Stochastic, Dynamic Modelling and Signal Processing: Time Variable and State Dependent Parameter Estimation

Peter Young

Chapter from *Nonlinear and Nonstationary Signal Processing* edited by W. J. Fitzgerald *et al.*, Cambridge University Press: Cambridge, 2000, 74-114.

1 Introduction

Previous publications (e.g. Young, 1978, 1983, 1993a,b, 1998a,b, 1999a,b; Young and Runkle, 1989; Young and Minchin, 1991; Young *et al.*, 1991; Young and Lees, 1993; Young and Beven, 1994; Young and Pedregal, 1997, 1998, 1999) have discussed an approach to nonstationary and nonlinear signal processing based on the identification and estimation of stochastic models with time variable (TVP) or state dependent (SDP) parameters. Here the term 'nonstationarity' is assumed to mean that the statistical properties of the signal, as defined by the parameters in an associated stochastic model, are changing over time at a rate which is 'slow' in relation to the rates of change of the stochastic state variables in the system under study. Although such nonstationary systems exhibit nonlinear behaviour, this can often be approximated well by TVP (or piece-wise linear) models, the parameters of which can be estimated using recursive methods of estimation in which the parameters are assumed to evolve in a simple stochastic manner (e.g. Young, 1984, 1999a). On the other hand, if the changes in the parameters are functions of the state or input variables (i.e. they actually constitute stochastic state variables), then the system is truly nonlinear and likely to exhibit severe nonlinear behaviour. Normally, this cannot be approximated in a simple TVP manner; in which case, recourse must be made to the alternative, and more powerful SDP modelling methods that are the main topic of this chapter.

The extension of the TVP estimation methods to allow for state dependency, as described here, involves two statistical stages.

- First, the non-parametric *identification* of the state dependency using recursive methods of time variable parameter estimation which allow for rapid (state dependent) parametric change. As we shall see, the standard methods of TVP estimation developed previously for nonstationary time series analysis need to be modified considerably in this SDP setting to allow for the much more rapid temporal changes that arise from the state dependency.

- Second, the parameterization of the identified non-parametric relationships, followed by the *statistically efficient estimation* of the (now normally constant) parameters that characterize these nonlinearities.

The first identification stage in this process exploits recursive *Fixed Interval Smoothing* (FIS) algorithms, combined with special data re-ordering and 'back-fitting' procedures, to obtain estimates of any state dependent parameter variations. These state dependencies are estimated in the form of non-parametric relationships (graphs) between the estimated rapid parameter variation and the associated state or input variable(s). Parameterization of these non-parametric relationships can be accomplished in various ways, from simple curve fitting based on weighted least squares methods (Young, 1993a; Young and Beven, 1994) to the use of neural networks or radial basis functions.

Having identified a structural form for the nonlinear model of the system based on the parameterized nonlinear relationships, this model is converted into a stochastic state space form. The final estimation phase of the nonlinear modelling then exploits *Maximum Likelihood* (ML) methods of estimation, based on Gaussian Assumptions for the stochastic disturbances and the application of *Prediction Error Decomposition* (Schweppe, 1965). If successful, this yields statistically efficient estimates of the constant parameters in the identified nonlinear state space model. The resulting model should then provide a parametrically efficient representation of the stochastic, nonlinear system that has considerable potential for use in subsequent signal processing, time series analysis, forecasting and automatic control system design. For example, the methodology described here exploits recursive estimation in an *off-line* manner but this sequential processing of the data facilitates the development of related *on-line adaptive* methods of signal processing, forecasting and control.

Although primarily concerned with nonlinear state dependent parameter models, as outlined above, the chapter also provides a sequel to a previous paper (Young, 1999a) that discusses the simpler class of 'linear' TVP regression relationships. These include the *Dynamic AutoRegressive eXogenous variables* (DARX) model, the constant parameter version of which is often used in the modelling of linear stochastic, dynamic systems. As a prelude to our discussion of SDP estimation, therefore, the next section 2 considers the alternative *Dynamic Transfer Function* (DTF) model, using a new instrumental variable method of FIS, and shows how this is much superior to the DARX model when measurement noise is present (the *errors-in-variables* situation). This leads naturally to the definition of the more complex SDP transfer function models of truly nonlinear, stochastic dynamic systems and the associated methods of statistical identification and estimation.

In order to illustrate the practical application and utility of both the TVP and SDP methods, the chapter also contains a number of simulation examples, as well as a practical study involving a re-analysis of the famous Nicholson blowfly data (Nicholson, 1954). Other, practical examples cited in the

references cover a variety of application areas from the environment through engineering to economics.

2 TVP Transfer Function Models: DTF Model Estimation

A previous paper (Young, 1999a) has discussed the estimation of time variable parameters in the various kinds of 'linear' regression model. One of these, the *Dynamic*[1] *Auto-Regressive eXogenous variables* (DARX) model is capable of modelling the input-output behaviour of stochastic, dynamic systems. The DARX model relating a single input variable u_t to an output variable y_t, can be written in the following form:

$$y_t = -a_{1,t}y_{t-1} - \cdots - a_{n,t}y_{t-n} + b_{0,t}u_{t-\delta} + b_{1,t}u_{t-\delta-1} + \cdots + b_{m,t}u_{t-\delta-m} + e_t \quad (1)$$

or, in transfer function terms,

$$y_t = \frac{B(L,t)}{A(L,t)}u_{t-\delta} + \frac{1}{A(L,t)}e_t \qquad e_t = N(0,\sigma^2). \quad (1a)$$

In these equations, L is the backward shift operator, i.e., $L^r y_t = y_{t-r}$, $A(L,t)$ and $B(L,t)$ are *time variable coefficient polynomials* in L of the following form:

$$\begin{aligned} A(L,t) &= 1 + a_{1,t}L + a_{2,t}L^2 + \cdots + a_{n,t}L^n \\ B(L,t) &= b_{0,t} + b_{1,t}L + b_{2,t}L^2 + \cdots + b_{m,t}L^m. \end{aligned} \quad (1b)$$

The term δ is a pure time delay, measured in sampling intervals, which is introduced to allow for any temporal delay that may occur between the incidence of a change in u_t and its first effect on y_t; and e_t is a zero mean, white noise input with Gaussian normal amplitude distribution and variance σ^2.

Unfortunately, the DARX model is limited in practical terms since it depends on the assumption of the above, rather specific, signal topology, with the noise entering the model through a restricted AR process with a polynomial $A(L,t)$ equal to that of the denominator polynomial in the main DTF between $u_{t-\delta}$ and y_t. A more general Dynamic Transfer Function (DTF) model, without the restrictions of the DARX, is the following:

$$y_t = \frac{B(L,t)}{A(L,t)}u_{t-\delta} + \xi_t \quad (2a)$$

where ξ_t represents uncertainty in the relationship arising from a combination of measurement noise, the effects of other unmeasured inputs and modelling

[1]The term 'dynamic' is used here for historical reasons (see Young, 1999a) to mean a time variable parameter ARX model.

error. Normally, ξ_t is assumed to be independent of u_t and is modelled as an AutoRegressive (AR) or AutoRegressive-Moving Average (ARMA) stochastic process (see e.g. Box and Jenkins, 1970; Young, 1984), although even this restriction can be avoided by the use of instrumental variable methods, as discussed below.

Equation (2a) can be written in the following vector equation form:

$$y_t = \mathbf{z}_t^T \mathbf{p}_t + \mu_t \tag{2b}$$

where,

$$\begin{aligned} \mathbf{z}_t^T &= [-y_{t-1} \ -y_{t-2} \ \ldots \ -y_{t-n} \ u_{t-\delta} \ \ldots \ u_{t-\delta-m}] \\ \mathbf{p}_t &= [a_{1,t} \ a_{2,t} \ \ldots \ a_{n,t} \ b_{0,t} \ \ldots \ b_{m,t}]^T \end{aligned} \tag{2c}$$

and $\mu_t = A(L,t)\xi_t$. For convenience of notation, let $\mathbf{p}_t$ be defined as follows,

$$\mathbf{p}_t = [p_{1,t} \ p_{2,t} \ \cdots \ p_{n+m+1,t}]^T \tag{2d}$$

with $p_{i,t}$, $i = 1, 2, \ldots, n+m+1$, relating to the TF model parameters $a_{i,t}$ and $b_{j,t}$ through (2c).

In order to estimate the assumed time variable model parameters in $\mathbf{p}_t$, it is necessary to make some assumptions about the nature of their temporal variability. Reflecting the statistical setting of the analysis and referring to previous research on this topic, it seems desirable if this is characterized in some stochastic manner. Normally, when little is known about the nature of the time variability, this model needs to be both simple but flexible. One of the simplest and most generally useful class of stochastic, state space models involves the assumption that the ith parameter, $p_{i,t}$, $i = 1, 2, \ldots, n+m+1$, is defined by a two-dimensional stochastic state vector $\mathbf{x}_{i,t} = [l_{i,t} \ d_{i,t}]^T$, where $l_{i,t}$ and $d_{i,t}$ are, respectively, the changing level and slope of the associated TVP. This selection of a two-dimensional state representation of the TVPs is based on practical experience over a number of years. Initial research by the author and others in the 1960s (Young, 1969, 1970a) tended to use a simple scalar random walk (RW) model for the parameter variations. However, later work in the 1980s (see above references) showed the value of modelling not only the level changes in the TVPs but also their rates of change, as in the definition of $\mathbf{x}_{i,t}$, above.

The stochastic evolution of each $\mathbf{x}_{i,t}$ (and, therefore, each of the $n+m+1$ parameters in $\mathbf{p}_t$) is assumed to be described by a Generalized Random Walk (GRW) process defined in the following State Space (SS) terms:

$$\mathbf{x}_{i,t} = \mathbf{F}_i \mathbf{x}_{i,t-1} + \mathbf{G}_i \boldsymbol{\eta}_{i,t} \qquad i = 1, 2, \ldots, m+n+1, \tag{2e}$$

where

$$\mathbf{F}_i = \begin{bmatrix} \alpha & \beta \\ 0 & \gamma \end{bmatrix}, \qquad \mathbf{G}_i = \begin{bmatrix} \delta & 0 \\ 0 & \varepsilon \end{bmatrix}$$

and $\boldsymbol{\eta}_{i,t} = [\eta_{1i,t}\ \eta_{2i,t}]^T$ is a 2×1, zero mean, white noise vector that allows for stochastic variability in the parameters and is assumed to be characterized by a (normally diagonal) covariance matrix $\mathbf{Q}_{\eta i}$. This general model comprises as special cases the *Integrated Random Walk* (IRW: $\alpha = \beta = \gamma = \varepsilon = 1$; $\delta = 0$); the scalar *Random Walk* (RW: scalar but equivalent to (2e) with $\beta = \gamma = \varepsilon = 0$; $\alpha = \delta = 1$: i.e. just the first equation in (2e), see below); the intermediate case of *Smoothed Random Walk* (SRW: $0 < \alpha < 1$; $\beta = \gamma = \varepsilon = 1$; and $\delta = 0$); the first order autoregressive process (AR(1): again scalar with $\beta = \gamma = \varepsilon = 0$; $0 < \alpha < 1$; $\delta = 1$; and, finally, both the *Local Linear Trend* (LLT: $\alpha = \beta = \gamma = \varepsilon = \delta = 1$)[2]; and *Damped Trend* (DT: $\alpha = \beta = \delta = \varepsilon = 1$; $0 < \gamma < 1$): see Harvey, 1984, 1989. The various, normally constant, parameters in this GRW model (α, β, γ, δ, ε and the elements of $\mathbf{Q}_{\eta i}$) are often referred to as *hyper-parameters*. This is to differentiate them from the TVPs that are the main object of the estimation analysis. However, the hyper-parameters are also assumed to be unknown *a priori* and need to be estimated from the data, as we shall see in the subsequent discussion.

The full GRW model (2e) was introduced in Jakeman and Young (1979, 1984): further discussion and practical examples appear in Young (1988), Young *et al.* (1989), Young and Ng (1989), Ng and Young (1990). Note that, in the case of the RW model, i.e.,

$$l_{i.t} = l_{i,t-1} + \eta_{1i,t}; \qquad l_{i,t} = p_{i,t}, \tag{3}$$

each parameter can be assumed to be time-invariant if the variance of the white noise input $\eta_{1i,t}$ is set to zero. Then, the stochastic TVP setting reverts to the more normal, constant parameter TF model situation. In other words, if RW models with zero variance white noise inputs are specified for the model parameters, the recursive *Instrumental Variable* (IV) estimation algorithm described below for the general stochastic TVP case will provide recursive estimates that are identical to those obtained with the normal recursive IV estimation algorithm for TF models with constant parameters (Young, 1984). Of course, there is some added value to the recursive solution even in this situation, since the user is provided with the recursive estimates over the whole interval $t = 1, 2, \ldots, N$. These can provide additional useful information on the model: for example, they show how the estimates are converging and can be used (e.g. Brown *et al.*, 1975; Young, 1984) to detect both the presence of potential parametric change and possible over-parameterization (i.e. the model contains too many parameters to provide unambiguous estimation results).

Clearly other, more general and higher order stochastic processes could be used to model the stochastic TVPs, provided such models can be identified satisfactorily from the data. For example the higher order IRWs (Double and

[2]Interestingly, the LLT model can be considered simply as the combination of the simpler RW and IRW models.

Triple Integrated Random Walk (DIRW, TIRW), etc.), the Integrated or Double Integrated AutoRegressive (IAR, DIAR: see Young, 1994) model, and even more general processes (e.g. Pedregal and Young, 1996, 1998). However, the more complex models introduce additional hyper-parameters that would have to be well identified from the data and optimized, thus introducing potential practical difficulties.

