ELSEVIER Computational Statistics & Data Analysis 28 (1998) 171–191

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Smoothing non-Gaussian time series with autoregressive structure

Gary K. Grunwald [a,*], Rob J. Hyndman [b]

[a] *Department of Preventive Medicine & Biometrics, Health Sciences Center, University of Colorado, 4200 E. 9th Avenue, School of Medicine B-119, Denver, CO 80262, USA*

[b] *Department of Mathematics, Monash University, Clayton, VIC 3168, Australia*

Received 1 December 1996 ; received in revised form 1 February 1998

Abstract

We consider nonparametric smoothing for time series which are clearly non-Gaussian and which are subject to an autoregressive random component. This generalizes methods for smoothing Gaussian series with autoregressive errors, but in the non-Gaussian case the autoregressive structure is not always additive. The problem can be formulated in a general way to include most common non-Gaussian autoregressive models. The amount of smoothing can be chosen by penalized likelihood methods, and we give simulations and parametric bootstrap methods for studying and empirically estimating the penalty function. We illustrate these methods, the generality of their application, and several data analytic methods with examples involving real data.

Keywords: AIC; Binary time series; Exponential family time series; Penalized likelihood; Poisson time series; Smoothing with correlated errors

1. Introduction

There has been much recent interest in smoothing Gaussian time series using nonparametric smoothers (splines or kernel smoothers, for instance) when the errors η_t in the general smoothing model

$$y_t = \mu_t + \eta_t \tag{1}$$

are correlated and μ_t is a smooth function of t. Diggle and Hutchinson (1989) considered the case when η_t follows a Gaussian autoregressive model of order 1

* Corresponding author.

0167-9473/98/$19.00
PII S0167-9473(98)00034-6

(AR(1)), and Hart (1991), Altman (1990) and Kohn et al. (1992) considered more general ARMA errors. Models like these which use a combination of longer-term smooth components and short-term correlation structure often give a useful and interpretable semiparametric model for a time series.

The problem of smoothing in the presence of correlation structure also arises in the analysis of time series of counts, proportions, strictly positive values, and other non-Gaussian settings. The motivation for models with smooth mean terms and correlated errors is perhaps even stronger for non-Gaussian series than for Gaussian series. The most common treatment of nonstationarity for Gaussian series is differencing, but because the non-Gaussian data may be counts or proportions or nonnegative values, it is not at all clear how to proceed with differencing. The purpose of this paper is to give models and methods for such situations. The methods in this paper are built from a few standard tools such as nonparametric smoothers, maximum likelihood, penalized likelihood, and autoregressive structure, all of which are easily interpreted and understood by practitioners.

In Section 2 we give a general class of parametric non-Gaussian autoregressive models to be used as a basis for semiparametric models involving smooth mean terms, and in Section 3 we modify these models to have a smoothly varying mean function. Estimation can be carried out using standard nonparametric smoothing methods, with a penalized likelihood approach to select the degree of smoothing. The form of the likelihood penalty function is very important, and in Sections 4.2 and 4.3 we give simulations and graphical methods for studying the effects of various penalties, and suggest a method based on parametric bootstrap sampling for empirically estimating the penalty. Throughout, and particularly in Sections 2.2 and 5, we illustrate the methods and the wide variety of models and data to which they apply with several examples. In Section 6 we mention some of the many topics needing further work, including higher order autoregressive structure, weighted smoothing, and relations to other models.

2. Models with general AR(1) structure

2.1. Model formulation

The usual AR(1) model with mean μ is

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + \varepsilon_t, \tag{2}$$

where $-1 < \phi < 1$ and ε_t are i.i.d. with mean 0 (Brockwell and Davis, 1991, for instance). If $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$ this is a Gaussian AR(1) model. However, many non-Gaussian time-series models cannot be written in this way in terms of i.i.d. innovations (the three examples below, for instance) but still possess the main features important to an autoregressive process. For this reason, we give a more general definition of linear autoregressive structure. We work with the first-order process (AR(1)), but the generalization to higher-order AR(p) processes is in many cases straightforward and is illustrated in Section 6.1.

Let $\{Y_t\}, t = 0, 1, \ldots$, be a time-homogeneous first-order Markov process on sample space $\boldsymbol{Y} \subseteq \mathbb{R}$, and assume we have available the (nondegenerate) conditional density $p(y_t \mid y_{t-1})$. Here, we assume conditioning on Y_1 and $\boldsymbol{\theta}$, a vector of length d_θ of any other parameters involved in p. Let $m(Y_{t-1}) \equiv E(Y_t \mid Y_{t-1})$, which is assumed to be in $\boldsymbol{M}$, the set of values of $m(Y_{t-1})$ for which p is a nondegenerate proper probability density.

We assume autoregressive structure represented by a linear conditional mean function

$$m(Y_{t-1}) = \phi Y_{t-1} + \lambda, \tag{3}$$

where ϕ and λ can take any values such that $m(Y_{t-1}) \in \boldsymbol{M}$ for all possible values of $Y_{t-1} \in \boldsymbol{Y}$. Since many non-Gaussian models require $0 \leq \phi < 1$, and since nonnegative autocorrelation is by far the most common case in practice, we make this assumption as well.

This setting gives a sufficiently general definition of an AR(1) process for our purposes. The linear Gaussian AR(1) model is a special case with p a normal density, $\boldsymbol{Y} = \mathbb{R}, \boldsymbol{M} = \mathbb{R}, \lambda = (1 - \phi)\mu$ and $\boldsymbol{\theta} = \sigma$. However, the above formulation encompasses many models which cannot be stated in the innovations form (2), even with ε_t non-Gaussian. For instance, the negative binomial model of Lampard (1968), the DAR(1) model of Jacobs and Lewis (1978), the EAR(1) model of Gaver and Lewis (1980), the Poisson AR(1) model of McKenzie (1988) and Al-Osh and Alzaid (1987), and of Heyde and Seneta (1972) and Venkataraman (1982), the nonnegative AR(1) model of Hutton (1990), the geometric and negative binomial models of McKenzie (1986), the product AR(1) models of Grunwald and Feigin (1995), the exponential family AR(1) models of Zeger and Qaqish (1988) (if an identity link is used in their Generalized Linear Model formulation), and the general AR(1) models of Sim (1994) are all included in this formulation.

Grunwald et al. (1997) have studied properties for models under this general formulation, some of which will be used here. In particular, the process mean is easily available: If $E(Y_1) = \lambda/(1 - \phi) \equiv \mu$ then $E(Y_t) = \mu$ for $t \geq 1$. Further, if $\mathrm{Var}(Y_t)$ is finite and constant in time, then the autocorrelation function (ACF) has the usual AR(1) form

$$\mathrm{Corr}(Y_t, Y_{t-k}) = \phi^k \quad \text{for } k = 0, 1, 2, \ldots . \tag{4}$$

This means that the sample ACF can be used as usual as a model diagnostic. We illustrate this in the examples below.

2.2. Examples

We now give three examples illustrating the wide variety of situations covered by the above formulation. In this section we give the data and context and describe a model with constant mean and linear AR(1) structure which could be used for the data. In later sections we show how the models can be modified to accommodate a smooth time-varying mean, and carry out the analysis.

