ISSN 1032-3813
ISBN 0 7326 1022 2

MONASH UNIVERSITY

AUSTRALIA

ADDITIVE NONPARAMETRIC REGRESSION WITH AUTOCORRELATED ERRORS

Michael Smith
Chi-Ming Wong
Robert Kohn

Working Paper 19/96
November 1996

DEPARTMENT OF ECONOMETRICS

Additive nonparametric regression with autocorrelated errors*

Michael Smith[a], Chi-Ming Wong[b] and Robert Kohn[c]
[a]Department of Econometrics, Monash University.
[b]Department of Information and Systems Management,
The Hong Kong University of Science and Technology.
[c]Australian Graduate School of Management,
University of New South Wales.

November 27, 1996

Summary

A Bayesian approach is presented for nonparametric estimation of an additive regression model with autocorrelated errors. Each of the potentially nonlinear components is modelled as a regression spline using many knots, while the errors are modelled by a high order stationary autoregressive process parameterised in terms of its autocorrelations. The distribution of significant knots and partial autocorrelations is accounted for using subset selection. Our approach also allows the selection of a suitable transformation of the dependent variable. All aspects of the model are estimated simultaneously using Markov chain Monte Carlo. It is shown empirically that the proposed approach works well on a number of simulated and real examples.

KEY WORDS: Autoregressive model; Bayesian analysis; Data transformation; Markov chain Monte Carlo; Regression spline; Subset selection.

*This is part of Michael Smith's PhD Thesis

1 Introduction

When a regression model is fitted to time series data the errors are likely to be autocorrelated, such as in the problems tackled by Engle, Granger, Rice and Weiss (1986) and Harvey Koopman (1993). Few approaches are currently available for estimating a regression model nonparametrically when the errors are autocorrelated, despite the fact that failure to take account of the autocorrelation can result in poor function estimates; see Section 4.1 and Altman (1990) for simulation evidence. Those authors that allow for autocorrelation in the errors usually only deal with univariate nonparametric regression with time as the independent variable; e.g. Altman (1990), Chu and Marron (1991) and Hart (1991, 1994). These estimators do not generalise to the case where the independent variable is not in time order.

This paper presents a comprehensive Bayesian approach for semiparametrically estimating an additive regression model when the errors are autocorrelated. Each potentially nonlinear component is modelled as a regression spline with many knots and the errors are modelled as a stationary autoregression parameterised by its partial autocorrelations. The distribution of significant knots in the regression spline and significant partial autocorrelations is accounted for by subset selection. The Bayesian analysis also allows a suitable transformation to be chosen for the dependent variable. The entire model is estimated simultaneously using Markov chain Monte Carlo.

To the best of our knowledge, even those papers that deal with nonparametric regression when the errors are correlated do not consider issues such as selecting the appropriate model for the errors and the transformation of the dependent variable. Furthermore, the approach in this paper can be made robust to outliers and can accommodate missing values of the dependent variable as in Barnett, Kohn and Sheather (1996a). It can also be extended to bivariate surface estimation, as is demonstrated in Smith and Kohn (1996b).

This paper links two lines of research. The first is by Smith and Kohn (1996a) who combine regression splines with Bayesian subset selection to nonparametrically estimate an additive regression model with independent errors. They show that in the univariate case their

approach acts as a variable bandwidth smoother and compares favourably with modern kernel weighted local linear smoothing. The second line of research is by Barnett *et. al.* (1996a) who propose a Bayesian approach for robustly estimating an autoregressive model, simultaneously choosing the order of the model and estimating its parameters and any missing observations. We note that the work of Smith and Kohn (1996a) is motivated by the Bayesian subset selection paper of George and McCulloch (1993), while Barnett *et. al.* (1996a) refine and extend the work of McCulloch and Tsay (1994).

The paper is organised as follows. Section 2 describes the model and the prior assumptions and Section 3 discusses estimation and the Markov chain sampler. Section 4 studies in detail the performance of the nonparametric estimator in the univariate case and compares it to previous estimators. Section 5 considers multiple regression examples and Appendix 1 shows how to implement the sampler.