The idea of assuming that the model parameters evolve over time as nonstationary stochastic variables may seem complex at first sight but it is, in fact, just a statistical device to allow for the estimation of parametric change. After all, the assumption of the RW model is simply a means of introducing into the estimation problem the freedom for the associated parameter to vary at each sample in time by a small random amount defined by the variance of the white noise input $\eta_{1i,t}$. And the more complex GRW models in (2e) are just a way of refining and adding to this freedom. In fact, it can be shown (Young and Pedregal, 1998) that the GRW assumptions on the parameter variations have an *implicit* but physically interpretable effect: they make the recursive parameter estimates, at any sample time t, depend only on the local data in the vicinity of this sample, with the selected GRW model defining the local weighting effect. In the case of the RW model, for instance, this weighting effect or 'kernel' has a Gaussian-like shape that applies maximum weight to the current data with declining weight each side. And the 'bandwidth' of the kernel is defined by the ratio of the variance of the white noise input $\eta_{1i,t}$ to the residual variance σ^2 (the Noise Variance Ratio (NVR): see later). This can be related to the more *explicit* use of localized data weighting in methods such as locally weighted kernel regression (e.g. Holst *et al.*, 1996; Young and Pedregal, 1996) and 'wavelet' methods (e.g., Daubechies, 1988) that are currently receiving so much attention in the statistical and signal processing literature.

Having introduced the GRW models for the parameter variations, an overall SS model can then be constructed straightforwardly by the aggregation of the subsystem matrices defined in (2e), with the 'observation' equation defined by the model equation (2b): i.e.,

$$\begin{aligned} \text{Observation Equation:} \quad & y_t = \mathbf{H}_t \mathbf{x}_t + \mu_t \qquad \text{(i)} \\ \text{State Equations:} \quad & \mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{G}\boldsymbol{\eta}_t \quad \text{(ii)}. \end{aligned} \tag{4a}$$

where,

$$\mathbf{x}_t = \left[x_{1,t}^T \; x_{2,t}^T \; \ldots \; x_{n+m+1,t}^T \right]^T. \tag{4b}$$

If $p = 2(n+m+1)$, then $\mathbf{F}$ is a $p \times p$ block diagonal matrix with blocks defined by the $\mathbf{F}_i$ matrices in (2e); $\mathbf{G}$ is a $p \times p$ block diagonal matrix with blocks defined by the corresponding subsystem matrices $\mathbf{G}_i$ in (2e); and $\boldsymbol{\eta}_t$ is a p-dimensional vector containing, in appropriate locations, the white noise input vectors $\boldsymbol{\eta}_{i,t}$ ('system disturbances' in normal SS terminology) to each of the

GRW models in (2e). These white noise inputs, which provide the stochastic stimulus for parametric change in the model, are assumed to be independent of the observation noise e_t and have a covariance matrix $\mathbf{Q}$ formed from the combination of the individual covariance matrices $\mathbf{Q}_{\eta,i}$. Finally, $\mathbf{H}_t$ is a $1 \times p$ vector of the following form,

$$\mathbf{H}_t = [-y_{t-1}\ 0\ -y_{t-2}\ 0\ \ldots\ y_{t-n}\ 0\ u_{t-\delta}\ 0\ \ldots\ u_{t-\delta-m}\ 0] \tag{4c}$$

that relates the scalar observation y_t to the state variables defined by (4a)(ii), so that it represents the DTF model (2b), with each parameter defined as a GRW process. In the case of the scalar RW and AR(1) models, the alternate zeros are simply omitted.

The SS formulation in equations (4) is particularly well suited for optimal recursive estimation in which the time variable parameters (acting as surrogate 'states' in this SS formulation) are estimated sequentially whilst working through the data in temporal order (usually termed 'forward-pass filtering'). In the off-line situation, where all the time series data are available for analysis, this filtering operation is accompanied by optimal recursive smoothing (see e.g. Young, 1984; Young *et al.*, 1998). Here, the estimates obtained from the forward-pass, filtering algorithm are updated sequentially whilst working through the data in *reverse* temporal order (usually termed 'backward-pass smoothing') using a backwards-recursive *Fixed Interval Smoothing* (FIS) algorithm, where the 'fixed interval' is the interval covered by the total sample size N.[3]

The reason for this two-pass approach is easy to understand. The forward-pass filtering estimate of $\mathbf{x}_t$, which defines the estimated TVPs, can be denoted by $\hat{\mathbf{x}}_{t|t}$ (or simply $\hat{\mathbf{x}}_t$, for convenience) since it represents the estimate at sample t given only the data up to and including sampling instant t. However, under our assumption that each of the parameters evolve stochastically according to the equation (2e), a superior 'smoothed' estimate $\hat{\mathbf{x}}_{t|N}$, exists and can be generated by the FIS algorithm, in which the estimate at t is based on all the data over the sampling interval $t = 1, 2, \ldots, N$. As a result, the phase lag associated with the forward-pass filtering estimate (since it cannot anticipate any change until the evidence for change in the series has been processed) is eliminated on the backward smoothing pass. Thus, any variation in the parameters is estimated as it occurs, without any lag effect (indeed, it may even be anticipated if the smoothing effect is substantial, as it can be in high noise situations). This proves particularly useful in operations such as interpolation over gaps in the data, estimation and removal of individual components from the data (signal extraction), and seasonal adjustment.

[3]On-line 'Fixed Lag Smoothing' is also possible, where the recursive estimation works in a forward-pass, filtering mode but with smoothed estimates provided at every sampling instant t over a finite interval of l samples into the past (i.e. over the interval $t - l$ to t), but this is not discussed here.

In our previous publications on this topic, a standard algorithmic approach to the problem has been utilized. This involves a the two step (prediction-correction) version of the forward-pass recursive filtering algorithm; followed by a version of the FIS algorithm which is stable numerically and which has a structure allowing for the handling of missing observations and outliers. More specifically, it is an adapted version (e.g. Young, 1988; Young *et al.*, 1998 and the references therein) combining Bryson and Ho's recursion for the Lagrange multipliers (Bryson and Ho, 1969) with the state update recursion of Norton (1975, 1986). It should be noted that the recursive filtering algorithm is closely related to the Kalman Filter (KF: 1960) and is often referred to as such. The difference is that the $\mathbf{H}_t$ matrix in the present, recursive TVP estimation context for the TF model (2b), is based on measured variables. In particular, the output variables y_{t-i}, $i = 1, 2, \ldots, n$, in $\mathbf{H}_t$ are affected by the noise ξ_t (the *errors-in-variables* problem); whereas, strictly, $\mathbf{H}_t$ in the KF has to be composed of exactly known (but, if necessary, time variable), deterministic coefficients.

This difference is important in the present TF context since it can be shown that the TVP estimates obtained from the standard recursive filtering/smoothing algorithm (see later, section 3.1) will be asymptotically biased away from their 'true' values. The level of this bias is dependent on the magnitude of the measurement noise and it can be problematic in high noise situations, particularly if the parameters are physically meaningful (see e.g. Young, 1984 for a discussion of this problem)[4]. For this reason, it is necessary to modify the standard algorithm to avoid these biasing problems. This can be achieved by attempting to model the noise μ_t in some manner (e.g. Norton, 1975, 1986). However, since μ_t is a complex, nonstationary, noise process, its complete estimation is not straightforward. A new, alternative approach, which does not require modelling μ_t, provided it is independent of the input u_t, is the recursive *TVP Instrumental Variable* (TVPIV) method.

In relation to the time series y_t, $t = 1, 2, \ldots, N$,the recursive TVPIV filtering/smoothing algorithm has the following form:

1. Forward-Pass Symmetric IV Equations (iterative)

Iterate the following recursive equations (5a) - 5(c) for $j = 1, 2, \ldots, I_T$, with $\hat{\mathbf{H}}_t = \mathbf{H}_t$ for $j = 1$:

Prediction:

$$\begin{aligned} \hat{\mathbf{x}}_{t|t-1} &= \mathbf{F}\hat{\mathbf{x}}_{t-1} \\ \hat{\mathbf{P}}_{t|t-1} &= \mathbf{F}\hat{\mathbf{P}}_{t-1}\mathbf{F}^T + \mathbf{G}\mathbf{Q}_r\mathbf{G}^T. \end{aligned}$$

[4]It is less important if the model is to be used within a forecasting context since the forecasts produced by the model are not biased (although they may not be statistically efficient).

Correction:

$$\begin{aligned} \hat{\mathbf{x}}_t &= \hat{\mathbf{x}}_{t|t-1} + \hat{\mathbf{P}}_{t|t-1}\hat{\mathbf{H}}_t^T \left[1 + \hat{\mathbf{H}}_t\hat{\mathbf{P}}_{t|t-1}\hat{\mathbf{H}}_t^T\right]^{-1} \{\, y_t - \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1} \,\} \\ \hat{\mathbf{P}}_t &= \hat{\mathbf{P}}_{t|t-1} + \hat{\mathbf{P}}_{t|t-1}\hat{\mathbf{H}}_t^T \left[1 + \hat{\mathbf{H}}_t\hat{\mathbf{P}}_{t|t-1}\hat{\mathbf{H}}_t^T\right]^{-1} \hat{\mathbf{H}}_t\hat{\mathbf{P}}_{t|t-1} \end{aligned} \tag{5a}$$

where,

$$\hat{\mathbf{H}}_t = [-\hat{x}_{t-1}, 0, -\hat{x}_{t-2}, 0, \ldots, \hat{x}_{t-n}, 0, \ldots, u_{t-\delta}, 0, \ldots, u_{t-\delta-m}, 0] \tag{5b}$$

$$\hat{x}_t = \frac{\hat{B}_{j-1}(L,t)}{\hat{A}_{j-1}(L,t)} u_{t-\delta}. \tag{5c}$$

The FIS algorithm is in the form of a backward recursion operating from the end of the sample set to the beginning [5].

2. Backward-Pass Fixed Interval Smoothing IV (FISIV) equations (single pass)

$$\begin{aligned} \hat{\mathbf{x}}_{t|N} &= \mathbf{F}^{-1}\left[\hat{\mathbf{x}}_{t+1|N} + \mathbf{G}\mathbf{Q}_r\mathbf{G}^T\mathbf{L}_t\right] \\ \mathbf{L}_t &= \left[\mathbf{I} - \hat{\mathbf{P}}_{t+1}\hat{\mathbf{H}}_{t+1}^T\hat{\mathbf{H}}_{t+1}\right]^T \left[\hat{\mathbf{F}}^T\hat{\mathbf{L}}_{t+1} - \hat{\mathbf{H}}_{t+1}^T\left\{\, y_{t+1} - \hat{\mathbf{H}}_{t+1}\hat{\mathbf{x}}_{t+1} \,\right\}\right] \\ \hat{\mathbf{P}}_{t|N} &= \hat{\mathbf{P}}_t + \hat{\mathbf{P}}_t\hat{\mathbf{F}}^T\hat{\mathbf{P}}_{t+1|t}^{-1}\left[\hat{\mathbf{P}}_{t+1|N} - \hat{\mathbf{P}}_{t+1|t}\right]\hat{\mathbf{P}}_{t+1|t}^{-1}\mathbf{F}\hat{\mathbf{P}}_t, \end{aligned} \tag{5d}$$

with $\mathbf{L}_N = 0$.

In these algorithms, the $p \times p$ Noise Variance Ratio (NVR) matrix $\mathbf{Q}_r$ and the $p \times p$ matrix $\hat{\mathbf{P}}_t$ are defined as follows,

$$\mathbf{Q}_r = \frac{\mathbf{Q}}{\sigma^2}; \qquad \hat{\mathbf{P}}_t = \frac{\mathbf{P}_t^*}{\sigma^2}, \tag{5e}$$

where $\mathbf{P}_t^*$ is the error covariance matrix associated with the state estimates which, in the present TVP context, define the estimated uncertainty in the parameters. For simplicity, it is normally assumed that the NVR matrix $\mathbf{Q}_r$ is diagonal, although this is not essential. The NVR parameters that characterize $\mathbf{Q}_r$ (as well as any other unknown hyper-parameters in the SS model (4)) are unknown prior to the analysis and clearly need to be estimated on the basis of the time series data y_t and u_t before the filtering and smoothing algorithms can be utilized. The optimization of these hyper-parameters is discussed in the next sub-section.

The main difference between the above algorithm (5a)–(5e) and the standard filtering/smoothing algorithms (see later section 3.1) is the introduction of 'hats' on the $\hat{\mathbf{H}}_t$ vector and the $\hat{\mathbf{P}}_t$ matrix, and the use of an iterative IV

[5] An alternative FIS algorithm is available in which, at each backwards recursion, the estimate $\hat{\mathbf{x}}_{t|N}$ is based on an update of the filtering estimate $\hat{\mathbf{x}}_t$ (see Young, 1984). This can be specified as an alternative to 5(d)

solution in the forward-pass algorithm. In (5b) $\hat{\mathbf{H}}_t$ is the *IV vector*, which is used by the algorithm in the generation of all the $\hat{\mathbf{P}}_t$ terms and is the main vehicle in removing the bias from the TVP estimates. The subscript $j-1$ on $\hat{A}_{j-1}(L,t)$ and $\hat{B}_{j-1}(L,t)$ indicates that the estimated DTF polynomials in the *Auxiliary Model*, (5c), which generates the *instrumental variables* $\hat{x}_t$ that appear in the definition of $\hat{\mathbf{H}}_t$, are updated in an iterative manner, starting with the least squares estimates of these polynomials. Iteration is continued for T iterations, until the forward pass (filtered) IV estimates of the TVPs are no longer changing significantly: normally only 3 or 4 iterations are required.

This new recursive-iterative IV approach to time variable parameter estimation is based on the IV algorithm for *constant* parameter TF models (e.g. Young, 1984 and the prior references therein), except that the *symmetric gain* version of the IV algorithm (Young, 1970b; 1984, p.183) is used, rather than the more usual asymmetric version. This is necessary in order that the standard recursive FIS algorithm in (5d) can be used to generate the smoothed estimates of the TVPs.

Maximum Likelihood (ML) Optimization of Hyper-Parameters

The approach to ML optimization based on *Prediction Error Decomposition* (PED) derives originally from the work of Schweppe (1965), who showed how to generate likelihood functions for Gaussian signals using the Kalman filter (see also Bryson and Ho, 1969; page 389). Its importance in the present UC context was probably first recognized by Harvey (1981) and Kitagawa (1981). Since then, it has become one of the two standard approaches to the problem (the other being the *Expectation and Minimization* (EM) algorithm: Dempster *et al.*, 1977).