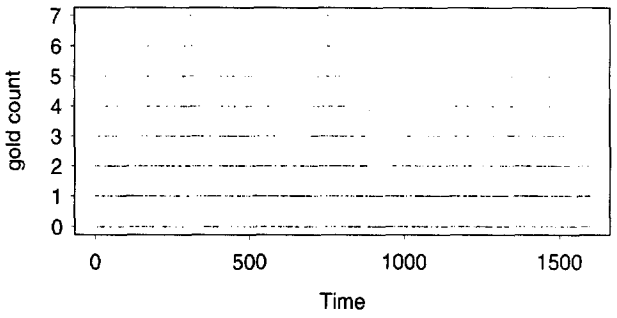

Fig. 1. Westgren's (1916) gold series, as given in Guttorp (1991).

2.2.1. *Example 1: Westgren's gold series*

Fig. 1 shows data from Westgren's (1916) experiment C as given in Guttorp (1991). There are 1598 observations with integer values from 0 to 7 inclusive. The values are the numbers of gold particles observed in the same small volume of a solution every 1.39 seconds. Particles may move in and out of the volume.

These data have been used or referred to by Chandrasekhar (1954), Heyde and Seneta (1972) and Guttorp (1991) as a series for which a model of the form

$$Y_t = \sum_{i=1}^{Y_{t-1}} B_{i,t}(\phi) + P_t(\lambda) \tag{5}$$

with $\lambda > 0$ is appropriate. $B_{i,t}(\phi)$ are Bernoulli random variables with $\mathrm{P}[B_{i,t}(\phi) = 1] = \phi$ which are independent for all t and i, and the $P_t(\lambda)$ are Poisson random variables with mean λ which are independent of all $B_{i,t}(\phi)$ and of each other.

In the stochastic processes literature, this model appears as an infinite server $(M/M/\infty)$ queue (Parzen, 1962) and as a Poisson branching process with immigration (Heyde and Seneta, 1972, for instance), and in the time series literature as a Poisson AR(1) model (McKenzie, 1985 or Al-Osh and Alzaid, 1987 for instance). It satisfies (3) and the other conditions in Section 2.1 but cannot be written in the i.i.d. innovations form (2). Although the model with $0 \leq \phi < 1$ and $\lambda > 0$ was constructed to be stationary and have Poisson marginal distribution $Y_t \sim \text{Poisson}(\mu)$, $p(y_t \mid y_{t-1})$ is readily available (Aly and Bouzar, 1994 for instance). No additional parameters are involved, so θ is null.

The sample autocorrelation function for the gold series is shown in Fig. 2. If model (5) is correct, it should decay exponentially by (4) but it does not. Using the parametric bootstrap diagnostic methods of Tsay (1992), the dotted lines show 2.5% and 97.5% quantiles of the ACF values from 100 simulated series from the model (5) fitted by maximum likelihood. The general decay with increasing lag is evident for small lags, but there is more autocorrelation in the series than the constant mean AR(1) model (5) can account for.

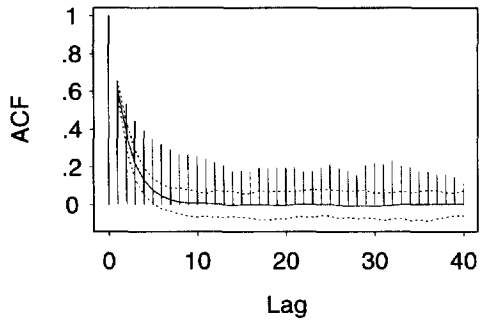

Fig. 2. The sample ACF for the gold series shows that the AR(1) model is not adequate.

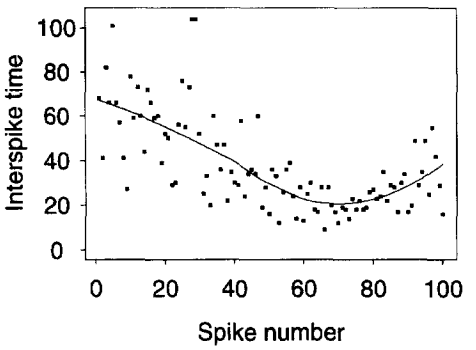

Fig. 3. Zeger and Qaqish's (1988) neurons data. The values are the interarrival times of spikes in the motor cortex of an unstimulated monkey. The smooth curve was estimated by the methods in Sections 3 and 4.

Table 1
Results of penalized likelihood fits to neurons series

Penalty	$\hat{\phi}$	$\hat{D}$	$\hat{\theta}$	AIC/BIC
AIC	0.069	6.3	8.3	778.9
BIC	0.085	6.0	8.2	794.9
AIC	0	6.3	8.3	779.5
BIC	0	6.1	8.2	795.6

2.2.2. *Example 2: Neuron interarrival times for a monkey*

Zeger and Qaqish (1988) give a series of interspike times of neurons in the motor cortex of a monkey. Fig. 3 shows the series. The data shown in Table 1 of Zeger and Qaqish (1988) contain several typographical errors. The correct data were obtained from S. Zeger (personal communication); the four corrected values are 60, 36, 60 and 25 at times 13, 36, 47 and 61. (This discrepancy also appears to explain why Li's (1991) results differed from Zeger and Qaqish's results in estimating the same model.)

These authors suggest an autoregressive model based on Gamma distributions:

$$(Y_t \mid Y_{t-1}) \sim \mathrm{G}(m(Y_{t-1}), \theta), \tag{6}$$

where $G(m,\theta)$ represents a random variable having Gamma distribution with mean m and variance m^2/θ with shape parameter $\theta \in (0,\infty)$. ($X \sim G(m,\theta)$ if $p(x) = \exp\{-\theta x/m\}(\theta/m)^\theta x^{\theta-1}/\Gamma(\theta)$ for $x > 0$, $\theta > 0$, $m > 0$). Since Zeger and Qaqish work in a Generalized linear model (GLM) context (McCullagh and Nelder, 1989) they use an inverse link function in specifying $m(Y_{t-1})$, but in this paper we assume an identity link and a linear AR(1) structure for $m(Y_{t-1})$ as in (3) with $\lambda > 0$. This linear AR(1) model cannot be written in the innovations form (2) with i.i.d. innovations, but still has AR(1) structure as defined in Section 2.1. Grunwald and Feigin (1995) have studied properties and estimation for this model. The smooth curve in Fig. 3 was estimated assuming a linear AR(1) random component by the methods described in Sections 3 and 4.

2.2.3. Example 3: Daily rainfall occurrence in Melbourne, Australia, 1980–1989

The series of daily rainfall occurrence in Melbourne, Australia, is binary with 0 indicating no rain and 1 indicating rain. Rain occurred on 1416, or 38.8%, of the 3653 days.

To model this binary series, we consider models of the form $(Y_t \mid Y_{t-1}) \sim B(m(Y_{t-1}))$ with $m(Y_{t-1}) = \phi Y_{t-1} + (1-\phi)\mu$ and $0 < \mu < 1$. $B(m)$ represents a Bernoulli random variable with mean m. Again, it is not possible to write this model in the innovations form (2), but it is of the AR(1) form described in Section 2.1. In Section 5 we generalize this model to allow for seasonality and other trends with a smooth time-varying mean μ_t.

3. Models with a time-varying mean

3.1. Formulation

We can generalize the models in Section 2.1 to allow a time-varying mean by letting $(Y_t \mid Y_{t-1})$ have density $p_t(y_t \mid y_{t-1})$ with

$$E(Y_t \mid Y_{t-1}) = m_t(Y_{t-1}) = \phi Y_{t-1} + \lambda_t, \tag{7}$$

where $\lambda_t = \mu_t - \phi\mu_{t-1}$. As before, ϕ, λ_t and μ_t can take any values such that $m_t(Y_{t-1}) \in \boldsymbol{M}$ for all possible values of Y_{t-1}. We sometimes use vector notation $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$. It is easy to show that for these models, if $E(Y_1) = \mu_1$ then $E(Y_t) = \mu_t$ for $t \geq 1$. McKenzie (1985) suggested this modification of the Poisson model of Example 1, but did not discuss estimation.