2 Model and prior assumptions

2.1 Autoregressive model for the errors

Suppose

$$y_t = f(x_t) + u_t\,, \quad t = 1, \ldots, n, \tag{2.1}$$

where y_t is the dependent variable, $f(x_t)$ is an unknown regression function of the independent variable x_t, and u_t is a stationary autocorrelated error sequence. The errors are modelled by the zero mean stationary autoregressive process of maximal order s,

$$u_t = \theta_1 u_{t-1} + \cdots + \theta_s u_{t-s} + e_t\,,$$

where e_t is independent $N(0, \sigma^2)$. There is little loss of generality in this assumption as most Gaussian stationary processes can be approximated by an autoregressive process of sufficiently high order. Moreover, it is straightforward to adapt the methods of the present paper to handle autoregressive-moving-average errors as in Barnett, Kohn and Sheather (1996b).

Let ψ_i be the ith partial autocorrelation of u_t, so that $-1 < \psi_i < 1$ for $i = 1, \ldots, s$ and $\psi_i = 0$ for $i > s$. We note that the partial autocorrelations $\psi_1, \ldots, \psi_s$ are a one-to-one transformation of $\theta_1, \ldots, \theta_s$. As in Monahan (1984) and Barnett *et. al.* (1996a) it is convenient to enforce stationarity by re-parameterising u_t in terms of $\psi = (\psi_1, \ldots, \psi_s)$. When x_t is time it is important to enforce stationarity of u_t so as not to confound the model for the errors with the nonparametric estimate of the function. For example, a random walk on the errors, $u_t = u_{t-1} + e_t$, acts as first order spline smoother.

As in Barnett *et. al.* (1996a), the following prior assumptions are made on σ^2 and ψ.

A1. $\log \sigma^2$ has a flat prior on the line, so that $p(\sigma^2) \propto 1/\sigma^2$. This is a commonly used prior for σ^2.

A2. Let κ_i be a binary variable determining the status of ψ_i. If $\kappa_i = 0$ then ψ_i is identically zero; if $\kappa_i = 1$ then ψ_i is uniformly distributed on $(-1, 1)$. We assume that, *a priori*, the ψ_i are independent of each other. This ensures that the u_t are both stationary and parsimoniously parameterised and that the prior distribution of $\psi | \kappa_i = 1$ is non-informative. We also assume that ψ_i is *a priori* independent of σ^2.

The maximal order s of the autoregression and the probabilities $p(\kappa_i = 1), i = 1, \ldots, s$, are prescribed by the user. In all our examples we take a descending prior on the order, where $p(\kappa_1 = 1) = 0.5, p(\kappa_2 = 1) = 0.4, p(\kappa_3 = 1) = 0.3, p(\kappa_4 = 1) = 0.2$, and $p(\kappa_i = 1) = 0.1$ for $i = 5, \ldots, s$. However, the results prove reasonably insensitive to the exact specification of $p(\kappa_i = 1)$.

2.2 Regression splines

The nonlinear regression function is assumed to be smooth and modelled (approximated) by the regression spline

$$b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \sum_{i=1}^{m} b_{i+3} (x - \tilde{x}_i)_+^3 , \tag{2.2}$$

where $\tilde{x}_1, \ldots, \tilde{x}_m$ are m knots placed along the domain of the independent variable x, such that $\min_t(x_t) < \tilde{x}_1 < \tilde{x}_2 < \cdots < \tilde{x}_m < \max_t(x_t)$ and $z_+ = \max(0, z)$. When $f(x)$ is observed with noise, the two most important problems in approximating $f(x)$ by (2.2) are how many knots to use and where to place them. If too few knots are used, or they are badly placed, then important features of the curve may be missed. If too many knots are used then the estimate of f will have high local variance. Smith and Kohn (1996a) solve this problem in the independent error case by introducing many knots from which significant knots are selected. We show how to extend their approach to the autocorrelated error case by rewriting (2.1) as a linear model.