In the case of the simpler DARX model (Young, 1999a), $\hat{\mathbf{H}}_t$ and $\hat{\mathbf{P}}_t$ in the recursive TVP least squares estimation algorithm (5a) are replaced by $\mathbf{H}_t$ and $\mathbf{P}_t$, respectively, and iteration is not required. With given initial values for the hyper-parameters, this algorithm will yield the one-step-ahead prediction errors (also termed the 'innovations' or 'recursive residuals') ε_t, where

$$\varepsilon_t = y_t - \mathbf{H}_t \hat{\mathbf{x}}_{t|t-1} \qquad t = 1, 2, \ldots, N. \tag{6}$$

If the first p observations are regarded as fixed, the log-likelihood function of $y_{p+1}, \ldots, y_N$ can be defined in terms of the standard 'regression' form of prediction error decomposition , i.e.,

$$\begin{aligned} \log L \;\; = \;\; & \frac{-(N-p)}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\sum_{t=p+1}^{N}\log(1+\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^t) \\ & -\frac{1}{2\sigma^2}\sum_{t=p+1}^{N}\frac{\varepsilon_t^2}{1+\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T}, \end{aligned} \tag{7}$$

where it can be shown that $\sigma^2(1 + \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T)$ is the variance of ε_t, so that the last term in (7) is based on the sum of squares of the normalized one-step-ahead prediction errors. Now the ML estimate of σ^2, conditional on the hyper-parameters, is given by

$$\hat{\sigma}^2 = \frac{1}{N-p}\sum_{t=p+1}^{N}\frac{\varepsilon_t^2}{1 + \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T}, \tag{8}$$

so that it can be estimated in this manner and 'concentrated out' of the expression (7) by substituting (8) into (7), to yield the following expression for the 'concentrated likelihood':

$$\log(L_c) = -\frac{N-p}{2}\log(2\pi + 1) - \frac{1}{2}\sum_{t=p+1}^{N}\log(1 + \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T) - \frac{N-p}{2}\log(\hat{\sigma}^2) \tag{9}$$

which needs to be maximized with respect to the unknown hyper-parameters in order to obtain their ML estimates.

Since (9) is nonlinear in the hyper-parameters, the likelihood maximization needs to be carried out numerically. Consequently, it is more convenient to remove the constant term (since it will play no part in the minimization) and multiply (9) by -2, to yield:

$$\log(L_c) = \sum_{t=p+1}^{N}\log(1 + \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T) + (N-p)\log(\sigma^2), \tag{10}$$

which then needs to be *minimized*. This minimization is accomplished by initiating the optimization with the hyper-parameter estimates either selected by the user or set to some default values (in both cases, ensuring that the resulting optimization does not converge on a local minimum). The recursive TVP estimation algorithm is used repeatedly to generate the one step ahead prediction errors ε_t and, thence, the log-likelihood value in (10) associated with the latest selection of hyper-parameter estimates made by the optimization algorithm. The optimization algorithm then adjusts its selection of hyper-parameter estimates in order to converge on those estimates that minimize (10). Further details of this and alternative ML optimization procedures are given, for example, in Harvey (1989) and Harvey and Peters (1990). Typical methods that can be used for numerical optimization are the '*fmins*' and '*fminu*' functions available in the MATLAB software system, or their equivalents, although more complex and efficient procedures are available.

In the case of the DTF model, The same basic PED approach to ML optimization can be used, but with $\mathbf{H}_t$ and $\mathbf{P}_t$ in equations (7) to (10) replaced by $\hat{\mathbf{H}}_t$ and $\hat{\mathbf{P}}_t$, respectively. The equations formed in this manner can then be considered as an IV version of PED and as an approximation to the standard TVP regression version. This approximation is justified by the link between

optimal IV estimation and the classical ML approach to TF model estimation (Young and Jakeman, 1979; Young, 1984) arising from the theoretical results of Pierce (1972). Its practical efficacy is demonstrated in the following simulation example.

Simulation Example 1

As a simple example of DTF modelling, consider the estimation of the parameters in the following first order TVP model:

$$x_t = \frac{b_0}{1 + a_{1,t} z^{-1}} u_{t-2} \qquad u_t = \mathrm{N}(0, 6.25)$$
$$y_t = x_t + \xi_t \qquad \xi_t = \mathrm{N}(0, 2.56),$$

or, written in the form of equation (2a):

$$y_t = \frac{b_0}{1 + a_{1,t}} u_{t-2} + \xi_t$$

where the $b_0 = 0.5$ is constant and $a_{1,t}$ varies sinusoidally, as $0.9 \sin(0.02t)$. Estimation is based on the measurements of y_t and u_t, $t = 1, 2, \ldots, 2000$, shown in the upper panels of Figure 1; the lower panel shows the variation of $a_{1,t}$. It will be noted that the overall noise/signal ratio on the output measurement y_t is high (0.71 by variance; 0.84 by standard deviation).

It is assumed that no information is available on the variation of the parameters and so RW models are chosen for both of the two unknown parameters. ML optimization of the NVRs, as described in the previous subsection, then yields $\mathbf{Q}_r = \mathrm{diag}[0.00105 \;\; 1.905 \times 10^{-20}]$, where it will be noted that the NVR for the b_0 parameter is insignificantly different from zero, indicating that the parameter is identified as being time invariant. This shows how, quite objectively, the ML optimization is able to identify the relative temporal variability of the model parameters from the input-output data, without any other *a priori* information. The FIS estimated TVP $\hat{a}_{1,t|N}$ is shown in Figures 2 and 3 (upper panels), where it is compared with the DARX estimates (lower panels). The superiority of the DTF estimates is particularly clear in Figure 2: not only are the DTF estimates much better than the equivalent DARX estimates, but the estimated standard errors (shown dashed) are more realistic. As in the case of similar situations with constant parameter models, the least squares DARX standard errors are too optimistic and, in contrast to the DTF standard errors, they do not encompass the true variation of the parameters. The DTF model with these estimated parameters explains the data well: the coefficient of determination (COD) based on the simulated model output compared with the noise free output is $R_T^2 = 0.93$ (93% of the data explained by the TVP model); whilst for the DARX model, this is reduced to $R_T^2 = 0.85$. The model residuals (innovations) for the DTF model are also superior: they

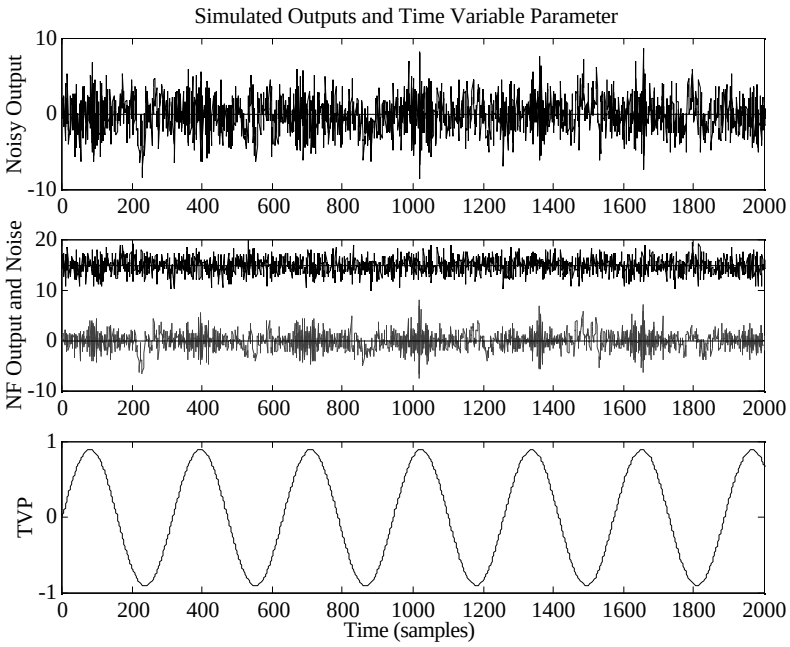

Figure 1. Simulated DTF model. Top panel: noisy output. Middle panel: noise free output and additive noise (+15). Lower panel: variation of $a_{1,t}$ parameter

have an approximately normal amplitude distribution; and, as required, both the autocorrelation function (acf) and the cross correlation function (ccf) between the residuals and the input u_t, are insignificant at all lags. In contrast, the ccf for the DARX model residuals shows significant correlation with u_t at some lags.

The advantage of the complete, off-line FIS estimation in relation to forward-pass (filtering) estimation is illustrated in the upper panel of Figure 3, which compares a 250 sample section of the FIS estimates for the DTF model (full line) with both the actual variation (dashed line) and the forward pass (IV, filtered) estimates generated by algorithm (5a) at the final iteration. In the lower panel, similar results are shown for the DARX estimates. It is clear that the FIS estimates in the DTF case are very close to the actual values and that, as expected, the associated forward pass (IV, filtered) DTF estimates are much less smooth and have a pronounced lag. Similar characteristics can be observed for the DARX smoothed and filtered estimates but, in addition, they also show clear evidence of the deleterious least squares biasing effects. A measure of the overall estimation accuracy is the COD for the error between

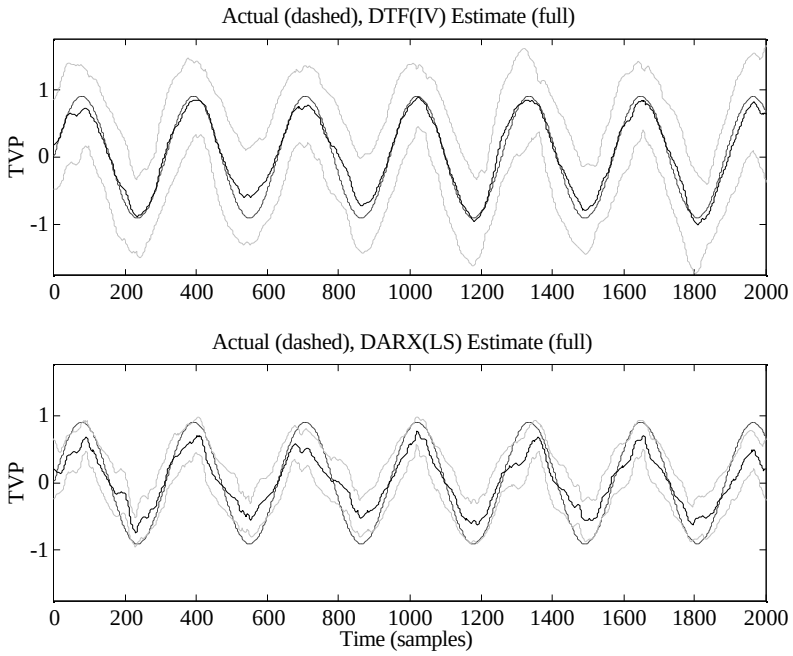

Figure 2. Simulated DTF model. Top panel: FISIV estimate of $a_{1,t}$ (full line) compared with actual variation (dashed), with standard error bound shown dotted. Lower panel: biased least squares FIS estimate of the TVP under the assumption of that the model is DARX (lines as for upper panel)

the estimated and actual TVPs: in the case of the DTF model, this yields an $R_T^2 = 0.96$ (i.e. 96% of the actual variation in the parameter explained by the TVP estimate) for the FIS estimates, and a much lower $R_T^2 = 0.85$ for the associated IV filtered estimates. For the DARX model, the values are lower still at $R_T^2 = 0.80$ and $R_T^2 = 0.70$, respectively.

Note finally, in this example, that the b_0 parameter has been kept constant to show how the ML hyper-parameter optimization is able to detect this fact (without the provision or use of any *a priori* information) and so inform the DTF recursive algorithm (5) that b_0 is time-invariant. However, if both parameters are allowed to vary in a similar manner (e.g. $a_{1,t}$ sinusoidal, as here, and $b_{0,t}$ with a similar frequency, cosine variation between 0.9 and -0.9), then the associated NVRs are optimized as $\mathbf{Q}_r$= diag [0.00101 0.00232], showing that the ML optimization has found strong evidence of temporal changes in both parameters. As a result, they are both estimated well (although, as might

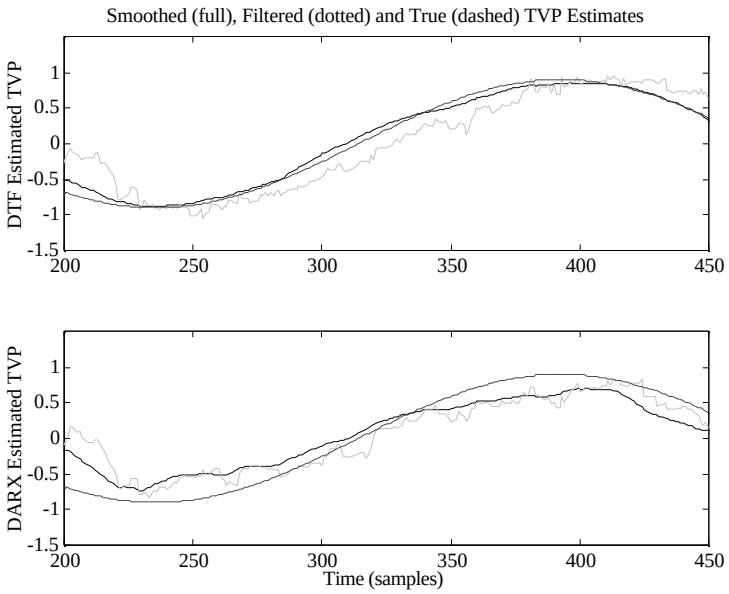

Figure 3. Simulated DTF model. Top panel: FISIV (full line) and forward-pass filtered (dotted line) estimates of $a_{1,t}$ compared with the actual variation (dashed). Lower panel: biased least squares estimates of $a_{1,t}$ under the assumption of that the model is DARX (lines as for upper panel)

be expected, there is some deterioration in the $\hat{a}_{1,t}$ estimate, when compared with the above results: see Young and McKenna, 1999).

3 SDP Transfer Function Models

As we have seen above, the approach to TVP estimation discussed in the last section works very well in situations where the parameters are slowly varying when compared to the observed temporal variation in the measured system inputs and outputs. Although such DARX and DTF models are nonlinear systems, since the same inputs, injected at different times, will elicit quite different output responses, the resultant nonlinearity is fairly mild. It is only when the parameters are varying at a rate commensurate with that of the system variables themselves that the model behaves in a heavily nonlinear or even chaotic manner. We will refer to *State Dependent Parameter* (SDP) models of this type as *State Dependent parameter ARX* (SDARX) and *State Dependent parameter TF* (SDTF) models respectively. The rest of the chapter

will show how these types of model can be used to represent a wide variety of nonlinear stochastic systems and time series.

