x}_{k \mid k}) V_{k-1} A(\hat{x}_{k \mid k})'$$

instead of the usual

$$P_k = A(\hat{x}_{k \mid k}) V_{k-1} A(\hat{x}_{k \mid k})' + B(\hat{x}_{k \mid k}) \Sigma_n B(\hat{x}_{k \mid k})',$$

for the stochastic case. Ref. [39] shows, by numerical examples, that even for artificially generated deterministic Lorenz chaos, the estimate of the state using nonlinear filtering without taking into account recent prediction errors leads to disastrous results.

The log-likelihood function assessment of the stochastic model with noise variance $\sigma^2$ smaller than the estimated $\hat{\sigma}^2$ can be also assessed in the same way. The state space model for this is the same as for the stochastic nonlinear state space model,

$$x_t = f(x_{t-1} | \theta) + n_t$$

and

$$z_t = Cx_t + e_t.$$

Here the system noise $\sigma^2_n$ variance could be set artificially smaller than the estimated variance $\hat{\sigma}^2$. Then the log-likelihood value,

$$l(\theta, \hat{\sigma}^2, \sigma^2_n \mid z_1, z_2, \ldots, z_N)$$

can be calculated in the same way as in section (4.4). Although the likelihood is smaller than the estimated model, it is usually much better than the deterministic model where $\sigma^2_n$ is set equal to 0.

4.7. The Mythology associated with Initial Value Sensitivity

Why is initial value sensitivity important for deterministic chaos? The answer used to be "Because it is the source of randomness". When even artificially generated deterministic chaos such as Lorenz chaos cannot be freed from small system noise because of small observation errors and computational errors in calculating the state transition of nonlinear dynamics there is nothing to distinguish it from data generated from a stochastic nonlinear time series model. It seems that the tone of the chaotists' argument has changed since 1990. Before 1990, hardly anyone dreamed about calling a Markov diffusion process defined by the Lorenz model differential equation with a very small Gaussian white noise "chaos". Nowadays system noise has been accepted by chaos modellers and a dynamic model driven by system noise is called "noisy chaos". Some dynamicists and nonlinear time series analysts still claim there is a difference between a "noisy chaos" model, a conventional Markov diffusion model or a nonlinear time series model. One clear raison-d'être for chaos models seems to be the argument of the initial value sensitivity of the model.

Naturally we could ask the same question as for non-deterministic chaos "Why is initial value sensitivity important in the stochastic case?" The answer is not so clear as in the deterministic case. Since noisy chaos belongs to the family of general nonlinear state space models, the problem at issue could
be discussed in the framework of the general nonlinear state space model,

\[ x_t = A(x_{t-1} | \theta)x_{t-1} + B(x_{t-1} | \theta)\eta_t, \]

and

\[ z_t = Cx_t + \epsilon_t. \]

Here the constant terms are dropped from the state dynamic and observation equations for simplicity. With the computational method given in the innovation approach, we can calculate the log-likelihood function \( l(\theta; \sigma_w^2, \sigma_v^2, x_0 | z_1, z_2, ..., z_N, V_0) \). Therefore, the best initial value \( \hat{x}_0 \) for the state is given by maximizing the log-likelihood function. Here the initial state error variance \( V_0 = E[(x(0) - x_0)(x(0) - x_0)'] \) needs to be specified to calculate the log-likelihood. With the innovation approach \( V_0 \) can also be made into a parameter so that we can estimate \( V_0 \) as a kind of indicator of the uncertainty of the initial state \( x_0 \). When \( x_0 \) is close to the true initial value, the likelihood function tends to choose a smaller error variance \( V_0 \). The way in which the uncertainty of the state estimate evolves in time is shown explicitly in the nonlinear filtering scheme for the nonlinear state space model. For example the state prediction error variance \( P_k = E[(x_k - x_{k+1})'(x_k - x_{k+1})'] \) is defined by the initial \( V_0 \), the transition matrix \( A(\hat{x}_{k+1} \mid k) \) and the variance \( B(\hat{x}_{k+1} \mid k)' \Sigma \, B(\hat{x}_{k+1} \mid k)' \) of the system noise, giving

\[ P_k = A(\hat{x}_{k-1} \mid k) V_{k-1} A(\hat{x}_{k-1} \mid k)' + B(\hat{x}_{k-1} \mid k)' \Sigma \, B(\hat{x}_{k-1} \mid k)' \]

This equation shows that regardless of the accuracy of the initial state estimate, i.e. regardless of whether \( V_0 \) is a large or very small, the future prediction error variance, e.g. \( P_1 \), will be affected by the term \( B(\hat{x}_{1 \mid 1}) \Sigma \, B(\hat{x}_{1 \mid 1})' \), where \( \hat{x}_{1 \mid 1} = x_{1 \mid 0} + K_1(z_1 - Cx_{1 \mid 0}) \). Then the state filtering error variance \( V_1 \) at the next time step and the Kalman gain \( K_1 \) are governed by the equations,

\[ V_1 = P_1 - K_1 C P_1 \]

and

\[ K_1 = P_1 C' (CP_1 C' + \sigma_e^2)^{-1}. \]