Let $r = m + 3$, $\beta = (b_0, \ldots, b_r)'$ be a vector of regression coefficients, $y = (y_1, \ldots, y_n)'$ be a vector of observations on the dependent variable, $\mathrm{x} = (x_1, \ldots, x_n)'$, and let 1 be a $n \times 1$ vector of ones. Let x^j and $(\mathrm{x} - 1\tilde{x}_j)_+^3$ be the $n \times 1$ vectors with ith elements x_i^j and $(x_i - \tilde{x}_j)_+^3$ respectively, and define the $n \times (r+1)$ design matrix $X = \{1, \mathrm{x}, \mathrm{x}^2, \mathrm{x}^3, (\mathrm{x} - 1\tilde{x}_1)_+^3, \ldots, (\mathrm{x} - 1\tilde{x}_m)_+^3\}$. By replacing $f(x)$ by (2.2), equation (2.1) can be expressed as the linear regression model

$$y = X\beta + u, \tag{2.3}$$

where $u = (u_1, \ldots, u_n)'$ is the vector of autoregressive errors. In this linear model context, selecting significant knots is equivalent to selecting significant variables in the linear regression model (2.3).

Let $\mathrm{var}(u) = \sigma^2\Omega_\psi$ be the variance matrix of u. To carry out subset selection on the columns of X, it is convenient to define the binary variables $\gamma_i, i = 0, \ldots, r$, determining which columns of X are in the regression. Let $\gamma_i = 0$ if b_i is identically zero and let $\gamma_i = 1$ otherwise. Put $\gamma = (\gamma_0, \ldots, \gamma_r)'$ and let β_γ and X_γ be the sub-vector of β and the sub-matrix of X, respectively, corresponding to the nonzero elements of γ. Given σ^2, ψ, and κ, we place the following prior on γ and β.

A3. The γ_i are independent *a priori* with $p(\gamma_i = 1) = 1/2$, while the conditional prior for

β_γ is

$$\beta_\gamma|\sigma^2,\gamma,\psi \sim N\left(0, c\sigma^2(X_\gamma'\Omega_\psi^{-1}X_\gamma)^{-1}\right).$$

The prior for γ means that we have no prior information on which columns of X to include. We usually take $c = 100$ as it works well in practice and makes the prior for β_γ almost diffuse relative to the information in the likelihood. We have checked using simulation on many examples that estimates based on this prior are relatively insensitive to values of c in the range $10 \leq c \leq 2000$. The prior for β_γ is similar to Zellner's (1986) g-prior when $\Omega_\psi = I$.

2.3 Data transformation

For a linear regression model with independent errors, Box and Cox (1964) show how to use a power transformation of the dependent variable to obtain an additive model with errors that are Gaussian and have a constant variance. Because our approach assumes an additive regression model with stationary errors, we make the choice of an appropriate transformation part of the Bayesian analysis. We consider a family of transformations $T_\lambda(y)$ indexed by $\lambda \in \Lambda$ such that $T_\lambda(y)$ is monotonically increasing in y for each $\lambda \in \Lambda$. An example is the family of power transformations $T_\lambda(y) = y^\lambda$ for $\lambda > 0$, $T_\lambda(y) = \log(y)$ for $\lambda = 0$, and $T_\lambda(y) = -y^\lambda$ for $\lambda < 0$.

Taking a transformation of y can make both the location and scale of $T_\lambda(y)$ different for each λ. Therefore, in a linear regression model Box and Cox (1964) advocate making the prior for β and σ^2 depend on λ. We take a different approach. Instead of working directly with $T_\lambda(y)$ we follow Smith and Kohn (1995) and work with the normalised transformation $y_\lambda = a_\lambda + b_\lambda T_\lambda(y)$. For each λ, the scalars a_λ and b_λ are chosen as follows. Let $y_{(i)}$ and $y_{(i),\lambda}$ be the ith ordered values of y_i and $y_{i,\lambda} = a_\lambda + b_\lambda T_\lambda(y_i)$. The scalars a_λ and b_λ are chosen so that $y_{(n/2)} = y_{(n/2),\lambda}$ and $y_{(3n/4)} - y_{(n/4)} = y_{(3n/4),\lambda} - y_{(n/4),\lambda}$. This means that the median and interquartile range of the $y_{i,\lambda}$ are approximately the same for all $\lambda \in \Lambda$. Such a choice of a_λ and b_λ is motivated by, but is a little different to, the transformation approach

of Emerson and Stoto (1983).

Because the centre and the scale of the $y_{i,\lambda}$ are the same for all λ we make the following assumption.

A4. The parameters $\sigma^2, \kappa, \psi, \gamma$, and β are *a priori* independent of λ.