3.1 SDARX Estimation

The SDARX model equation takes the following form,

$$y_t = \mathbf{z}_t^T \mathbf{p}_t + e_t \qquad e_t = \mathrm{N}(0, \sigma^2), \tag{11a}$$

where now

$$\begin{array}{rcl} \mathbf{z}_t^T & = & [-y_{t-1} \ -y_{t-2} \ \cdots \ -y_{t-n} \ u_{t-\delta} \ \cdots \ u_{t-\delta-m}] \\ \mathbf{p}_t & = & [a_1(\boldsymbol{\chi}_t) a_2(\boldsymbol{\chi}_t) \cdots a_n(\boldsymbol{\chi}_t) b_0(\boldsymbol{\chi}_t) \cdots b_m(\boldsymbol{\chi}_t)]^T, \end{array} \tag{11b}$$

while $a_i(\boldsymbol{\chi}_t)$, $i = 1, 2, \ldots, n$ and $b_j(\boldsymbol{\chi}_t)$, $j = 0, 2, \ldots, m$ are the state dependent parameters, which are assumed to be functions of a non-minimal state vector $\boldsymbol{\chi}_t^T = \left[\mathbf{z}_t^T \ \mathbf{U}_t^T\right]$. Here $\mathbf{U}_t = [U_{1,t} \ U_{2,t} \ \cdots \ U_{r,t}]^T$ is a vector of other variables, not necessarily direct functions of y_t and u_t that may affect the relationship between these two primary variables (see Young, 1993a; Young and Beven, 1994).

The simpler signal topology for this 'affine' model (see Co, 1996) means that FIS estimation can be based on the standard recursive least squares (RLS) filtering/smoothing equations (e.g. Young, 1984, 1999a): i.e. the following RLS form of the algorithm (5a)–(5e):

1. Forward Pass Recursive LS Equations (single pass)

Prediction:

$$\begin{array}{rcl} \hat{\mathbf{x}}_{t|t-1} & = & \mathbf{F}\hat{\mathbf{x}}_{t-1} \\ \hat{\mathbf{P}}_{t|t-1} & = & \mathbf{F}\hat{\mathbf{P}}_{t-1}\mathbf{F}^T + \mathbf{G}\mathbf{Q}_r\mathbf{G}^T. \end{array}$$

Correction:

$$\begin{array}{rcl} \hat{\mathbf{x}}_t & = & \hat{\mathbf{x}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}_t^T \left[1 + \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T\right]^{-1} \left\{ y_t - \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1} \right\}. \\ \mathbf{P}_t & = & \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}_t^T \left[1 = \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T\right]^1 \mathbf{H}_t\mathbf{P}_{t|t-1} \end{array}$$

2. Backward Pass IV Smoothing Equations

$$\begin{array}{rcl} \hat{\mathbf{x}}_{t|N} & = & \mathbf{F}^{-1}\left[\hat{\mathbf{x}}_{t+1|N} + \mathbf{G}\mathbf{Q}_r\mathbf{G}^T\mathbf{L}_t\right] \\ \mathbf{L}_t & = & \left[\mathbf{I} - \mathbf{P}_{t+1}\mathbf{H}_{t+1}^T\mathbf{H}_{t+1}\right]^T \left[\mathbf{F}^T\mathbf{L}_{t+1} - \mathbf{H}_{t+1}^T\left\{ y_{t+1} - \mathbf{H}_{t+1}\hat{\mathbf{x}}_{t+1} \right\}\right] \\ \mathbf{P}_{t|N} & = & \mathbf{P}_t + \mathbf{P}_t\mathbf{F}^T\mathbf{P}_{t+1|N}^{-1}\left[\mathbf{P}_{t+1|N} - \mathbf{P}_{t=1|t}\right]\mathbf{P}_{t+1|t}^{-1}\mathbf{F}\mathbf{P}_t, \\ \mathbf{L}_N & = & 0. \end{array}$$

Since the parameter vector $\mathbf{p}_t$ is potentially state-dependent, it may vary at a rate commensurate with the temporal variations in y_t, u_t and $U_{i,t}$, and so it cannot be assumed that the simple GRW model (2e) is appropriate to describe the parametric variation over time. At first sight, it would appear that the stochastic state space model should include prior information on the nature of the parameter variation, if the TVP estimation methodology discussed in previous sections is to work satisfactorily. Fortunately, it is possible to remove this requirement if we resort to the rather unusual procedure, at least within a time series context, of sorting the data in a *non-temporal order*. Then, if the ordering is chosen so that the SDP variations associated with the *sorted* series are smoother and less rapid, it is more likely that a simple GRW process can be utilized to describe their evolution.

For example, if the time series are sorted in some common 'ascending order of magnitude' manner (i.e. the *sort* operation in MATLAB), then the rapid natural variations in y_t and u_t are effectively eliminated from the data and replaced, in the sorted data space, by much smoother and less rapid variations. And if the SDP's are, indeed, related to these variables, then they will be similarly affected by the sorting. Following FIS estimation, however, these SDP estimates can be 'unsorted' (a trivial *unsort* operation to reverse MATLAB's *sort*) and their true, rapid variation will become apparent. Of course, the nature of the sorting will affect the estimation and it seems likely that there will be an optimum sorting which results in minimum variance estimates. However, such optimum sorting will naturally depend upon the nature of the state dependency and its definition would require some sort of iterative estimation procedure. In practical terms, therefore, the common ascending order sorting and un-sorting operations seem the most straightforward and will be utilized here.

One obvious requirement with this new approach to SDP estimation is that the sorting of data, prior to FIS estimation, must be *common to all of the variables in the relationship* (11a). If an ascending order strategy is selected, therefore, it is necessary to decide upon which variable in the model the sorting should be based. The simplest strategy is to sort according to the ascending order of the 'dependent' variable y_t. Depending upon the nature of each SDP in the vector $\mathbf{p}_t$, however, a single variable sorting strategy, such of this, may not produce satisfactory results. If this is the case, then a more complicated, but still straightforward, 'backfitting' procedure can be exploited. Here, each parameter is estimated *in turn*, based on the 'modified dependent variable' series obtained by subtracting all the other terms on the right hand side of (11a) from y_t. At each such backfitting iteration, the sorting can then be based on the single variable associated with the current SDP being estimated.

Since the SDP estimates resulting from this backfitting algorithm are themselves time series, it will be noted that the algorithm constitutes a special form of non-parametric estimation and, as such, can be compared with other

non-parametric methods, such as the Generalized Additive Modelling (GAM) approach of Hastie and Tibshirani (1996). However, in both conceptual and algorithmic terms, the SDP approach described here is significantly different from this earlier approach and seems more appropriate to the estimation of nonlinear, stochastic, dynamic models. Moreover, the recursive methodology, on which SDP estimation is based, is couched in optimal maximum likelihood terms that seem more elegant and flexible than the scatter-plot smoothing procedures used by Hastie, Tibshirani and others.

As previously, let $\mathbf{p}_t = [p_{1,t}\ p_{2,t}\ \ldots\ p_{n+m+1,t}]^T$, with the $p_{i,t}$, $i = 1, 2, \ldots, n+m+1$, relating to the TF model parameters $a_{i,t}$ and $b_{j,t}$ through (11b).The backfitting algorithm for the SDP model (11) then takes the following form:

Backfitting Algorithm for SDP Models

1. Assume that FIS estimation has yielded prior TVP estimates $\hat{p}^0_{i,t|N}, i = 1, 2, \cdots, m+n+1$ of the SDPs. [6]

2. Iterate: $i = 1, 2, \cdots, m+n+1; k = 1, 2, \cdots, k_c$

   (i) form the modified dependent variable $y^i_t = y_t - \sum_{j \neq i} z_{j,t}.\hat{p}^k_{j,t|N}$

   (ii) sort[7] both y^i_t and $z_{i,t}$ according to the ascending order of $z_{i,t}$

   (iii) obtain an FIS estimate $\hat{p}^k_{i,t|N}$ of $p_{i,t}$ in the modified dependent variable relationship $y^i_t = p_{i,t}.z_{i,t}$

3. Continue 2. (each time forming the modified dependent variable and then sorting according to the current right hand side variable $z_{i,t}$, prior to FIS estimation), until iteration k_c, when the individual SDPs (which are each time-series of length N) have not changed significantly according to some chosen criterion. The smoothing hyper-parameters required for FIS estimation at each stage are optimized by Maximum Likelihood (ML), as explained earlier in section 2.1 and discussed further below.

Note that the ML optimization can be carried out in various ways: after every complete iteration (each involving $m+n+1$ FIS operations) until convergence is achieved; only at the initial complete iteration, with the hyper-parameters maintained at these values for the rest of the backfitting; or just on the first two iterations. The latter choice seems most satisfactory in practice, since very little improvement in convergence occurs if optimization is continued after this stage. Normally, convergence is completed after only a few iterations, although it can be more lengthy in some circumstances (see later discussion in the Conclusions, section 5).

[6] As a default, these can be simply the constant least squares parameter estimates, since the convergence of the backfitting procedure is not too sensitive to the prior estimates, provided they are reasonable: see simulation example 2 below

[7] depending on the nature of the state dependency, sorting may need to be with respect to another variable in $\boldsymbol{\chi}_t$

Simulation Example 2

As a simulation example of SDARX modelling, consider the following forced logistic growth equation,

$$y_t = 2.0y_{t-1} - 2.0y_{t-1}^2 + u_t + e_t \qquad u_t = \mathrm{N}(0, 0.08) \quad e_t = \mathrm{N}(0, 0.008) \tag{12a}$$

or,

$$y_t = a_1(y_{t-1}).\, y_{t-1} + b_0(u_t).\, u_t + e_t \tag{12b}$$

where,

$$a_1(y_{t-1}) = 2.0 - 2.0y_{t-1} \qquad b_0(u_t) = 1.0 \quad \forall t, \tag{12c}$$

Here, u_t is a measured input and e_t is white 'system' noise. The unforced equivalent of this model (i.e. $u_t = 0$ for all t: see simulation example 3) is, in fact, the example used by the author in the first publication on SDP modelling (Young, 1978) although, at that time, the more powerful FIS algorithms described in this chapter had not been developed and simple recursive (filtering) estimation was utilized.

The estimation is based on 1000 samples of u_t and y_t, a 100 sample segment of which (for clarity) is shown in Figure 4. This is a fairly low noise situation with about 10% noise level (by standard deviation). ML optimization at the second iteration yields an NVR matrix $\mathbf{Q}_r = \mathrm{diag}[0.000357\ 2.5 \times 10^{-17}]$, showing how the optimization has, once again, identified that the parameter $b_0(u_t)$ associated with the input u_t is constant and that only the 'lag' parameter $a_1(y_{t-1})$ is varying. This value of $\mathbf{Q}_r$ is then utilized for the subsequent 3 iterations of the backfitting procedure that leads to convergence of the SDP estimates.

The results obtained in this manner are shown in Figures 5 and 6. Fig. 5 provides plots of the estimated parameter variation against the associated non-minimal state variable, in each case. The left hand panel shows the estimated relationship between $\hat{a}_{1,t|N} = \hat{a}_1(y_{t-1} \mid N)$ and y_{t-1}; while the right hand panel shows the estimated $\hat{b}_{0,t|N} = \hat{b}_0(u_t \mid N)$ as a function of u_t. As expected, these correspond to those used in the simulated model: $\hat{a}_1(y_{t-1} \mid N)$ is a clear linear function of y_{t-1} with slope and intercept close to –2 and 2, respectively, in correspondence with equation (12c); while $\hat{b}_0(u_t \mid N)$ is constant for all t, at an estimated value close to unity. The actual values of the parameters are shown as dashed lines and the standard error bounds are shown dotted. Figure 6 presents the associated estimates of the nonlinear functions: $\hat{f}_1(y_{t-1}) = \hat{a}_1(y_{t-1} \mid N).\, y_{t-1}$ and $\hat{f}_2(u_t) = \hat{b}_0(u_t \mid N).\, u_t$, again with actual values shown dashed and the standard error bounds show dotted.