In the literature on smoothing Gaussian series with autocorrelated errors, the standard model for AR(1) errors is $Y_t = \mu_t + \eta_t$ where $\eta_t = \phi\eta_{t-1} + \varepsilon_t$ with ε_t i.i.d. $N(0,\sigma^2)$ (Diggle and Hutchinson, 1989; Altman, 1990; Hart, 1991 and others). This can be written $(Y_t \mid Y_{t-1}) \sim \text{Normal}(m_t(Y_{t-1}), \sigma^2)$ with $m_t(Y_{t-1}) = \phi Y_{t-1} + \mu_t - \phi\mu_{t-1}$, motivating (7).

3.2. Estimation

Under relatively mild assumptions on $p(y_t \mid y_{t-1})$, maximum likelihood can be used to estimate ϕ, μ and θ in models with constant mean. Grunwald and Feigin (1995) give an application of the general methods from Billingsley (1961) to the model in Example 2. For some constant mean models, explicit forms are available. For time-varying mean models (7), the log-likelihood conditional on $\hat{\mu}$ and y_1,

$$LL(\phi, \theta \mid \hat{\mu}, y_1) = \sum_{t=2}^{n} \log(p_t(y_t \mid y_{t-1}, \hat{\mu})) \tag{8}$$

can be numerically maximized. Estimates of μ can be found using standard smoothing methods such as kernel, locally weighted polynomial or spline smoothers. For the models with time-varying mean given in the previous section, $E(Y_t) = \mu_t$ so applying a nonparametric smoother to the graph of y_t versus t is easily interpreted as an estimate of $E(Y_\tau \mid \tau = t) = E(Y_t) = \mu_t$. For a given smoother, the family of estimated curves $\hat{\mu}$ is parameterized by a smoothing parameter d_μ, which we take to be the equivalent degrees of freedom of the smoother as defined by $d_\mu = \text{trace}(S)$ where S is the smoothing matrix, i.e. $\hat{\mu} = Sy$ (other definitions are possible – see Hastie and Tibshirani (1990) for further discussion). For the constant mean model, $d_\mu = 1$. We let D denote the total equivalent parameters of the model, so $D = d_\mu + d_\theta + 1$.

Several other theoretical and practical issues arise in estimating $\hat{\mu}$. The variances of the observations typically change over time and depend on the mean function μ and correlation structure. We discuss this issue further in Section 6.2. It may also be necessary in the course of numerical optimization to impose the model constraints on μ_t and λ_t stated in Section 3.1. For instance, the constraint $\lambda_t = \mu_t - \phi\mu_{t-1} > 0$ in a model for non-negative counts can be imposed artificially by setting $\lambda_t = \max(\varepsilon, \mu_t - \phi\mu_{t-1})$ where $\varepsilon > 0$ is a small number.

Other methods have been proposed for smoothing non-Gaussian time series. For instance, West et al. (1985), Kitagawa (1987), Fahrmeir (1992), and Kashigawi and Yanagimoto (1992) discuss methods based on state space models. These methods require a large amount of specialized programming and a knowledge of state space modelling that is not likely to be available to many statistical consultants. The smoothers are also defined less directly and so their properties are not as easily studied. By allowing standard nonparametric smoothers such as splines, kernel smoothers and locally weighted polynomials, the methods described in this paper use familiar, easily understood and widely available tools whose performance is better understood and can be tailored to the specific application (Hastie and Loader, 1993 for instance).

4. Bandwidth selection

4.1. Penalized likelihood methods

There is much ongoing discussion about the best methods for selecting the amount of smoothing. Two broad classes of methods, penalized likelihood and crossvalida-

tion, have been suggested for selecting the number of parameters in a general statistical model. In Gaussian settings, where residuals are natural, crossvalidation or generalized crossvalidation based on "leave-one-out" residuals is a natural concept. In non-Gaussian settings, where residuals are less natural and the likelihood is used for inference, penalized likelihood methods such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) seem to us more natural. Hastie and Tibshirani (1990) have suggested a version of generalized crossvalidation for use in non-Gaussian settings, but extending these ideas to correlated situations, which requires specialized computations with the correlation matrix in the Gaussian case (see Altman 1990 or Hart, 1991 for instance), appears difficult and involves several approximations. Also, as shown by van der Linde (1994), leave-one-out methods for correlated data require modification, since even when a point is left out, information about it remains through the correlation structure. Furthermore, as Hastie and Loader (1993) points out, penalized likelihood methods provide a sensitivity analysis to the smoothing parameter which are not directly available with crossvalidation methods. We illustrate this below. AIC and BIC are also the standard tools in selecting the order of Gaussian time series models (Brockwell and Davis, 1991).

The log-likelihood for $(\phi, \boldsymbol{\mu}, \boldsymbol{\theta})$, or equivalently, for a specified smoother, for $(\phi, D, \boldsymbol{\theta})$, is

$$LL(\phi, \boldsymbol{\mu}, \boldsymbol{\theta} \mid y_1) = LL(\phi, D, \boldsymbol{\theta} \mid y_1) = \sum_{t=2}^{n} \log(p_t(y_t \mid y_{t-1})).$$

Maximizing this over $(\phi, \boldsymbol{\mu}, \boldsymbol{\theta})$ simply gives interpolation for $\hat{\boldsymbol{\mu}}$, so various methods have been considered to penalize the log-likelihood for adding parameters. The two most popular are (Harvey, 1989).

$$\begin{aligned} \text{BIC} &= -2LL(\phi, D, \boldsymbol{\theta}) + \log(n-1)D, \\ \text{AIC} &= -2LL(\phi, D, \boldsymbol{\theta}) + 2D. \end{aligned} \tag{9}$$

In practice, the model is estimated for values of D on a grid to minimize AIC or BIC. If for each value of D, maximum likelihood estimates conditional on $\hat{\boldsymbol{\mu}}$ are calculated as in Section 3.2, the maximizing $(\hat{\phi}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})$ is the maximum penalized likelihood estimate.

Fig. 4 shows the AIC (left panel) and BIC (right panel) for the gold series of Example 1 plotted against D for smoothing spline fits. Fig. 5 shows the series with smoothing spline curves selected by the AIC, with $d_\mu = 25$ (dark solid line) and BIC, with $d_\mu = 10$ (light solid line). Thus, the form of the penalty function can make a huge difference in the estimated curve.

It also has been noted by several authors (Altman, 1990; Hart, 1991; Herrmann, Gasser and Kneip, 1992 for instance) that smoothing parameter selection methods that do not take into consideration the correlation when observations are correlated tend to select smoothing parameters which undersmooth, attributing the autocorrelation to the smooth mean term. Similar behavior is noted throughout this study. For instance, the dotted line in Fig. 5 has $d_\mu = 120$ degrees of freedom and was selected by the AIC assuming no autocorrelation ($\phi = 0$).

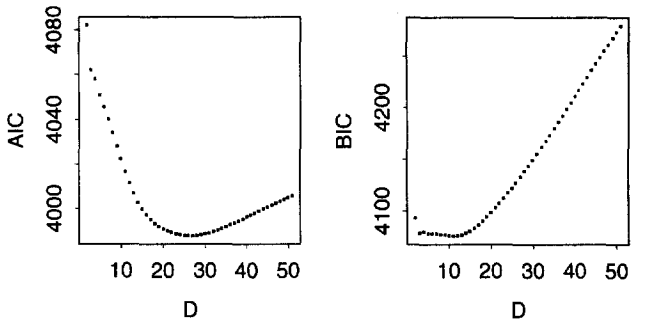

Fig. 4. AIC and BIC for the gold series versus number of model parameters D.