It is clear from these observations that, after several steps, the initial state value \( x_0 \) and its uncertainty \( V_0 \) are not important factors for the state prediction error variance \( P_k \) because of the involvement of \( B(\hat{x}_{k \mid k}) \Sigma \, B(\hat{x}_{k \mid k})' \) at each time step \( k \). Of course, things are different for the deterministic case, where the prediction error variance evolves in time without \( B(\hat{x}_{k \mid k}) \Sigma \, B(\hat{x}_{k \mid k})' \). Here the prediction error variance \( P_k \) and the filtering error variance \( V_k \) are equivalent, \( P_k = V_k \), a function of \( V_0 \) given by,

\[ P_k = A(\hat{x}_{k \mid k}) V_{k-1} A(\hat{x}_{k \mid k})' \]

\[ = A(\hat{x}_{k \mid k}) A(\hat{x}_{k-1 \mid k-1}) V_{k-2} A(\hat{x}_{k-2 \mid k-2})' \]

\[ = A(\hat{x}_{k \mid k}) A(\hat{x}_{k-1 \mid k-1}) \times A(\hat{x}_{k \mid k}) V_{k-1} A(\hat{x}_{k \mid k})' \]

\[ = A(\hat{x}_{k \mid k}) \Sigma \, A(\hat{x}_{k \mid k})'. \]

This equation shows how \( P_k \), the future uncertainty of the state, is determined quantitatively by the initial state uncertainty \( V_0 \). If \( V_0 = 0 \) and \( \Sigma = 0 \), the initial state is exact and the system is deterministic, then theoretically speaking the prediction error should be zero forever. If \( V_0 \neq 0 \), the error \( (x_k - x_{k \mid 0}) \) is passed on to the next step after being deformed by the transition matrix \( A(\hat{x}_{k \mid k}) \). Whether the error is amplified or reduced depends on the local eigenvalues \( \lambda_1(\hat{x}_{k \mid k}), \lambda_2(\hat{x}_{k \mid k}), ..., \lambda_d(\hat{x}_{k \mid k}) \), of the transition matrix \( A(\hat{x}_{k \mid k}) \) at time point \( k \).

Estimating the initial state \( x_0 \) of chaos should not be a taboo topic which statisticians are not allowed to tough. Some analysts claim that chaos models leads to chaotic likelihood functions, in spite of the proven success of Ozaki's general method for the maximum likelihood estimation of nonlinear stochastic dynamical
systems. To show that the likelihood function really is useful in the estimation of "chaos" models, the authors have recently shown numerical examples where the maximum likelihood method is used to estimate Lorenz and Rikitake chaos. In a similar manner to the estimation of ordinary statistical models such as the linear ARMA model, the initial state \( x_0 \) can be estimated using the maximum likelihood method given in the previous section (4.4). In this case, the likelihood function \( l(\theta, \sigma_n^2, \sigma_w^2, x_0 \mid z_1, z_2, \ldots, z_N) \) just becomes a function of the system parameter \( \theta \), the system noise variance \( \sigma_w^2 \), the observation noise variance \( \sigma_n^2 \), and the initial state \( x_0 \). Some numerical examples demonstrating the estimation of the initial state for Rikitake two-disk dynamo chaos are also shown in [40].

The concept of shadowing is also connected with the myth of the initial value sensitivity of chaos. Some analysts expect that when the system noise \( n_t \) of stochastic chaos

\[
x_t = f(x_{t-1}, \theta) + n_t
\]

is small, the trajectory of the deterministic chaos model,

\[
x_t = f(x_{t-1}, \theta)
\]

will coincide with the stochastic trajectory given an appropriate initial value. The idea is called shadowing and is discussed in various works in the mathematical and physical literature. If we consider their argument from the state space framework, these researchers are expecting to be able to use shadowing to find a deterministic model (2) whose trajectory is equivalent to that of stochastic chaos (19). If this were true, we could find a deterministic trajectory which follows the same path as the stochastic chaos (19) simply by adjusting the initial state \( x_0 \) while the other parameter vector \( \theta \) is fixed to be the same as for the stochastic model and the system noise variance \( \sum_n \) is fixed to be zero. This claim is rather optimistic because the parameter space of the log-likelihood function for the stochastic model is less restrictive than that of the deterministic chaos. The likelihood of a less restrictive model always give a better likelihood than that of a restrictive model. If deterministic chaos gave zero prediction error \( v_k = z_k - Cx_{k-1} = 0 \) for all \( k > 0 \) only by adjusting \( x_0 \) for the deterministic model (2), the prediction error variance \( \sigma_k^2 \) would become zero and \( -2 \) log-likelihood would become \(-\infty\). For those experienced applied time series analysts, this kind of claim is absurd. They know it never happens in real data analysis and is simply out of the question. Even without having had experience of data analysis, researchers may now easily check that this shadowing never happens by using artificially generated data and the numerical procedure for the maximum likelihood method given in the section (4.4).

**ACKNOWLEDGEMENTS**

The authors are grateful to the Takahashi Foundation for Industrial and Economic Research, TWAS (UNESCO), the Institute of Statistical Mathematics Co-operative Research Committee, and the Ministry of Education, Science and Culture, Japan, for supporting this joint work.

**REFERENCES**