Figure 7 shows results similar to those in Figure 5, again based on a 1000 sample data set, but with a much higher noise level of 69% by standard deviation (48% by variance). Although the uncertainty on the estimates is greater, as would be expected, the general nature of the state dependent relationships is

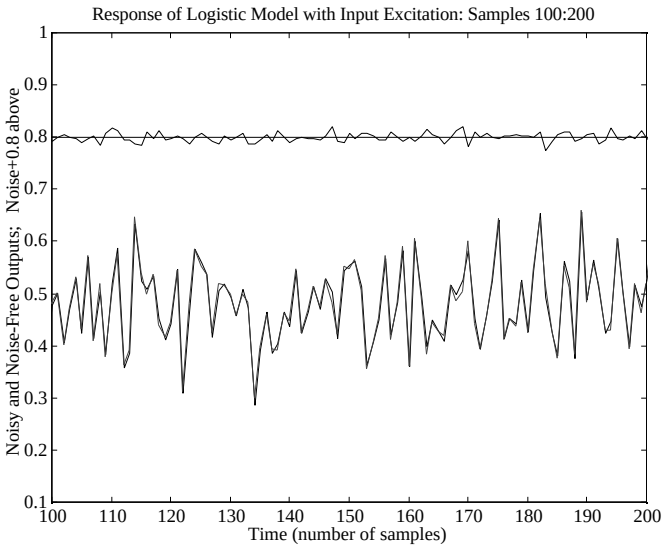

Figure 4. Simulated forced logistic growth model. Noise-free output (full line); noisy output (dashed line); and noise (+0.8) above

still clear and the identification of a forced logistic growth equation is quite unambiguous. In this noisy situation, the NVR matrix $\mathbf{Q}_r$= diag[9.89×10^{-7} 0.0]

3.2 SDTF Estimation

In the SDTF model, it is assumed that noise can enter as either system or measurement noise, or both. For example, in the case of the forced logistic equation (12), it could take the form:

$$\begin{array}{ll} x_t = 2.0x_{t-1} - 2.0x_{t-1}^2 + u_t + e_t & u_t = \mathrm{N}(0, 0.08) \quad e_t = \mathrm{N}(0, 0.008) \\ y_t = x_t + \xi_t & \xi_t = \mathrm{N}(0, 0.08) \end{array} \tag{13a}$$

or,

$$y_t = a_1(y_{t-1}).\, y_{t-1} + b_0(u_t).\, u_t + \zeta_t \tag{13b}$$

where,

$$a_1(y_{t-1}) = 2.0 - 2.0y_{t-1} \qquad b_0(u_t) = 1.0 \quad \forall t, \tag{13c}$$

and the noise ζ_t is a complex nonlinear function of e_t, ξ_t and y_t. In this situation, estimates obtained under the assumption that the model is of the simpler SDARX form are nominally biased to a level dependent on the noise/signal ratio. Fortunately, however, this bias is often fairly small, even for quite high

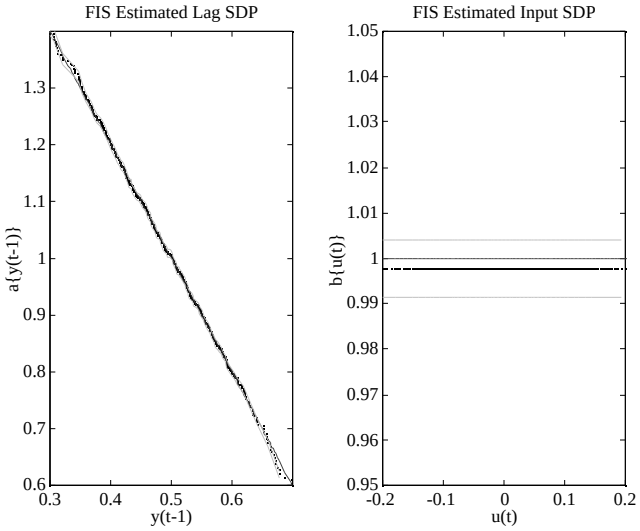

Figure 5. Simulated forced logistic growth model. Left panel: FIS estimate $\hat{a}_1(y_{t-1}|N)$ vs y_{t-1}. Right panel: FIS estimate $\hat{b}_o(u_t|N)$ vs u_t. True values shown as dashed lines and standard error bands shown as dotted lines.

noise levels and, in consequence, it does not interfere substantially with the identification of any state dependent relationships (note that the analysis is aimed at *identifying* the form of the nonlinearity and more efficient statistical estimation follows this identification step). For example, in the above example, the estimation results for a measurement noise level similar to that of the system noise level used to obtain the results in Figure 7 (69% by standard deviation; 48% by variance) are quite acceptable and only a little worse than those shown in Figure 7.

Nevertheless, it would be advantageous if a truly bias-free estimation method was available in the SDTF model case and research is continuing on the development of an IV backfitting algorithm which exploits the methodology discussed above. The main problem with such an approach is in maintaining the stability of the auxiliary model (see equation (5c)) that generates the instrumental variables: in the case of nonlinear models with chaotic properties, for example, only small uncertainties can lead to wide differences in response and possible instability. Consequently, other approaches that extend the model to include (SDP) noise terms are also being investigated.

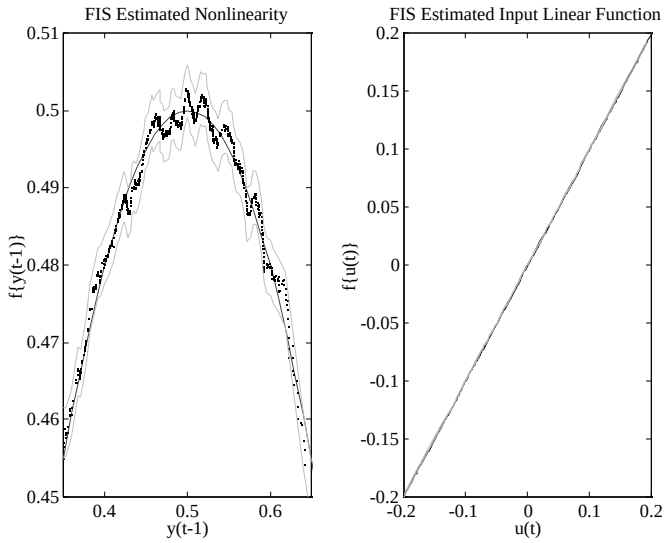

Figure 6. Simulated forced logistic growth model. Left panel: FIS estimate of feedback nonlinearity $\hat{f}_1(y_{t-1})$. Right panel: FIS estimate of input nonlinearity $\hat{f}_2(u_t)$. True values shown as dashed lines and standard error bands shown as dotted lines.

Final Parametric Estimation

If any underlying parametric state dependency can be identified in non-parametric form using the recursive estimation methods discussed in previous sections, then it is often possible to continue further and parameterize the identified nonlinear relationships in terms of a finite set of constant parameters. For example, given the non-parametric estimation results shown in Figures 5, 6 and 7, it is clear that the forced logistic growth equation, with three unknown parameters, provides a parsimonious representation of the data. It is straightforward, therefore, to obtain estimates of these parameters by simply fitting appropriate parameterized functions to the estimated state dependent relationships using weighted or ordinary least squares (WLS: see Young, 1993a, 1998a; Young and Beven, 1994). In this case, the WLS estimates of the two parameters of the linear relationship in (12c) obtained from the low noise results (Figure 5) are: 2.002(0.0013) and $-2.004(0.0025)$. The corresponding estimates in the high noise case (Figure 7) are: 2.039(0.008) and $-2.08(0.14)$; and similar results are obtained in the SDTF case. However, it is more satisfactory to use ML estimation of the parameters in (12c) directly from the data. In the

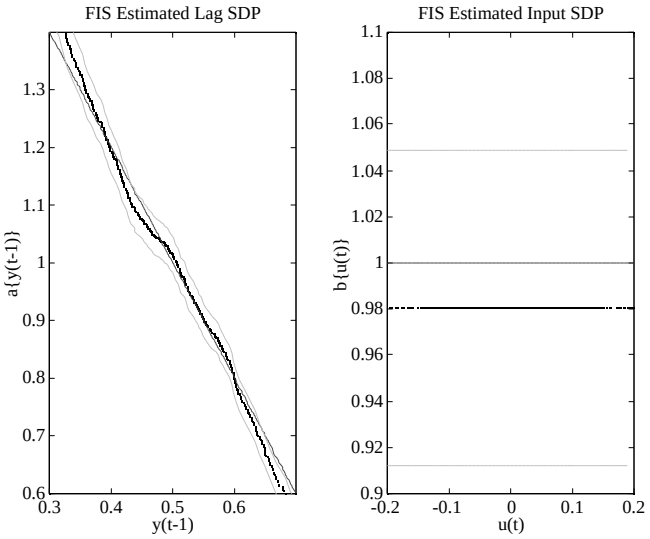

Figure 7. Simulated forced logistic growth model. As for Figure 5 but with higher level of noise on the data

low noise case, this yields 2.0332(0.008), −2.0679(0.017) and 0.9813(0.008); while in the high noise case, the estimates are 2.0375(0.036), −2.1485(0.075) and 0.9834(0.033). So, in all cases, the estimates are very good (although with signs of very small positive bias) and the finally estimated model produces responses that are insignificantly different from noise free output of the actual system, even in the high noise case. This is shown in Figure 8, where the error shown above (+0.8) is between the model output and the noise-free output. The noisy output on which the estimation is based is shown with circle points.

3.3 Estimation of Purely Stochastic Systems

A special example of the SDP model (11) is the following *State Dependent parameter Auto-Regressive* (SDAR) model:

$$y_t = -a_{1,t}y_{t-1} - a_{2,t}y_{t-2} - \cdots - a_{n,t}y_{t-n} + e_t \tag{14a}$$

or

$$y_t = \mathbf{z}_t^T \mathbf{p}_t + e_t, \tag{14b}$$

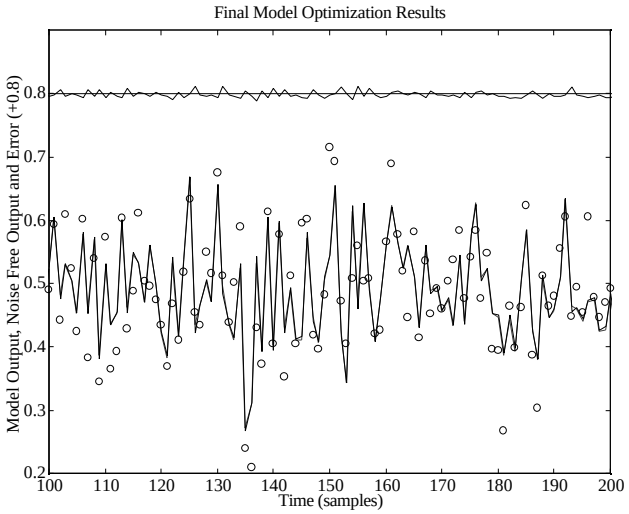

Figure 8 .Simulated forced logistic growth model: final parametric estimation results. Comparison of estimated model output (full line) and noise free output (dashed), with the error (+0.8) shown above. The circle points show the noisy measured output used in the estimation.

where

$$\begin{aligned} \mathbf{z}_t^T &= [-y_{t-1} \ -y_{t-2} \ \cdots \ -y_{t-n}] \\ \mathbf{p}_t &= [a_1(\boldsymbol{\chi}_t) \ a_2(\boldsymbol{\chi}_t) \ \cdots \ a_n(\boldsymbol{\chi}_t)]^T, \end{aligned} \tag{14c}$$

Clearly, the same SDARX estimation methods discussed previously can be applied to this model, a simple example of which is the chaotic version of the logistic equation. Typical simulation results for this model are discussed below.

Simulation Example 3

In order to consider the effects of measurement noise, the following Noisy SDAR (NSDAR) version of the chaotic logistic growth model (cf. equations (12)) will be used in this example:

$$\begin{aligned} x_t &= 4.0x_{t-1} - 4.0x_{t-1}^2 + e_t \quad & e_t &= \mathrm{N}(0, 0.0064) \\ y_t &= x_t + \xi_t & \xi_t &= \mathrm{N}(0, 0.0012) \end{aligned} \tag{15a}$$

or

$$y_t = a_1(y_{t-1}).\, y_{t-1} + \zeta_t \qquad a_1(y_{t-1}) = 4.0 - 4.0y_{t-1}. \tag{15b}$$

As in equation (13b), the noise ζ_t is a complex nonlinear function of e_t, ξ_t and y_t, and the noise level is about 10% (by standard deviation). As expected, without this measurement noise present, SDP estimation is straightforward with excellent, low variance SDP estimates that identify the nature of the nonlinearity without any difficulty.

Even with the measurement noise, this is a simpler estimation problem than in the forced logistic growth model since there is only one SDP, $a_1(y_{t-1})$, and ML optimization of the associated scalar NVR quickly yields $\mathbf{Q}_r = 0.000016$. The subsequent SDP estimation results are illustrated in Figure 9, where the top left hand panel shows the FIS estimated nonlinear function with its characteristic quadratic shape; while the top right hand panel shows the estimated state dependency of the SDP estimate $\hat{a}_1(y_{t-1} \mid N)$. In both plots, the true relationship is shown as a dashed line, while the estimated standard error bounds are shown dotted.

The most notable feature of the results in Figure 9 is the larger errors in the SDP estimate at low values of y_{t-1}. Since the measurement noise variance is constant, this is the region where the measurement noise is having its most deleterious biasing effect. Even with these errors, however, the WLS estimates of the parameters (3.884(0.111) and -3.857(0.135), respectively) are quite close to the true values (4 and -4). The results are better still, however, if the noise is itself made state dependent, in the sense that it is set proportional to the signal level (i.e. $\xi_t^{\mathrm{sd}} = [x_t/\max(x_t)].\,\xi_t$). This is a common situation with real data and it significantly reduces the noise effect at low signal levels where the bias is largest. As a result, the WLS estimates are improved to 3.960(0.013) and -3.946(0.016), respectively. Indeed, in this state dependent noise situation, the results are still good even if the variance of ξ_t is doubled, so that the noise level on the data is visibly quite large, as shown in Figure 10. The resulting SDP estimation results are presented in the lower panels of Figure 9, which identify clearly that the data were generated by the logistic model with chaos inducing parameter values.

4 Further Examples

In this section, we present two further examples. The first is another, more difficult simulation example based on the so-called *Cosine Map* model. The second is a brief description of a practical example: namely the analysis of the famous set of times series data on the Australian sheep blowfly *Lucilia Cuprina* by Nicholson (1954). This example is described in more detail elsewhere (Young, 1998a; Young and Fawcett, 1999). Other, practical examples, in a variety of application areas from the environment to economics, are described in Young (1993a, 1996, 1998a,b, 1999a; Young and Beven, 1994; Young and Pedregal, 1997, 1999; Young *et al*, 1996).

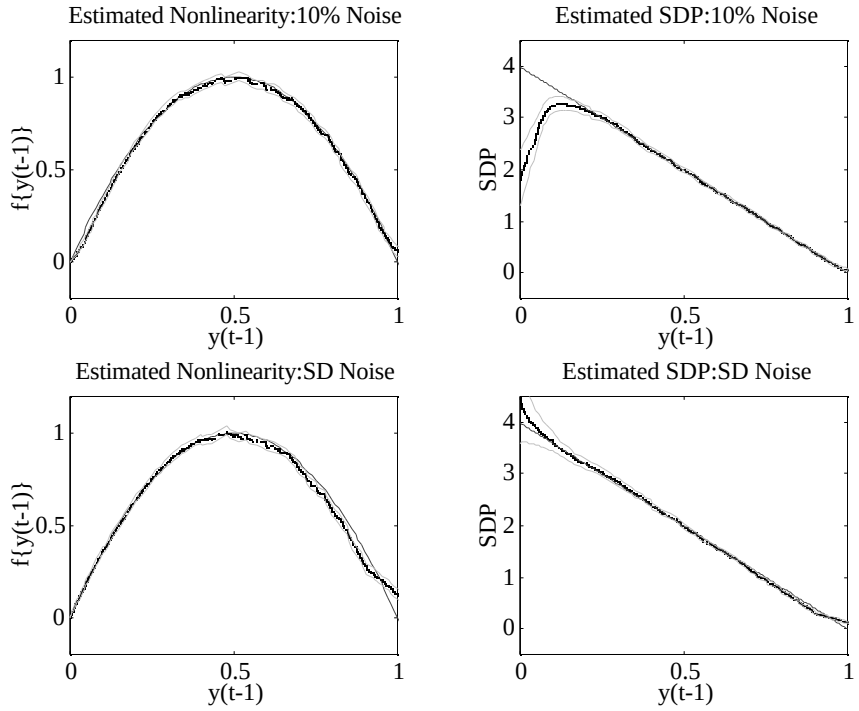

Figure 9. Simulated chaotic logistic model. Left panels: estimated nonlinear functions with ordinary (upper) and state-dependent (lower) noise. Right panels: estimated state dependent parameters with ordinary (upper) and state-dependent (lower) noise

4.1 Identification and Estimation of the Cosine Map Model

The cosine map model (e.g. Zhan-Qian and Smith, 1998) takes the form

$$y_t = \cos(2.8y_{t-1}) + 0.3y_{t-2} + e_t \qquad e_t = \mathrm{N}(0, 0.01), \tag{16}$$

and a typical 2000 sample simulation of the model is presented in Figure 11, which shows the time response in the upper panel and the phase plane ($y_t \sim y_{t-1}$) plot in the lower panel.