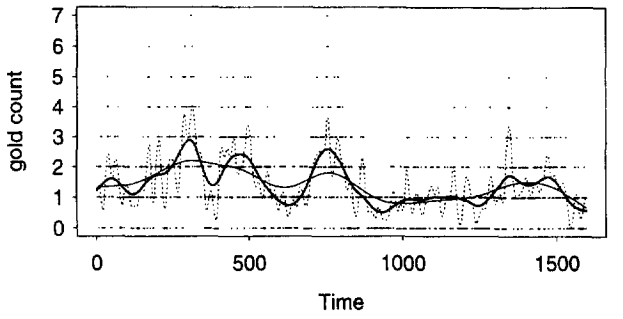

Fig. 5. The gold series with curve estimates selected by AIC (dark line), BIC (light line) and AIC assuming no autocorrelation (dotted line).

4.2. Simulations

Which is "best", AIC or BIC? There has been much work on the choice of penalty functions for order selection in ARMA time series models, but the situation even in that setting is still quite contradictory, as the discussion in Brockwell and Davis 1991, Section 9.3, summarizes. Theoretical arguments concerning the AIC and BIC are asymptotic and it is not clear if such arguments are practically helpful in the present context because the series are extremely non-Gaussian, the number of parameters involved is fairly large compared to the length of the series, and the positive correlation between observations tends to reduce the effective sample size.

Given a particular series of length n, the problem is to find a penalty function pen(D) such that the $(\hat{\phi}, \hat{D}, \hat{\theta})$ which minimize

$$-2LL(\phi, D, \theta) + \text{pen}(D) \tag{10}$$

give a "good" estimate of μ. As a criterion, we use the mean-squared error $\text{MSE}(\hat{\mu}) = E\left[(1/n)\sum_{t=1}^{n}(\mu_t - \hat{\mu}_t)^2\right]$. For a given series, we need not model the dependence of pen(D) on n, only on D, and both the AIC and BIC in (9) use penalty functions linear in D. We will restrict attention to such penalty functions: $\text{pen}(D) = fD$ for $f > 0$. (Chen et al. (1993) make this assumption also.)

Simulations can be performed efficiently to yield enough information for a careful study of the effect of different f values on MSE($\hat{\mu}$). For a fixed model $(\phi^*, \mu^*, \theta^*)$, which we refer to as the *base model*, proceed as follows:

Part I:

1. Simulate series $y^{(i)}$ for $i = 1,\ldots,I$ from base model $(\phi^*, \mu^*, \theta^*)$.
2. Take a sequence of possible values of D, say $D(j)$, $j = 1,...,J$. For each value of $D(j)$, and for each series $y^{(i)}$,
   2.1. smooth $y^{(i)}$ to obtain $\hat{\mu}^{(i)}(j)$,
   2.2. compute $\text{ASE}(\hat{\mu}^{(i)}(j)) = (1/n)\sum_{t=1}^{n}(\mu_t^* - \hat{\mu}_t^{(i)}(j))^2$, and
   2.3. find $\hat{\phi}^{(i)}(j)$ and $\hat{\theta}^{(i)}(j)$ which minimize $-2LL(\phi, D(j), \theta)$. These are maximum likelihood estimates conditional on $\hat{\mu}^{(i)}(j)$, or equivalently on $D(j)$.

We now have the set of points $(\hat{\phi}^{(i)}(j), D(j), \hat{\theta}^{(i)}(j))$ representing the minimizing values of $-2LL(\phi, D, \theta)$ for the given set of $D(j)$, and the values of $-2LL(\phi, D, \theta)$ and $\text{ASE}(\hat{\mu}^{(i)}(j))$ on this set. Since the penalty function is linear and depends only on D, the minimum of any penalized likelihood of the form (10) lies in this set of points. So, the penalties can be applied later, as below. Note also that the many curve estimates $\hat{\mu}^{(i)}(j)$ need not be stored after $\text{ASE}(\hat{\mu}^{(i)}(j))$ is calculated in step 2.2.

Part II:

3. Take a sequence f_k, $k = 1,\ldots,K$ of possible values of f. For each series $y^{(i)}$:
   3.1. For each penalty f_k, find the $D(j)$ which minimizes

$$-2LL(\hat{\phi}^{(i)}(j), D(j), \hat{\theta}^{(i)}(j)) + f_k D(j)\,.$$

   Call this $\hat{D}_k^{(i)}$.
   3.2. Call the values of ϕ and θ corresponding to this minimum $\hat{\phi}_k^{(i)}$ and $\hat{\theta}_k^{(i)}$.
4. $(\hat{\phi}_k^{(i)}, \hat{\mu}_k^{(i)}, \hat{\theta}_k^{(i)})$ represents the optimal model for series $y^{(i)}$ under penalty f_k. Summarize these to illustrate the performance of the penalty f_k. For instance, minimize $\text{ASE}_k = (1/I)\sum_{i=1}^{I}\text{ASE}(\hat{\mu}_k^{(i)})$ with respect to k.

Simulations were done using the model of Example 1 with $I = 100$ series having known mean function $\mu_t^* = \sin(\pi t/h) + 2$ with $h = 80$ and each of three values of ϕ, 0.25, 0.50 and 0.75. All work was done in SPLUS, with a short FORTRAN routine for speeding the recursive calculation of the log-likelihood. The simulations used $D(j) = j$ for $j = 1,\ldots,120$, and were summarized for f_k values $\{1.0, 1.25, 1.50, \ldots, 12.00\}$. Fig. 6 shows $\text{ASE}(\hat{\mu}_k^{(i)})$ versus f_k as light lines for each individual series and the average, ASE_k, as solid dark line. ASE_k is greater for larger ϕ^*, as expected. For each value of ϕ^*, the minimum ASE_k is achieved with f near 2, which corresponds to the AIC. The sensitivity of ASE_k to the penalty is greater for larger ϕ^*. Note that this sensitivity information is more practically useful than the sensitivity of AIC and BIC to the number of parameters, as shown in Fig. 4, because MSE($\hat{\mu}$) is of more direct practical interest.

These simulations were repeated for another known mean function, $\mu_t^* = \sin(\pi t/h) + 2$ with $h = 400$, with the results shown in Fig. 7. The lines for $\phi^* = 0.25$ and

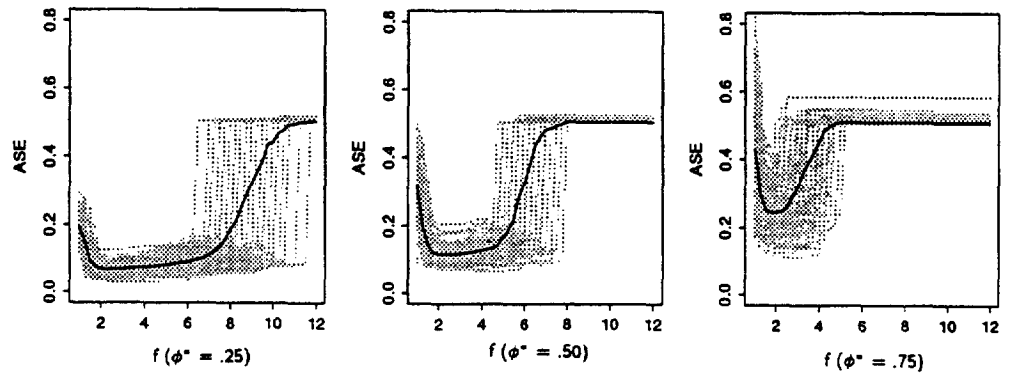

Fig. 6. ASE_k from simulations with base model $\mu_t^* = \sin(\pi t/80) + 2$ and $\phi^* = 0.25$ (left panel), $\phi^* = 0.50$ (middle panel) and $\phi^* = 0.75$ (right panel).