We also limit the set of possible transformations Λ to be discrete and small because it is necessary to integrate out λ in some steps of the Markov chain Monte Carlo sampler in Section 3.

3 Sampling scheme and parameter estimation

3.1 Sampling scheme

The complexity of the Bayesian model means that it is necessary to use Markov chain Monte Carlo sampling to estimate the regression function and the autoregressive parameters. We refer the reader to Gelfand and Smith (1990) and Tierney (1994) for the application of Markov chain Monte Carlo sampling in statistics. To describe the sampling scheme it is convenient to use the notation $\alpha|\beta$ to mean that α is generated conditional on β. Some steps of the sampler generate from the exact conditional distribution which we write as $p(\alpha|\beta)$. Other steps generate from an approximation to the conditional distribution combined with a Metropolis-Hastings step (see Tierney, 1994). We write the approximation to the conditional distribution as $q(\alpha|\beta)$.

The sampling scheme is first presented and each of its steps is then briefly described. Implementation details are given in Appendix 1. Let $\psi = (\psi_1, \ldots, \psi_s)$, $\kappa = (\kappa_1, \ldots, \kappa_s)$, and $\gamma = (\gamma_0, \ldots, \gamma_r)$.

Sampling Scheme: Starting with some initial values $\kappa^{[0]}, \psi^{[0]}, \beta^{[0]}, \gamma^{[0]}$, and $\lambda^{[0]}$, the sampling scheme iteratively generates the parameters using the following conditioning: (i) $\beta_\gamma|\kappa, \psi, \gamma, \lambda, y$;

(ii) $\psi_i, \kappa_i | \kappa_{j \neq i}, \psi_{j \neq i}, \gamma, \beta, \lambda, y$, for $i = 1, \ldots, s$; (iii) $\gamma_i | \gamma_{j \neq i}, \kappa, \psi, y$, for $i = 0, \ldots, r$; (iv) $\lambda | \kappa, \psi, \gamma, y$.

The sampling scheme is invariant to the posterior distribution $p(\beta, \gamma, \psi, \kappa, \lambda | y)$ as each part either generates directly from a conditional distribution or uses a Metropolis-Hastings step. It can be readily checked that the sampling scheme is also irreducible and aperiodic. Therefore, by Tierney (1994) the sampler converges to the correct posterior distribution.

The error variance σ^2 is integrated out of the sampling scheme. In Step (i), β_γ is generated from its conditional distribution $p(\beta_\gamma | \kappa, \psi, \gamma, \lambda, y)$ which is multivariate t. In Step (ii), ψ_i and κ_i are generated as a block. The binary indicator κ_i is generated first from $p(\kappa_i | \kappa_{j \neq i}, \psi_{j \neq i}, \gamma, \beta, \lambda, y)$ using numerical integration to determine the conditional probability that $\kappa_i = 1$. The partial autocorrelation ψ_i is then generated using a normal approximation to its conditional distribution; we use the approximation
$q(\psi_i) \propto p(y_{s+1}, \ldots, y_n | y_1, \ldots, y_s, \kappa, \psi, \gamma, \beta, \lambda)$ which is $t-$distributed in ψ_i. The generation of ψ_i is completed using a Metropolis-Hastings step. It is necessary to generate ψ_i and κ_i simultaneously as generating them one at a time produces a reducible sampling scheme which does not converge. In Step (iii), the binomial density $p(\gamma_i | \gamma_{j \neq i}, \kappa, \psi, y)$ is obtained explicitly, by evaluating the probability that $\gamma_i = 0$ and 1 up to a scale factor and then normalising. The transformation parameter λ is generated from its multinomial conditional probability as for each $\lambda \in \Lambda$, the conditional probability of λ can be evaluated up to a scale factor and then normalised.

The family of transformations Λ is taken discrete and small to make it easy to integrate λ out when generating γ_i and also to generate λ. It is necessary to integrate λ out when generating γ as the sampler which conditions γ on λ converges too slowly to be practical.