In the latter graph, the noise free response is shown as a full line with the noisy response plotted as dots. This a typical stochastic model that exhibits underlying chaotic response characteristics. It provides a testing example for the SDP approach, however, because it is not in the assumed affine form: in

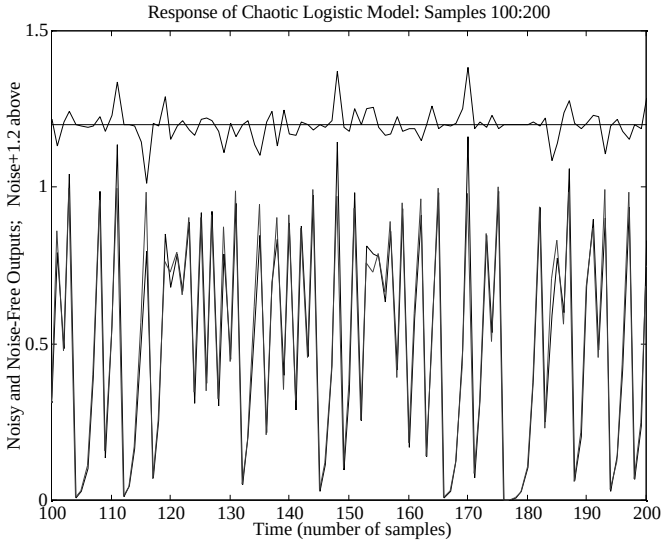

Figure 10. Simulated chaotic logistic model. Noisy (full line) and noise-free (dashed line) output, with the state-dependent noise (+1.2) shown above

particular, the SDP term $a_1(y_{t-1}).\, y_{t-1}$ in the most appropriate SDAR model,

$$y_t = a_1(y_{t-1}).\, y_{t-1} + a_2(y_{t-2}).\, y_{t-2} + e_t \qquad e_t = \mathrm{N}(0, \sigma^2), \tag{17}$$

is not able to represent the equivalent term $\cos(2.8y_{t-1})$ in (16) exactly, since $\cos(2.8y_{t-1})/y_{t-1}$ has a singularity at $y_{t-1} = 0$. Despite this difficulty, SDAR estimation yields excellent results, as illustrated in Figures 12 to 14. These were obtained with the ML optimized NVR matrix $\mathbf{Q}_r = \mathrm{diag}[0.0057\ 1.28 \times 10^{-7}]$ and it is clear, yet again, that the optimization has successfully identified that the potential state dependency resides in the first lag parameter $a_{1,t} = a_1(y_{t-1})$, while the second lag parameter $a_{2,t} = a_2(y_{t-2})$ is effectively time-invariant.

Figure 12 shows the FIS estimate of the cosine nonlinearity subsequent to the convergence of the backfitting procedure, which took 6 iterations in this case. Except for the region around the singularity at $y_{t-1} = 0$ (see below), the estimation is very good. The associated $\hat{a}_2(y_{t-2} \mid N) = 0.291(0.0038)$, for all t, is estimated as being time-invariant despite the fact that the NVR, $\mathbf{Q}_r(2,2) = 1.28 \times 10^{-7}$, is not too small in this case. Figure 13 compares the actual, simulated phase plane plot for the data used in the estimation (left hand panel) with a similar plot based on data from a typical random

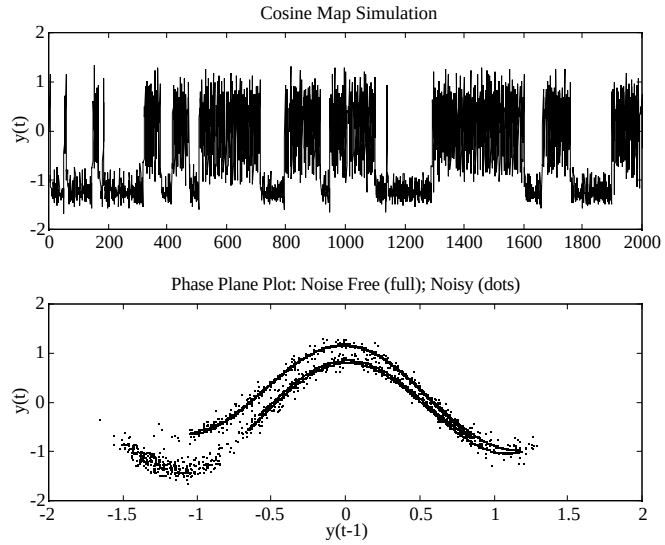

Figure 11. Simulated cosine map model. Upper panel: measured output. Lower panel: phase plane (embedding) plot of y_t vs $y_{t-1}$with noise-free response shown as a full line

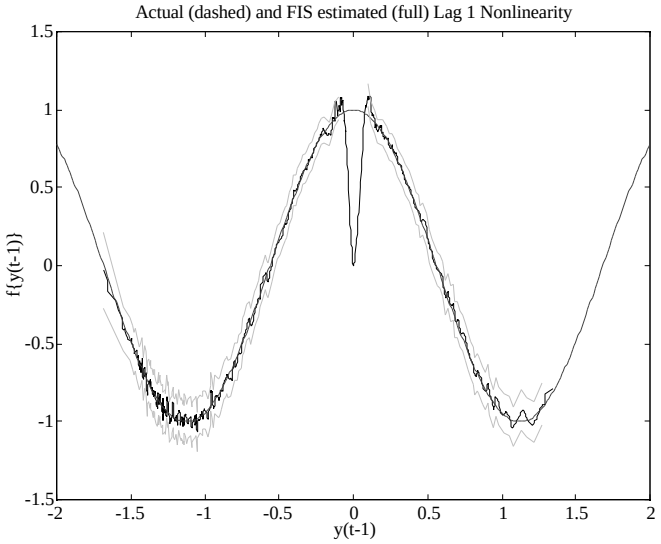

Figure 12. Simulated cosine map model. Comparison of FIS estimated and actual cosine nonlinearity

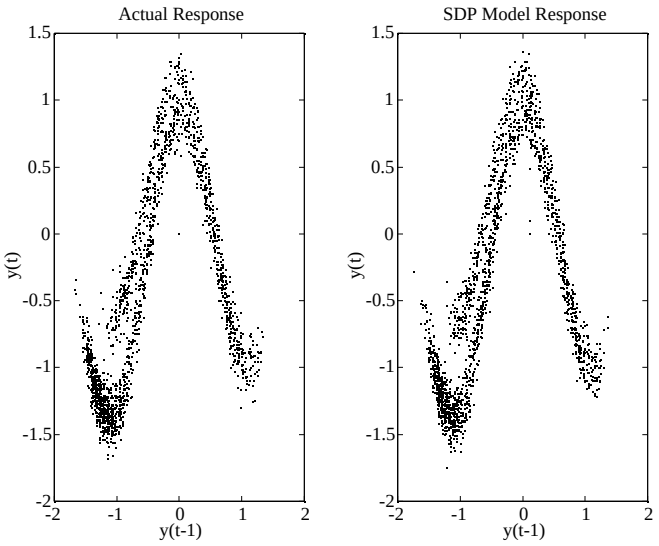

Figure 13. Simulated cosine map model. Comparison of the phase-plane plot for the data used in the estimation (left panel) with the phase-plane obtained from a random realization of the estimated SDAR model

realization of the SDAR model (right hand panel). Similar agreement is found in both the time plots of the two series and the histograms.

Finally, Figure 14 compares the FIS estimate $\hat{a}_1(y_{t-1} \mid N)$, plotted as a function of y_{t-1}, with the theoretical function given by $a_1(y_{t-1}) = \cos(2.8y_{t-1})/y_{t-1}$ (which migrates to $\pm\infty$ at the point $y_{t-1} = 0$). Over the most important region $-0.1 > y_{t-1} > 0.1$, the estimate is very accurate. Not surprisingly, it becomes inaccurate close to the singularity, but the algorithm is robust enough to handle this well, without impairing the SDP estimates elsewhere. And the overall cosine shape of the nonlinear function is clearly estimated accurately in Figure 12.

Finally, on the basis of the above SDP results, the form of the nonlinear equation is identified correctly as:

$$y_t = \alpha \cos(\beta y_{t-1}) + \gamma y_{t-2} + e_t, \tag{18}$$

and the optimized ML estimates of the, now constant, parameters α, β and γ in this model are $\hat{\alpha} = 0.998(0.004)$, $\hat{\beta} = 2.797(0.004)$ and $\hat{\gamma} = 0.303(0.003)$.

4.2 The Nicholson Blow y Data Revisited

Figure 15 is the best known example of the data collected so laboriously by Nicholson (1954) in his investigation of the Australian sheep blowfly *Lucilia*

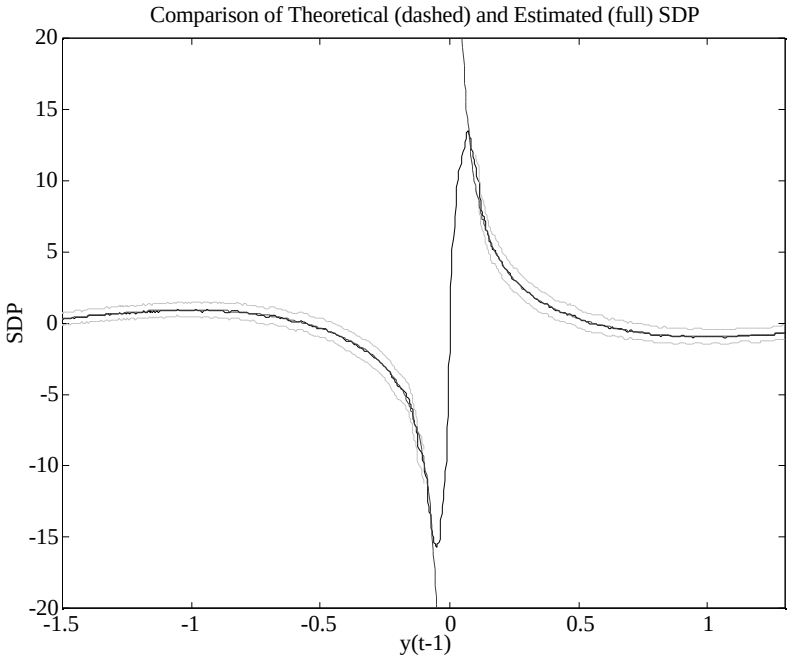

Figure 14. Simulated cosine map model. Comparison of the FIS estimate $\hat{a}_1(y_{t-1} \mid N)$ as a function of y_{t-1}, with the theoretical function given by $a_1(y_{t-1}) = \cos(2.8y_{t-1})/y_{t-1}$ (which migrates to $\pm\infty$ at the point $y_{t-1} = 0$).

Cuprina. It is clear that, in this particular experiment where the food (liver) supplied to the blowflies was limited to 0.5g per day, the adult blowfly population y_t (upper graph) and the eggs laid per day by the blowflies u_t (lower graph) vary in an apparently systematic fashion that is redolent of nonlinear, limit cycle behaviour. Not surprisingly, these data have received much attention in the scientific literature (e.g. May, 1973, 1976; Banks, 1994; Gurney *et al.*, 1980; and the references therein).

In contrast to the analysis described in the previous references, however, the SDP modelling approach used here makes no *a priori* assumptions about the nature of the blowfly system but starts with a relatively non-prejudicial analysis of the data, recognising only that the eggs and blowfly series are causally related in some manner. The data in Figure 15 are sampled daily[8] and the most obvious relationship between u_t and y_t, namely the blowfly response to the egg production rate, seems quite linear and can be described

[8]The data shown in Figure 15 have been digitized at a daily sampling interval from the graphical plots in Nicholson's paper. The original data, which are no longer available (McNeil, 1996), were collected every two days.