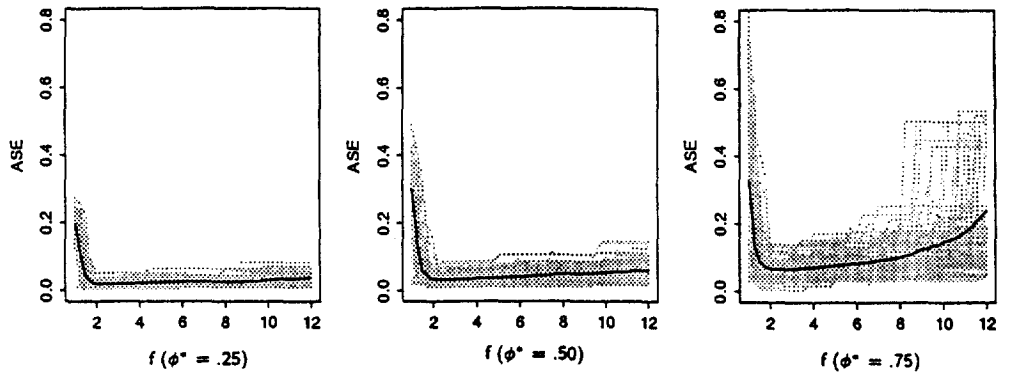

Fig. 7. ASE_k from simulations with base model $\mu_t^* = \sin(\pi t/400) + 2$ and $\phi^* = 0.25$ (left panel), $\phi^* = 0.50$ (middle panel) and $\phi^* = 0.75$ (right panel).

$\phi^* = 0.50$ eventually increase quickly for very large penalties ($f > 15$), but this is not shown as it is well outside the range of interest.

4.3. *Empirical selection of penalty functions*

The μ^* and ϕ^* used in the above simulations were chosen to cover the sort of models expected to be good for the gold series, based on trying various smoothers on the series and comparing ACFs of simulated series with that of the series. Since the goal is to simulate from a model as close as possible to the "true" model, using any "reasonable" model as the base model $(\phi^*, \mu^*, \theta^*)$ from which to run a simulation experiment as in the previous section should give the necessary information about optimal penalties, *without using the known values of* h^*, ϕ^*. We thus add an initial step to the algorithm in Section 4.2:

0. Use some "reasonable" penalty to obtain a base model $(\hat{\phi}, \hat{\mu}, \hat{\theta}) = (\phi^*, \mu^*, \theta^*)$ from which to run the simulation.

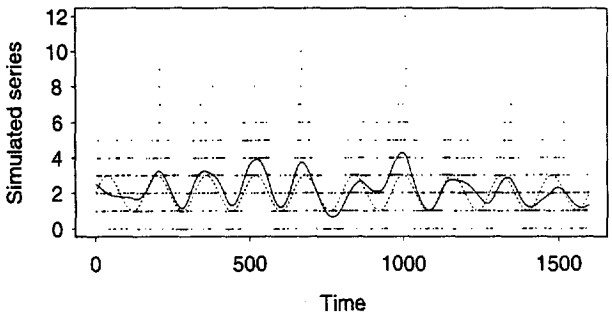

Fig. 8. Simulated series with true mean curve $\mu_t = \sin(\pi t/80) + 2$ (dotted curve) and estimated curve with $d_\mu = 31$ (solid curve).

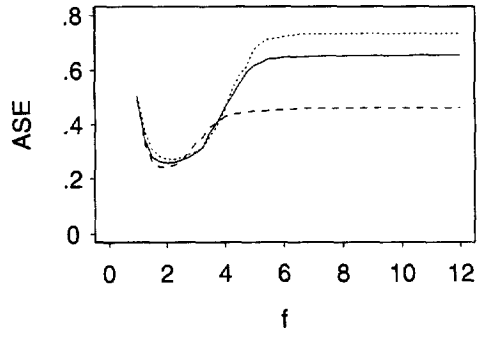

Fig. 9. ASE($\hat{\mu}$) versus f_k for $f = 1.5$ (dotted line), $f = 2$ (solid line) and $f = 4$ (dashed line) from simulations with base model μ^* the fitted curve with $d_\mu = 31$.

This approach is somewhat similar to the parametric bootstrap method of Chen et al. (1993) for empirically estimating penalties for selecting the order of an autoregressive process. Both problems involve estimating the dimension of a model.

To study this approach for, say, $h = 80$ and $\phi = 0.75$ (this should be the most challenging of the models considered since it is most sensitive to the form of the penalty function), we began with a series $\boldsymbol{y}$ simulated from the fitted model, but we now use no information about how the series was generated in estimating an appropriate penalty. For $\boldsymbol{y}$ shown in Fig. 8, AIC ($f = 2$) selects $\hat{D} = 31$ as optimal, yielding $\hat{\phi} = 0.749$ and $\hat{\boldsymbol{\mu}}$ shown in Fig. 8.

Using this fitted model as the base model, Fig. 9 shows ASE_k versus f_k as the solid line. ASE_k is minimized by $f = 2$. To study the effect of the base model, we also used $f = 1.5$ and $f = 4$ to obtain base models, giving $(\hat{\phi},\hat{D}) = (\phi^*,D^*) = (0.748,34)$ and $(\hat{\phi},\hat{D}) = (\phi^*,D^*) = (0.754,25)$ respectively. The results, shown as dotted and dashed lines in Fig. 9, again indicate that the AIC gives optimal estimation of $\boldsymbol{\mu}$ for models like these, and this conclusion was reached *without using the knowledge of the true model* $h = 80, \phi = 0.75$.

We suspect the theory required to determine conditions under which such an algorithm does actually select the optimal penalty will be very difficult. From experi-

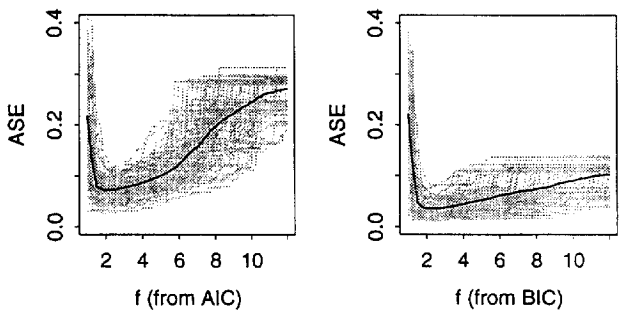

Fig. 10. Empirical selection of the amount of smoothing for the gold series using base models from AIC (left panel) and BIC (right panel).

ences with various simulations and situations, we can say a few things. Determining a base model using too large a penalty often results in penalized likelihood selecting the constant mean model, $d_\mu = 1$, as base model. This can happen if BIC is used to obtain $(\phi^*, \hat{\mu}^*, \theta^*)$ from which to simulate, particularly if the signal is weak. If a constant mean model is in fact best, using a penalty $f < 2$ will still give a reasonable model from which to simulate. Thus, using several penalties to determine base models as above is advisable. The series ACF also provides much guidance.