In general, the sampler is first run for a warmup period at the end of which it is assumed that it has converged to the posterior distribution $p(\beta, \gamma, \kappa, \psi, \lambda | y)$. It is then run for a further period called the sampling period whose output is used for estimation. Let $\beta^{[k]}, \gamma^{[k]}, \kappa^{[k]}, \psi^{[k]}, \lambda^{[k]}$, $k = 1 \ldots, K$, be the iterates of $\beta, \gamma, \kappa, \psi, \lambda$ during the sampling period.

3.2 Estimation

Estimation is done in two stages. The parameter λ is first estimated as the mode of an estimate of $p(\lambda|y)$; let $\hat{\lambda}_M$ be this estimate of the mode. In the second stage the unknown regression function and the autoregressive parameters are estimated conditional on $\hat{\lambda}_M$. This two stage approach is used by Box and Cox (1964, 1982) who advocate estimating the regression on a given scale (that is by conditioning on the estimate of λ) rather than averaging over the distribution of λ.

The modal estimate $\hat{\lambda}_M$ is obtained as follows. The Markov chain sampler in Section 3.1 is run to obtain the iterates $\beta^{[k]}, \gamma^{[k]}, \kappa^{[k]}, \psi^{[k]}, \lambda^{[k]}$, $k = 1 \ldots, K$. Based on these iterates the probability $p(\lambda|y)$ is estimated by

$$\hat{p}(\lambda|y) = \frac{1}{K} \sum_{k=1}^{K} p(\lambda|\beta^{[k]}, \gamma^{[k]}, \kappa^{[k]}, \psi^{[k]}, y) \tag{3.1}$$

The estimate $\hat{\lambda}_M$ is the maximum of $\hat{p}(\lambda|y)$ for $\lambda \in \Lambda$.

The Markov chain Monte Carlo sampler is run again, this time conditioning on $\hat{\lambda}_M$, and not executing Step (iv) of the sampling scheme. New iterates $\beta^{[k]}, \gamma^{[k]}, \kappa^{[k]}, \psi^{[k]}$ are obtained. The following estimates of $E(\beta|y), E(\psi|y)$, and $E(\kappa_i|y) = p(\kappa_i = 1|y)$ are used to estimate the regression function and the autoregressive parameters;

$$\begin{aligned}
\hat{\beta} &= \frac{1}{K} \sum_{k=1}^{K} E(\beta|\gamma^{[k]}, \psi^{[k]}, \kappa^{[k]}, \hat{\lambda}_M, y)\,, \\
\hat{\psi} &= \frac{1}{K} \sum_{k=1}^{K} \psi^{[k]}\,, \\
\hat{\kappa}_i &= \frac{1}{K} \sum_{k=1}^{K} p(\kappa_i = 1|\psi_{j\neq i}^{[k]}, \kappa_{j\neq i}^{[k]}, \hat{\lambda}_M, y)\,, \quad i = 1, \ldots, s\,.
\end{aligned}$$

The estimates $\hat{\beta}$ and $\hat{\kappa}$ are called mixture estimates as they are based on the conditional distributions of β and κ. The estimate $\hat{\psi}$ is called an empirical estimate as it is based on the raw iterates $\psi^{[k]}$. The regression function is estimated by plugging $\hat{\beta}$ into (2.2). The estimate $\hat{\theta}$ of the autoregressive parameter θ is obtained from $\hat{\psi}$.

4 Univariate nonparametric regression

4.1 Independent variable not time

In this section we show by simulation that taking account of the autocorrelation in the errors can give substantially more accurate nonparametric estimates than those obtained if the autocorrelation is ignored. Previous simulation studies by Diggle and Hutchinson (1989) and Altman (1990) deal only with the univariate case where time is the independent variable. They show that if the errors are treated as independent then any autocorrelation in the errors is incorporated into the nonparametric estimate of the regression function and can result in a very rough estimate. What is insufficiently discussed in the literature is that, even if the independent variable is not time, modelling the autocorrelation in the errors gives more efficient nonparametric estimates as it reduces the effective error variance. This is illustrated by the following example. Consider the model (2.1) with u_t the first order stationary Gaussian autoregressive process $u_t = \phi u_{t-1} + e_t$, with e_t independent $N(0, \sigma^2)$ and ϕ known. Then, $\text{var}(u_t) = \sigma^2/(1-\phi^2)$. Equation (2.1) can be written as

$$y_t = \phi y_{t-1} + f(x_t) - \phi f(x_{t-1}) + e_t$$

which is an additive regression with error variance equal to σ^2. This shows that ignoring the structure in the errors increases the error variance by a factor of $1/(1-\phi^2)$ which is large for ϕ close to 1.