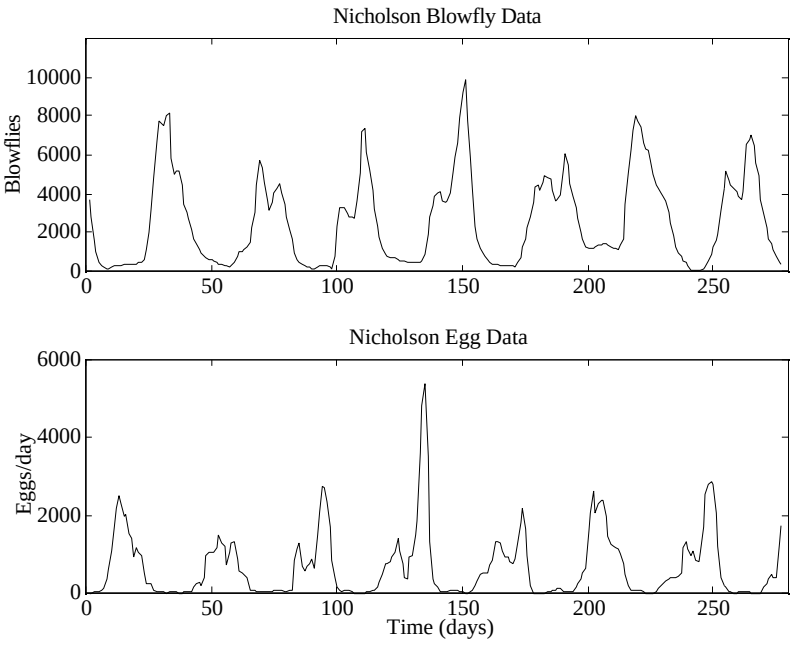

Figure 15. The Nicholson blowfly data example. Upper panel: daily numbers of adult blowflies. Lower panel: numbers of eggs laid per day.

well by the following first order, *constant* parameter, discrete-time TF:

$$y_t = \frac{b_0}{1 + a_1 z^{-1}} u_{t-\delta} + \xi_t \tag{19a}$$

or

$$y_t = a_1 y_{t-1} + b_0 u_{t-\delta} + \eta_t, \tag{19b}$$

where the identified time delay of $\delta = 15$ days accounts for the development of the eggs through a larval stage, prior to the emergence of the adult blowflies. The estimates of the parameters and their standard errors are obtained by constant parameter SRIV estimation (e.g. Young, 1984) as $\hat{b_0} = 0.865(0.031)$ and $\hat{a}_1 = 0.759(0.01)$. This simple linear model explains the experimental data very well with $R_T^2 = 0.86$ and the model residuals, with ξ_t modelled as a constant parameter, AR(3) process, satisfy the usual statistical diagnostic tests.

But the modelling story does not end with this linear model of the 'forward-path' dynamics: of much more interest is the nature of the 'feedback-path' dynamics, namely the mechanism by which the blowflies produce their eggs.

In contrast to the forward-path, this mechanism is obviously nonlinear, with the blowflies producing eggs only when their numbers are low and the food supply per blowfly is plentiful. There are two obvious ways in which the data can be analyzed to infer the nature of these nonlinear feedback dynamics. Most straightforwardly, the model (19) can be modified directly to acknowledge that the egg production rate u_t is a nonlinear function of the blowfly population y_t: this then yields a SDAR-type model with lag terms in y_t and $y_{t-\delta}$.

Alternatively, *since the time delay in the system is large*, it is possible to investigate the feedback-path dynamics separately by considering directly the relationship between y_t, now considered as the 'input' in the feedback path, and u_t, as the output, i.e.,

$$u_t = b_0(y_t).\, y_t + e_t. \tag{20}$$

In other words, we are able to investigate the presence of a simple *static* nonlinearity in the feedback path that has direct ecological significance. SDP estimation yields the estimate $\hat{b}_0(y_t \mid N)$ shown in Figure 16 as a function of y_t, with the standard error bounds shown dotted. Also shown as a dash-dot line is the WLS estimate of the nonlinearity based on the following exponential-type parameterization:

$$u_t = g.\, y_t.\, f(y_t) + e_t \qquad f(y_t) = \exp\left(\frac{-1}{N_0 \frac{fd}{y_t}}\right) \qquad e_t = \mathrm{N}(0, \sigma^2), \tag{21}$$

where fd is the food supplied per day to the blowfly colony (here 0.5g); while g and N_0 are unknown parameters with WLS estimates of $\hat{g} = 4.916(0.26)$ and $N_0 = 1451(94)$, respectively. The results in Figure 16 were obtained with an optimized NVR $= 8.16 \times 10^{-9}$.

Other parameterizations than (21) are clearly possible (e.g. Young, 1998a; Young and Fawcett, 1999) but this particular one was chosen here because it conforms to the prior analysis of Gurney *et al.* (1980). In particular, the present model, i.e.,

$$y_t = a_1 y_{t-1} + g.\, y_{t-\delta}.\exp\left(\frac{-1}{N_0 \frac{fd}{y_t}}\right) + \mu_t, \tag{22}$$

where μ_t represents the overall residual coloured noise, is closely related to the *deterministic, continuous-time differential equation* suggested by Gurney *et al.* However, their analysis is deterministic and only semi-quantitative, in the sense that they do not use statistical identification and estimation at all. Rather, they simply speculate on the form of this nonlinear system and show that, for a range of parameter values, its *general*, deterministic dynamic behaviour conforms reasonably with that exhibited by the Nicholson data.

Having identified a suitable nonlinear model for the blowfly data in (22), it is now possible to move on to the final ML estimation stage in the analysis. In

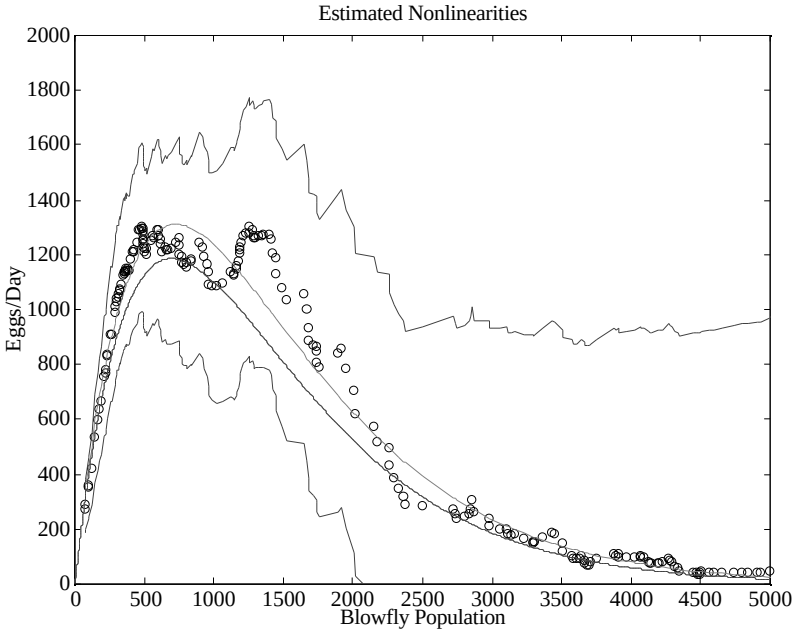

Figure 16. The Nicholson blowfly data example. FIS estimated feedback nonlinearity (circle points) compared with the weighted least squares estimate (dash-dot line) and final ML estimate (dashed line) of the parameterized nonlinear function. The standard error bands on the FIS estimate are shown as dashed lines

this case, the nature of the coloured noise on the data requires that the model (22) is enhanced to include stochastic elements (Young and Fawcett, 1999). ML optimization of the resulting stochastic model yields the following result:

$$\begin{aligned} x_t &= 0.818x_{t-1} + u_{t-15} + \xi_t \\ u_t &= 4.63x_t.\exp\left(\frac{-1}{1392\frac{0.5}{x_t}}\right) \\ \xi_t &= 1.137\xi_{t-1} - 0.491\xi_{t-2} + \varepsilon_t \\ y_t &= x_t + e_t \qquad e_t = \mathrm{N}(0,\sigma^2), \end{aligned} \tag{23}$$

where $\sigma^2 = 1.52 \times 10^5$ and $\varepsilon_t = \mathrm{N}(0, 44.5\sigma^2)$. In this ML optimization stage, the model (23) is formulated as the associated 17th order, discrete-time, stochastic state equation (15 orders account for the pure 'larval' time delay and 2 represent the noise dynamics). The optimization is then performed using prediction error decomposition based on a state dependent parameter

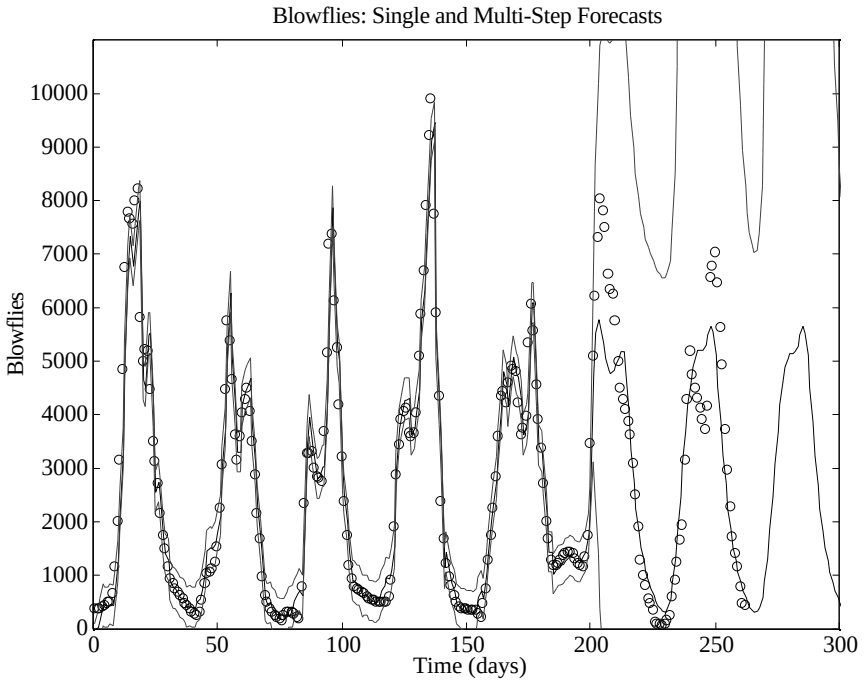

Figure 17. The Nicholson blowfly data example. Kalman filter forecasting results: forecast (full line) compared with the data (circle points), with the standard error band shown as dashed lines. Up to day 200, the forecasts are one-step-ahead; after this, the forecast is a true *ex-ante* multi-step-ahead forecast, based only on the first 200 days

implementation of the Kalman filter *that does not require linearization* (Young and Fawcett, 1999). The nonlinearity associated with this model is shown as the dashed line in Figure 16 and we see that it is similar to the WLS estimate from the non-parametric estimation stage in the identification analysis, laying well within the standard error bounds.

The residuals of the model (23) have satisfactory correlation properties but they are very heteroscedastic (clearly dependent on the blowfly population). Also, the Hessian associated with the parameter and hyper-parameter estimates suggests that some of the parameters have quite large standard errors. Despite this, the model performs well in forecasting and validation tests. Figure 17, for example, provides typical forecasting results, with one step-ahead forecasts up to the 200th day and true, multi-step *ex-ante* forecasts thereafter.

The explicit inclusion of fd in (22) and (23) is useful because it allows us

to evaluate the ecological realism of the model still further by examining its prediction of what should happen if the food supply is modified. In particular, Nicholson (1954) showed experimentally that, when the food supply was reduced from 0.5g per day to 0.1g per day, the average adult population dropped from 2520 to 527. In the case of the model (23) the average populations in these two same situations are 2657 and 548 respectively; a remarkable level of agreement in the circumstances (since the average values here are based on the deterministic limit cycle data produced by (23) when the stochastic inputs are removed, rather than the actual data). Certainly this result provides a suitable initial validation of the model.

These results tend to confirm and further quantify the deterministic analysis of Gurney *et al.*. However, the statistical diagnostics and associated stochastic simulations suggest that the stochastic model (23) requires a little further work, taking into account the heteroscedasticity and the poor definition of the optimized parameters, before a fully satisfactory stochastic model is confirmed for the blowfly data. Such future research could be based, for instance, on Markov Chain Monte Carlo (MCMC) methods (e.g. Ruanaidh and Fitzgerald, 1996; Gamerman, 1997) or on a simpler stochastic approach, such as that suggested recently by Durbin and Koopmans (1999), where the assumption of Gaussian disturbances is not necessary. Nevertheless, as it stands, the model (23) is clearly a reasonable one in predictive terms, as shown by the results in Figure 17. And the fact that this forecasting performance is good, despite the limitations of the stochastic model, is testament to the well known robustness of the Kalman filter when it is used as a basis for forecasting.

5 Conclusions

Recursive estimation has a long and rich history: from its beginnings in Gauss's original derivation of recursive least squares (Gauss, 1823; see Appendix 2 of Young, 1984), through its re-discovery by Plackett (1950) and Kalman's seminal work on stochastic state estimation (Kalman, 1960), to the burgeoning of research on recursive estimation in a whole range of different academic disciplines between 1960 and the present. In the last ten years, however, the advent of fast computers and the desire of theorists to extend the boundaries of time series analysis has led to an explosion of research on Monte Carlo-based numerical methods, from either classical (e.g. Durbin and Koopman, 1999) or Bayesian (e.g. Ruanaidh and Fitzgerald, 1996; Gamerman, 1997) perspectives.

The motivation of this more recent research is clearly to extend the 'Gaussian' methods of standard recursive estimation to non-Gaussian and nonlinear time series, using models in which the *stochastic inputs* are non-Gaussian. But models of non-Gaussian and nonlinear processes do not necessarily require the assumption of non-Gaussian inputs. As we have seen in this chapter, a

fairly wide class of non-Gaussian and nonlinear time series can be represented by time variable (TVP) and state dependent (SDP) parameter, nonlinear, stochastic models with *Gaussian* inputs. When it is possible (and the methods do seem quite widely applicable), this is clearly advantageous, since it allows for the use well tried and robust algorithms that are computationally much less demanding than even the 'classical' non-Gaussian methods (see Young, 1999c; Durbin and Koopman, 1999).

As far as the author is aware, the idea of SDP modelling originated in his 1978 paper on the modelling ofr badly defined dynamic systems (Young, 1978), and was then taken up by Priestley in a series of papers and a book on the subject (Priestley, 1988). These earlier publications do not, however, exploit the power of recursive fixed interval smoothing (FIS), which provides the main engine for the developments described in this Chapter. The combined Kalman filter (KF) and FIS (or FISIV: see Section 2) algorithms, as used here, clearly have a more powerful potential than their more conventional usage would suggest. Moreover, the fact that these same algorithms can function with system matrices characterized by time, or even state dependent parameters, extends their range of applicability to a considerable extent. Thus, the non-parametric models described in this chapter can provide the basis for rather novel non-parametric or state dependent parameter KF–FIS design, with implications for both modelling and nonlinear optimization based on prediction error decomposition.

The fact that the FIS algorithm can function well as a non-parametric estimator means that it provides a powerful, recursive alternative to other, more conventional, methods of smoothing, such as regularization, smoothing splines, kernel smoothers and locally weighted kernel regression (see Young and Pedregal, 1998). It also provides a non-parametric method for transforming random variables (e.g. Gaussian to non-Gaussian), or identifying the nature of a parametric transform between random variables.