5. Examples

5.1. *Example* 1 (*continued*): *Westgren gold series*

Fig. 10 (the solid line) shows results for the gold series with $I = 100$ simulated series using AIC ($f = 2$) to determine a base model. This gave $\phi^* = 0.576$, $D^* = 26$, and μ^* as shown in the solid dark line in Fig. 5. In fact, the left panel of Fig. 10 shows that $f = 2$, the AIC, is essentially best for estimating μ in this case.

Physically, μ in a model like Eq. (5) is related to a time changing immigration rate, perhaps indicating changing conditions during the experiment. This has not been detected in previous analyses. Note also the revised estimate of ϕ, compared to $\hat{\phi} = 0.615$ assuming no correlation. The latter includes the correlation due to the time varying mean. Exploratory analyses indicate that there may be some variation in ϕ with time, though substantially less than the variation in the mean. Hyndman and Wand (1996) give one possible approach for Gaussian series, but even in that setting modelling such time-varying correlation is a much more difficult problem which we do not attempt here.

Using the BIC ($f = \log(1597) \approx 7.4$) to estimate the base model for the gold series gives $\phi^* = 0.595$ and $D^* = 11$ (the thin solid line in Fig. 5). The result of simulations from this base model, shown in the right panel of Fig. 10, again suggests $f \approx 2$, so even starting with the BIC base model, we come to prefer the model selected by AIC.

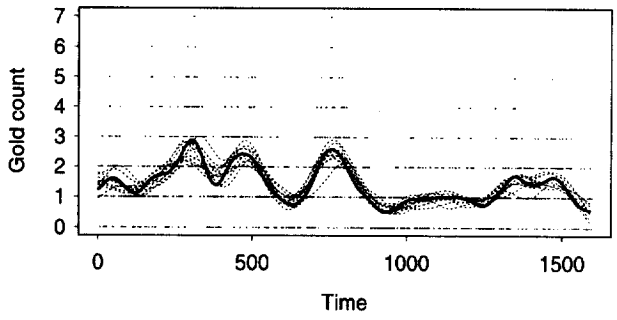

Fig. 11. Ten estimated mean curves for data simulated from the model fitted to the gold series using the penalty $f = 2$. The dark line represents the estimated time-varying mean for the gold series using $d_\mu = 25$ as chosen by empirical estimation of the penalty function.

The bootstrap simulation method yields uncertainty information as a by-product. For instance, Fig. 11 shows 10 curves estimated with penalty $f = 2$, indicating the variability in $\hat{\mu}$. This also provides an additional check that the procedure does well at recovering the base curve μ^* (curves estimated using the BIC penalty show extreme oversmoothing). Further rough checks of this nature can be made by simulating a few series from the fitted model and checking that their graphs, ACF's and so on look like those of the original series. The standard deviation of the 100 values of $\hat{\phi}^{(i)}$ for penalty $f = 2$ in the simulation was 0.019, and their normal QQplot is very straight.

5.2. *Example* 2 (*continued*) : *Neurons from a monkey's brain*

The ACF of the data (not shown) indicates somewhat more correlation than can be described by an AR(1) structure. Here, we consider the AR(1) model based on Gamma distributions described in Section 2.2, generalized as in Section 3 to have a time-varying mean. Proceeding with estimation as described in previous sections gives the top two rows of Table 1. Here, we have used a locally weighted quadratic smoother (LOESS in SPLUS with degree 2, as in Cleveland, et al. 1992) to reduce bias at the ends of the series (see Hastie and Loader, 1993 for a discussion of such bias issues).

Here, the AIC and BIC penalties choose very similar models, the resulting curve estimates are virtually identical, and the empirical estimation of the penalty function indicates little difference between AIC and BIC. Fig. 3 shows a locally weighted quadratic smooth with $d_\mu = 4.3$.

The very low values of $\hat{\phi}$ indicate that nearly all of the correlation is being allocated to the smooth mean function, with little being modelled by the AR(1) structure. The AIC and BIC are only slightly changed by setting $\phi = 0$ (bottom two rows of Table 1), and the resulting smooths are again virtually identical. Grunwald and Feigin (1995) suggested a method for assessing the distributional assumption in a model with Gamma random component. In such a model, $(Y_t \mid Y_{t-1}) \sim \mathrm{G}(m_t(Y_{t-1}), r)$

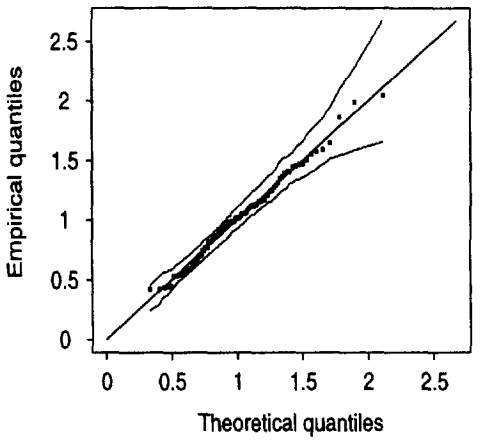

Fig. 12. A QQ-type diagnostic plot for the model fitted to the neuron series indicates no lack of fit.

so $\hat{S}_t \equiv Y_t/\hat{m}_t(Y_{t-1}) \sim \mathrm{G}(1,\hat{r})$ gives a standardized variate on which to base a QQ-type model diagnostic. If the model is adequate, sorted $\hat{S}_t$ versus quantiles of $\mathrm{G}(1,\hat{r})$ should be roughly straight and 1 : 1. Fig. 12 shows such a plot with a 1 : 1 line and 2.5% and 97.5% pointwise quantile bands based on 200 simulated samples of size 99 from $\mathrm{G}(1,\hat{r})$. The fit appears good. For this series, these simple nonparametric smooths with a Gamma random component give alternative models that would in many cases be more useful than the inverse-linked, constant mean AR(2) or the AR(1) with linear trend models fitted by Zeger and Qaqish (1988). We chose to use a Gamma random component for comparability with these previous analyses, but other choices such as log-normal could also prove adequate and could be fitted using the same methods.

5.3. *Example* 3 (*continued*) : *Daily rainfall in Melbourne*

Fig. 13 shows the ACF of the observed rainfall occurrences for lags 1 to 365. The use of the ACF is rather unusual for a binary series, but it contains much information about the series. The theoretical ACF for the Bernoulli AR(1) model decays exponentially, and as in Example 1 the parametric bootstrap method (Tsay, 1992) gives the necessary uncertainty information for interpretation. Here, 100 simulations were made from the fitted i.i.d. Bernoulli model of Example 3 ($\phi = 0$) and for each the sample ACF was computed. The dotted lines show the 2.5% and 97.5% points of the sample correlations for lags 1 to 365. There is a clear seasonal pattern and more positive autocorrelation in the first two lags than would be seen in i.i.d. Bernoulli sequences. (These are quite close to the usual $\pm 2/\sqrt{n} = \pm 0.033$ asymptotic 95% limits on the ACF. Grunwald et al. (1997) showed that these asymptotic limits hold very generally, and in particular in this example.)

Consider fitting the Bernoulli AR(1) model of Example 3, with time-varying mean as described in Section 3.1. Smoothing the series using a cubic smoothing spline,

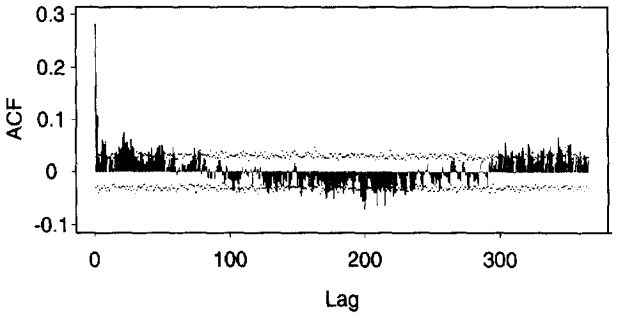

Fig. 13. ACF for Melbourne daily rainfall occurrence, with 95% confidence limits for an i.i.d. Bernoulli series computed using a parametric bootstrap.