Finally, it is clear that the simulated and real examples presented in the chapter, combined with those discussed in other cited references, demonstrate the efficacy of the proposed SDP approach to modelling for a fairly wide and practically useful class of nonlinear stochastic systems. However, the proposed technique is relatively new and it raises a variety of interesting theoretical questions and possibilities for extending the approach to an even richer class of nonlinear stochastic systems. For example:

- How can the approach be extended to handle multivariable state dependencies, where the SDPs may be functions of several state variables?
- What is the best method of handling the *errors-in-variables* problem and the estimation bias that occurs when the proposed SDP modelling approach is applied to *errors-in-variables* TF models? An Instrumental variable (IV) method, such as that used successfully in the case of DTF

models (section 2), has been devised to handle this problem in the case of well-behaved nonlinear models. But alternative approaches will be required in the case of sensitive chaotic models.

- Although no convergence problems have been encountered so far in the evaluation of the proposed SDP estimation procedure, what conditions are required for convergence of the backfitting procedure? Hastie and Tibshirani (1996) use a similar backfitting procedure for estimation of their Generalized Additive Model (GAM). It needs to be established whether their conclusions as regards convergence (which are not entirely persuasive, in any case) are applicable to the models and backfitting procedure described in this chapter. Unlike the GAM, for instance, the nonlinear functions in the SDP models are factorized into the product of the SDP and the model variable; and the SDP is estimated by optimal FIS smoothing (rather than the more conventional scatter-plot smoothing used by Hastie and Tibshirani).

- The backfitting procedure does not provide complete covariance information on the SDP estimates. Could this be distorting the standard errors on the estimates (e.g. the standard errors in Figure 5 seem very small)? In more general terms, what are the full *theoretical* statistical properties of the SDP estimates obtained by backfitting?

- Finally, what are the identifiability conditions on the SDP models? It is clear that problems analogous to collinearity in constant parameter model estimation can occur and that backfitting convergence will be affected by such problems. Also, in the case on input-output models, the nature of the input signals will affect the identifiability of the model parameters. It is necessary to explore these factors further and establish what other factors may affect the indentifiability of the model.

Regardless of the answers to these questions, however, the SDP approach to the identification of nonlinearities in stochastic systems appears to hold great promise. In contrast to other approaches, such as neural networks and NARMAX models, for example, it attempts to identify the type of nonlinearity and, therefore, the form of the nonlinear model, prior to the estimation of the parameters in the finally identified model. This helps to ensure that the final nonlinear model is efficiently parameterized (parsimonious) and it should avoid the over-parameterization that normally accompanies neural network and, to a lesser extent, the 'black-box' NARMAX models. Indeed, the SDP approach has been developed as a primary tool in *Data-Based Mechanistic* (DBM) modelling (e.g. Young, 1993a,b; 1998a,b, 1999b; Young and Beven, 1994; Young and Pedregal, 1997; and the prior references therein), where its ability to obtain parametrically efficient and physically meaningful models is essential.

SDP estimation also provides a non-parametric model that can be useful in its own right. As we have seen, the SDP model can be simulated easily in programs such as Simulink, thus removing the need for the final parametric estimation in some applications, such as simulation, forecasting and automatic control. In the latter case, for instance, it is clearly possible to develop state estimation and control system design methods based on this new class of nonlinear models, in parametric or non-parametric form (the latter providing a completely new way of considering control and estimation system design). Research on such developments is continuing and has so far led to encouraging initial results.

Acknowledgments

Part of this research was supported by the EPSRC under grant GR/K77884, and the BBSRC under grant EO6813. The author is also grateful to the EPSRC and the Newton Institute for Mathematical Sciences in Cambridge for support during his stay at the Newton Institute, where most of this chapter was written.

References

Banks, R.B. (1994) *Growth and Diffusion Phenomena*, Springer.

Box, G.E.P., and Jenkins, G.M. (1970) *Time Series Analysis: Forecasting and Control*, (revised edition 1976), Holden-Day.

Brown, R.L., Durbin, J. and Evans, J.M. (1975) 'Techniques for testing the constancy of regression relationships over time', *Jnl. Royal Stat. Soc., Series B* **37** 141-192.

Bryson, A.E. and Ho, Y.C. (1969) *Applied Optimal Control*, Blaisdel.

Co, T. (1996) 'Parameter estimation of nonlinear systems using modulating functions methods', in *Identification in Engineering Systems*, M.I. Friswell and J.E. Mottershead (eds.), University of Wales, Swansea, 87-96.

Daubechies, I. (1988) 'Orthonormal bases of compactly supported wavelets', *Communications on Pure Applied Mathematics* **41** 906-966.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', *Jnl. Royal Stat. Soc., Series B* **39** 1-38.

Durbin, J. and Koopmans, S.J. (1999) 'Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives', *Jnl. Royal Stat. Soc., Series B* **62**, in press.

Gamerman, D. (1997) *Markov Chain Monte Carlo*, Chapman and Hall.

Gauss, K.F. (1821, 1823, 1826) 'Theoria combinationis observationum erroribus minimis obnoxiae', Parts 1, 2 and supplement, *Werke* **4** 1-108.

Gurney, W.S.C., Blythe, P.C. and Nisbet, R.M. (1980). 'Nicholson's blowflies revisited', *Nature* **287** 17-21.

Harvey, A.C. (1981) *Time Series Models*, Phillip Allen, Oxford.

Harvey, A.C. (1984) 'A unified view of statistical forecasting procedures' (with comments), *Jnl. of Forecasting* **3** 245-283.

Harvey, A.C. (1989) *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Harvey, A.C. and Peters, S. (1990) 'Estimation procedures for structural time series models', *Journal of Forecasting* **9** 173-204.

Hastie, T.J. and Tibshirani, R.J. (1996) *Generalized Additive Models*, Chapman and Hall.

Holst, U., Hössjer, O., Björklund, C., Ragnarsson, P. and Edner, H. (1996) 'Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements', *Environmetrics* **7** 410-416.

Jakeman, A.J. and Young, P.C. (1979) 'Recursive filtering and the inversion of ill-posed causal problems', CRES Report No. AS/R28/1979, Centre for Resource and Environmental Studies, Australian National University.

Jakeman, A.J. and Young, P.C. (1984) 'Recursive filtering and the inversion of ill-posed causal problems', *Utilitas Mathematica* **35** 351-376.

Kalman, R.E. (1960) 'A new approach to linear filtering and prediction problems', *ASME Transactions Journal Basic Engineering* **82D** 35-45.

Kitagawa, G. (1981) 'A non-stationary time series model and its fitting by a recursive filter', *Jnl. of Time Series Anal.* **2** 103-116.

May, R.M. (1973) *Stability and Complexity in Model Ecosystems*, Princeton University Press.

May, R.M. (ed.) (1976) *Theoretical Ecology*, Blackwell Science.

McNeil, D. (1996) *Personal Communication*.

Ng, C.N. and Young, P.C. (1990) 'Recursive estimation and forecasting of nonstationary time-series', *Jnl. of Forecasting* **9** 173-204.

Nicholson, A.J. (1954) 'An outline of the dynamics of animal populations', *Australian Zoological Jnl.* **2** 9-65.

Norton, J.P. (1975) 'Optimal smoothing in the identification of linear time-varying systems', *Proceedings Institute Electrical Engineers* **122** 663-668.

Norton, J.P. (1986) *An Introduction to Identification*, Academic Press.

Pedregal, D.J. and Young, P.C. (1996) 'Modulated cycles, a new approach to modelling seasonal/cyclical behaviour in unobserved component models', Centre for Research on Environmental Systems and Statistics (CRES), Tech. Note No. TR/145.

Pedregal, D.J. and Young, P.C. (1998) 'Extensions of trend models in unobserved component models', Centre for Research on Environmental Systems and Statistics (CRES), Tech. Note No. TR/156.

Pierce, D.A. (1972) 'Least squares estimation in dynamic disturbance time-series models' *Biometrika* **59** 73-78

Plackett, R.L. (1950) 'Some theorems in least squares', *Biometrika* **37** 149-157.

Priestley, M.B. (1988) *Nonlinear and Nonstationary Time Series Analysis*, Academic Press.

Ruanaidh, J.J. and Fitzgerald, W.J. (1996) *Numerical Bayesian Methods Applied to Signal Processing*, Springer.

Schweppe, F. (1965) 'Evaluation of likelihood function for Gaussian signals', *IEEE Transactions on Information Theory* **11** 61-70.

Young, P.C. (1969) 'Applying parameter estimation to dynamic systems: Part l, Theory', Control Engineering, 16, 10, 119-125; 'Part lI Applications', 16, 11, 118-124.

Young, P.C. (1970a) 'An instrumental variable method for real-time identification of a noisy process', *Automatica* **6** 271-287.

Young, P.C. (1970b) *Differential equation error method of real-time process identification*, PhD Thesis, Cambridge University.

Young, P.C. (1978) 'A general theory of modeling for badly defined dynamic systems', in *Modeling, Identification and Control in Environmental Systems*, G.C. Vansteenkiste (ed.), North Holland, 103-135.

Young, P.C. (1983) 'The validity and credibility of models for badly defined systems', in *Uncertainty and Forecasting of Water Quality*, M.B. Beck and G. Van Straten (eds.), Springer , 69-100.

Young, P.C. (1984) *Recursive Estimation and Time-Series Analysis*, Springer.

Young, P.C. (1988) 'Recursive extrapolation, interpolation and smoothing of non-stationary time series', in *Identification and System Parameter Estimation*, C.F. Chen (ed.), Pergamon Press, 33-44.

Young, P.C. (1993a) 'Time variable and state dependent modelling of nonstationary and nonlinear time series', in *Developments in Time Series Analysis*, T. Subba Rao (ed.), Chapman and Hall, 374-413.

Young, P.C. (1993b) *Concise Encyclopedia of Environmental Systems*, Pergamon Press.

Young, P.C. (1994) 'Time-variable parameter and trend estimation in non-stationary economic time series', *Jnl. of Forecasting* **13** 179-210.

Young, P.C. (1996) 'A general approach to identification, estimation and control for a class of nonlinear dynamic systems', in *Identification in Engineering Systems*, M.I. Friswell and J.E. Mottershead (eds.), University of Wales, Swansea, 436-445.

Young, P.C. (1998a) 'Data-based mechanistic modelling of environmental, ecological, economic and engineering systems', *Environmental Modelling and Software* **13** 105-122.

Young, P.C. (1998b) 'Data-based mechanistic modelling of engineering systems', *Journal of Vibration and Control* **4** 5-28.

Young, P.C. (1999a) 'Nonstationary time series analysis and forecasting', *Progress in Environmental Science* **1** 3-48.

Young, P.C. (1999b) 'Data-based mechanistic modelling , generalised sensitivity and dominant mode analysis', *Computer Physics Communications* **115** 1-17.

Young (1999c) 'Comments on time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives', *Jnl. Royal Stat. Soc., Series B* **62**, in press.

Young, P.C. and Beven, K.J. (1994) 'Data-based mechanistic modelling and the rainfall-flow nonlinearity', *Environmetrics* **5** 335-363.

Young, P.C. and Fawcett, C. (1999) 'Data-based mechanistic modelling and the Nicholson blowfly data', Centre for Research on Environmental Systems and Statistics (CRES), Tech. Note No. TR140.

Young, P.C. and Jakeman, A.J. (1979) 'Refined instrumental variable methods of recursive time-series analysis: part I, single input single output systems' *Int. Jnl. of Control* **29** 1-30.

Young, P.C. and Lees, M.J. (1993) 'The Active Mixing Volume (AMV): a new concept in modelling environmental systems', Chapter 1 in *Statistics for the Environment*, V. Barnett and K.F. Turkman (eds.), Wiley, 3-44 .

Young, P.C. and McKenna, P. (1999) 'An instrumental variable approach to recursive fixed interval smoothing and the estimation of time variable parameter transfer function models', Centre for Research on Environmental Systems and Statistics (CRES), Tech. Note No. TR141.

Young, P.C. and Minchin, P. (1991) 'Environmetric time-series analysis: modelling natural systems from experimental time-series data', *Int. Jnl. Biol. Macromol.* **13** 190-201.

Young, P.C. and Ng, C.N. (1989) 'Variance intervention', *Jnl. of Forecasting* **8** 399-416.

Young, P.C. and Pedregal, D.J. (1996) 'Recursive fixed interval smoothing and the evaluation of LIDAR measurements', *Environmetrics* **7** 417-427.

Young, P.C. and Pedregal, D.J. (1997) 'Data-based mechanistic modelling', in *System Dynamics in Economic and Financial Models*, C. Heij, B. Hanzon and K. Paggman (eds.), Wiley, 169-213.

Young, P.C. and Pedregal, D.J. (1998) 'Recursive and en-bloc approaches to signal extraction', *Journal of Applied Statistics* **26** 103-128.

Young, P.C. and Pedregal, D.J. (1999) 'Macro-economic relativity: government spending, private investment and unemployment in the USA 1948-1998', *Jnl. Structural Change and Economic Dynamics*, in press.

Young, P.C. and Runkle, D.E. (1989) 'Recursive estimation and modelling of nonstationary and nonlinear time series', *Adaptive Systems in Control and Signal Processing* **1**, IFAC/Inst. Measurement and Control, London, 49–64.

Young, P.C., Lane, K., Ng, C.N. and Parker, D. (1991) 'Recursive forecasting, smoothing and seasonal adjustment of non-stationary environmental data', *Journal of Forecasting* **10** 57-89.

Young, P.C., Ng, C.N. and Armitage, P.(1989) 'A systems approach to economic forecasting and seasonal adjustment', *International Journal on Computers and Mathematics with Applications* **18** 481-501.

Young, P., Parkinson, S. and Lees, M. (1996) 'Simplicity out of complexity: Occam's Razor revisited', *Journal Of Applied Statistics* **23** 165-210.

Young, P.C., Pedregal, D.J. and Tych, W. (1998) 'Dynamic harmonic regression', Centre for Research on Environmental Systems and Statistics (CRES), Tech. Note No. TR96; in press *Journal of Forecasting*, 1999.

Zhan-Qian, J.L. and Smith, R.L. (1998) 'Estimating local Lyapunov exponents by local polynomial regression', University of North Carolina, Department of Statistics, http://www.stat.unc.edu./poscript/rs/chaos.