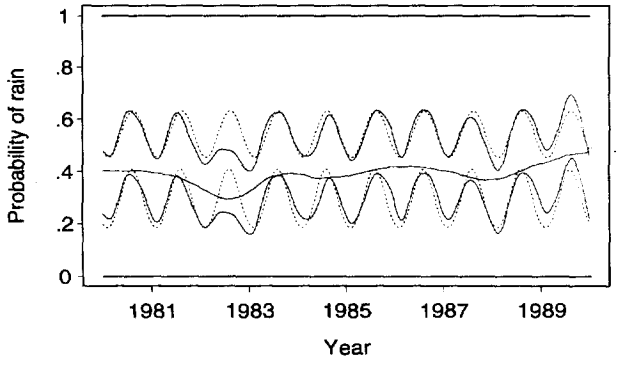

Fig. 14. Estimated conditional probability curves for a Bernoulli AR(1) model with time-varying mean.

the empirical penalty selection method again selects the AIC penalty ($f = 2$), and this gives $\hat{\phi} = 0.24$ and $\hat{D} = 34$. In the Markov chain parameterization, this model says that $P(Y_t = 1 \mid Y_{t-1} = 0) = (1 - \phi)\mu_t, P(Y_t = 1 \mid Y_{t-1} = 1) = \phi + (1 - \phi)\mu_t$ and $E(Y_t) = \mu_t$. Thus, $\hat{\phi} = 0.24$ describes the increase in probability of rain following a wet day as compared with following a dry day, while $\hat{\mu}_t$ is an estimate of the unconditional probability of rain on day t. Fig. 14 shows the two estimated conditional probability curves as the solid oscillating lines. Aside from the seasonal pattern, the most striking feature of these curves is the anomaly in 1982, when the probability of rain was quite low. In fact, 1982 was a time of extreme drought in Victoria associated with a strong El Niño event. It is remembered well since it was followed in February 1983 by severe bushfires over much of Victoria.

It is interesting to compare these results with other analyses. The smooth curve through the middle of Fig. 14 was computed by a moving average of length 365 to remove the seasonal component. For binary data the moving average represents moving proportions. The MA(365) was repeated to give a smoother curve. The conditional probabilities from the time-varying Bernoulli AR(1) model capture the drought and upward trend near the end of the period.

Another comparison can be made with the methods of Stern and Coe (1984), variants of which have become one standard approach in the rainfall literature. By

grouping the 10 years together, computing proportions of rainy days following dry days and following wet days in each of the 73 five day periods, and fitting sinusoids to these Binomial responses using a Generalized Linear Model (McCullagh and Nelder, 1989) and a logit link function, we obtained estimates of conditional probabilities for the seasonal effect, averaged over the 10 years, as shown by the dotted lines in Fig. 14 (the same pattern is repeated each year). The agreement with the Stern and Coe method is quite good with the exception of the anomalies in 1982 and to a lesser extent 1989, which the Stern and Coe method do not attempt to model. From our model it can be seen that, in terms of probability of rainfall, the drought of 1982 was relatively short and primarily during non-summer months. The advantage of the time varying Bernoulli AR(1) analysis is that it includes the seasonal effect, deviations from this seasonal effect, trends, and the autocorrelation structure in the same simple model. The main disadvantage in comparison with the Stern and Coe method is the assumption of constant ϕ, which requires the same temporal pattern following wet or dry days, quite a strong assumption. The Stern and Coe model does allow these to differ. The dotted lines in Fig. 14 indicate that the assumption of constant ϕ may not be too far off for these data, though further exploratory analyses do indicate some aspects of lack of fit.

6. Further topics

6.1. Higher-order autoregressive models

We have thus far considered only first-order autoregressive structure. In some cases, such as the gold series of Example 1, the physical situation makes it clear that an AR(1) model with time-varying mean is much more natural and interpretable than a higher order AR model. In other cases higher-order models could be useful. For instance, in modelling occurrence of rainfall a second-order model of the sort considered in Example 3 corresponds to a second-order Markov chain, and says that the probability of rain today is influenced by rainfall occurrence yesterday and the day before. Fitting this model with constant mean and with time varying mean to the Melbourne daily rainfall series using the AIC penalty gives the results in Table 2.

For constant mean models the AR(2) is slightly better, as the second parameter attempts to capture the seasonal variation. The time-varying mean models are far

Table 2
First and second order autoregressive models for daily Melbourne rainfall

Model	$\hat{D}$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\mu}$	AIC
Constant mean AR(1)	2	0.281	0	0.388	4595.7
Constant mean AR(2)	2	0.272	0.033	0.388	4591.7
Time varying AR(1)	34	0.242	0	—	4534.2
Time varying AR(2)	34	0.242	0.0004	—	4535.7

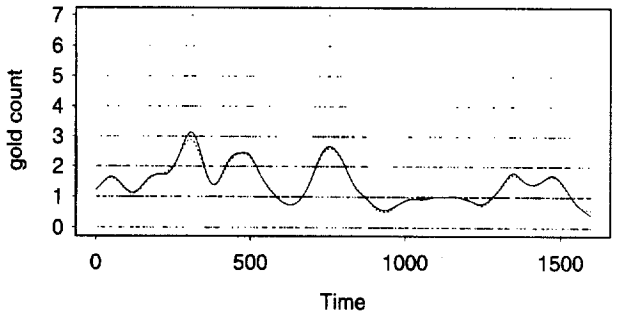

Fig. 15. Unweighted (dotted line) and weighted (solid line) spline smoothing for the gold series, both with 25 degrees of freedom.

superior, and there the second-order model is not as good as the first-order model. The description in Fig. 13 is still preferred.

6.2. *Variance weighted smoothing*

In this paper we have used linear smoothers to estimate $\boldsymbol{\mu}$. However, because of the nonnormal distributions, the variances of the observations typically change over time and depend on the mean function $\boldsymbol{\mu}$ and correlation structure. For uncorrelated non-Gaussian series, Hastie and Tibshirani (1990) and Green and Silverman (1994) show how weighted smoothers can be used to include these effects. For the time-varying mean models of Section 3.1, it is not difficult to derive the corresponding variances recursively. For instance, for the model of Example 1, if $\mathrm{Var}(Y_1) = \mu_1$ then $\mathrm{Var}(Y_t) = \mu_t$ for $t \geq 1$. Fig. 15 shows the spline smooth weighted with these variances as in Hastie and Tibshirani (1990) (solid line) and unweighted spline smooth (dotted line), both with $d_\mu = 25$. In examples we've tried, practical differences seem small. All of the methods given in this paper continue to apply if weighted smoothing is used.

Even for smoothing Gaussian series with correlation, it is not clear how to modify the smoothers to include correlation, as Hastie and Tibshirani (1990) and Green and Silverman (1994) remark. The effect of ignoring correlation is similar to that in regression, where in the presence of correlated errors estimates remain unbiased but are not fully efficient. For linear smoothers, the usual asymptotic arguments (under mild regularity conditions) give the Mean Square Error at time t as

$$\mathrm{MSE}(\hat{\mu}_t) \approx K(\mu_t^{(2)})^2 + \boldsymbol{s}'\Sigma(\boldsymbol{\mu})\boldsymbol{s},$$

where $\mu_t^{(2)}$ denotes the second derivative of μ_t with respect to t, $\boldsymbol{s}$ is the vector of weights for unweighted smoothing at time t, K is a constant depending on the smoothing matrix and the bandwidth, and $\Sigma(\boldsymbol{\mu})$ is the covariance matrix of $\boldsymbol{Y}$. The first term, representing the squared bias of $\hat{\mu}_t$, is unaffected by the correlation, even when it depends on $\boldsymbol{\mu}$.

6.3. Generalized additive models

In some cases, for instance Examples 2 and 3, the model (7) is of the form of a semi-parametric Generalized Additive Model (GAM) of Hastie and Tibshirani, 1990. We are presently exploring the possibility of fitting the model in this form using the backfitting algorithm for GAMs. The methods in this paper would still be useful for bandwidth selection.

6.4. Link functions

In the Generalized Linear Models literature, link functions are often used with non-Gaussian data. Zeger and Qaqish (1988) suggested this for autoregressive models and studied the resulting models, some of which do not have desirable properties. In some cases, such as Example 1, the physical setting clearly suggests that no link function be used. In general, the case for link functions in non-Gaussian autoregressive models is unresolved. If link functions were used, most of the methods in this paper would continue to apply, though smoothers would be modified as in Green and Silverman (1994), and the interpretation of the ACF would be more difficult.

Acknowledgements

This work was supported by Australian Research Council grant No. SG6935488. The authors are grateful to Kais Hamza and Geoff Laslett for several helpful discussions.

References

Al-Osh, M.A., Alzaid, A.A., 1987. First-order integer-valued autoregressive (INAR(1)) process. J. Time Series Anal. 8, 261–275.

Aly, A.A., Bouzar, N., 1994. Explicit stationary distributions for some Galton-Watson processes with immigration. Commun. Statist. – Stochastic Models 10, 499–517.

Altman, N.S., 1990. Kernel smoothing of data with correlated errors. J. Amer. Statist. Assoc. 85, 749–759.

Billingsley, P., 1961. Statistical Inference for Markov Processes. The University of Chicago Press, Chicago.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, 2nd ed. Springer, New York, 1991.

Chandrasekhar, S., 1954. Selected papers on noise and stochastic processes. In: Wax, N. (Ed.), Stochastic Problems in Physics and Astronomy. Dover, New York.

Chen, C., Davis, R.A., Brockwell, P.J., Bai, Z.D., 1993. Order determination for autoregressive processes using resampling methods. Statist. Sinica 3, 481–500.

Cleveland, W.S., Grosse, E., Shyu, W.M., 1992. Local regression models. In: Chambers, J.M., Hastie, T.J. (Eds.), Statistical Models in *S*. Wadsworth and Brooks, Pacific Grove, CA.

Diggle, P.J., Hutchinson, M.F., 1989. On spline smoothing with autocorrelated errors. Aust. J. Statist. 31, 166–182.

Fahrmeir, L., 1992. State space modeling and conditional mode estimation for categorical time series. In: Brillinger, D., Caines, P., Geweke, J., Parzen, E., Rosenblatt, M., Taqqu, M.S. (Eds.), New Directions in Time Series Analysis, Part 1. Springer, New York.

Gaver, D.P., Lewis, P.A.W., 1980. First order autoregressive gamma sequences and point processes. Adv. Appl. Probab. 12, 727–745.

Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, Chapman & Hall, London.

Grunwald, G.K., Feigin, P.D., 1995. First order product autoregressive processes. Mimeo, Dept. of Statistics, The University of Melbourne, Parkville, VIC.

Grunwald, G.K., Hyndman, R.J., Tedesco, L., Tweedie, R.L., 1997. A unified view of linear AR(1) models. Mimeo, Dept. of Mathematics and Statistics, Monash University, Clayton, VIC.

Guttorp, P., 1991. Statistical Inference for Branching Processes. Wiley, New York.

Hart, J.D., 1991. Kernel regression estimation with time series errors. J. Roy Statist. Soc. Ser. B, 53, 173–187.

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

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

Hastie, T.J., Loader, C., 1993. Local regression: automatic kernel carpentry (with discussion). Statist. Sci. 8, 120–143.

Herrmann, E., Gasser, T., Kneip, A., 1992. Choice of bandwidth for kernel regression when residuals are correlated. Biometrika 79, 783–795.

Heyde, C.C., Seneta, E., 1972. Estimation theory for growth and immigration rates in a multiplicative process. J. Appl. Probab 9, 235–256.

Hutton, J.L., 1990. Non-negative time series models for dry river flow. J. Appl. Probab 27, 171–182.

Hyndman, R.J., Wand, M.P., 1997. Nonparametric autocovariance function estimation. Aust. J. Statist, 39, 313–324.

Jacobs, P.A., Lewis, P.A.W., 1978. Discrete time series generated by mixtures I: correlational and runs properties. J. Roy Statist. Soc. Ser. B 40 (1978) 94–105.

Kashigawi, N., Yanagimoto, T., 1992. Smoothing serial count data through a state space model. Biometrics 48, 1187–1194.

Kitagawa, G.C., 1987. Non-Gaussian state-space modeling of nonstationary time series (with discussion). J. Amer. Statist. Assoc. 82, 1032–1063.

Kohn, R., Ansley, C.F., Wong, C.M., 1992. Nonparametric spline regression with autoregressive moving average errors. Biometrika 79, 335–346.

Lampard, D.G., 1968. A stochastic process whose successive intervals between events form a first order Markov chain – I. J. Appl. Probab 5, 648–668.

Li, W.K., 1991. Testing model adequacy for some Markov regression models for time series. Biometrika 78, 83–89.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. 2nd ed. Chapman & Hall, London.

McKenzie, E., 1985. Some simple models for discrete variate time series. Water Resources Bult. 21, 645–650.

McKenzie, E., 1986. Autoregressive moving-average processes with negative-binomial and geometric narginal distributions, Adv. Appl. Probab 18, 679–705.

McKenzie, E., 1988. Some ARMA models for dependent sequences of Poisson counts, Adv. Appl. Probab 20, 822–835.

Parzen, E., 1962. Stochastic Processes. Holden-Day, San Francisco.

Sim, C.H., 1994. Modelling non-normal first-order autoregressive time series. J. Forecasting 13, 369–381.

Stern, R.D., Coe, R., 1984. A model fitting analysis of daily rainfall data (with discussion). J. Roy Statist. Soc. Ser. A 147, 1–34.

Tsay, R.S., 1992. Model checking via parametric bootstraps in time series analysis. Appl. Statist. 41, 1–15.

van der Linde, A., 1994. On cross-validation for smoothing splines in the case of dependent observations, Austral. J. Statist. 36, 67–73.

Venkataraman, K.N., 1982. A time series approach to the study of the simple subcritical Galton-Watson process with immigration. Adv. Appl. Probab 14, 1–20.

West, M., Harrison, P.J., Migon, H.S., 1985. Dynamic generalized linear models and Bayesian forecasting (with discussion). J. Amer. Statist. Assoc. 80, 73–96.

Westgren, A., 1916. Die Veranderungsgeschwindigkeit der lokalen Teilchenkonzen-tration in kolloiden Systemen (Erste Mitteilung). Ark. Mat. Astron. Fys. 11, 1–25.

Zeger, S.L., Qaqish, B., 1988. Markov regression models for time series: a quasi-likelihood approach. Biometrics 44, 1019–1031.