This gain in efficiency is illustrated by simulation. One hundred observations were generated from the model (2.1) for each of the three functions

$$f_1(x) = 2x-1, f_2(x) = \sin(10\pi x), \quad \text{and} \quad f_3(x) = \{\phi(x; 0.15, 0.05) + \phi(x; 0.6, 0.2)\}/4\,; \quad (4.1)$$

$\phi(x; \mu, \sigma)$ is the Gaussian density, with mean μ and standard deviation σ, evaluated at x. The independent variable x_t was generated uniformly on $(0, 1)$, and the errors u_t were generated from the second order autoregression $u_t = 0.9u_{t-1} - 0.9u_{t-2} + e_t$, with $e_t \sim N(0, 0.5^2)$. This was replicated one hundred times, and for each function and each replication three

nonparametric Bayesian estimators of the regression function were computed. The first estimator fits a second order autoregression to the errors with κ_1 and κ_2 fixed at 1 and $s = 2$. That is, the error structure is assumed to be known except for the values of the parameters θ_1, θ_2, and σ^2. The second estimator fits an autoregression of order $s = 6$ to the errors and selects the significant partial autocorrelations. The third estimator treats the errors as independent.

For all three estimators the knots are selected to follow the density of the independent variable, with one knot every fifth observation of the ranked predictor. Extensive simulations by the authors, some of which appear in Smith and Kohn (1996a), indicate that such a scattering of potential knot sites is more than adequate to capture all but the most oscillatory function. Each replicate for each function and each estimate consisted of 500 iterations with the first 100 iterations discarded to ensure convergence.

The numerical criterion we use for judging the quality of an estimate $\hat{f}(x)$ of an unknown function $f(x)$ is the integrated squared error (ISE)–the integral of $\{f(x) - \hat{f}(x)\}^2$ over the domain of $f(x)$. We approximate this integral by taking a grid of 400 equally spaced points $z_i = i/400, i = 1, \ldots, 400$, and compute the ISE as

$$ISE = \frac{1}{400} \sum_{i=1}^{400} \{f(z_i) - \hat{f}(z_i)\}^2 .$$

Figure 1 presents the boxplots of $\log_e(ISE)$ for the three estimators and shows the substantial loss in efficiency when the autocorrelation in the errors is ignored. This is because the standard deviation of u_t is 2.6 times the standard deviation of e_t. The plots also show that the second estimator, which does not assume that the order of the autoregression is known, performs almost as well as the first estimator which assumes that the order is known.

To visually assess the performance of the full estimator (that is, where the autoregressive order is also estimated), we sorted the ISE scores from highest to lowest. We selected the $10th$, $51st$, and $90th$ highest scores as examples of poor, median and good fits for each of the three functions. These estimates, the actual function and the corresponding data are plotted in Figure 2. Also plotted is the nonparametric estimate which assumes that

the errors are independent. The plots confirm that when autocorrelation in the errors is ignored the function estimates are very poor, whereas the estimate that takes account of the autocorrelation performs well.

To ensure that 100 iterations suffice to ensure convergence, the output of a number of individual runs was studied. Convergence, as measured by the value of the iterates of the posterior density $p(\gamma^{[k]}, \kappa^{[k]}, \psi^{[k]}, \beta^{[k]}|y)$ consistently occurred within two dozen iterations, with the same estimates of the function obtained using a number of different starting values. In addition, the plots in Figure 2 show the high quality of the fits obtained for the second estimator. If the schemes had not converged to the correct joint posterior distribution these plots would either be highly biased (when important knots are omitted) or have a high variance (when redundant knots are retained).

4.2 Time as the independent variable

Much of the literature on nonparametric regression with autocorrelated errors deals exclusively with the univariate case with time as the independent variable, e.g. Diggle and Hutchinson (1989), Altman (1990), Hart (1991, 1994), and Kohn, Ansley, and Wong (1994). We therefore study by simulation the performance of the Bayesian regression spline estimator for this case. For comparison, we also look at the smoothing spline estimators in Kohn, Ansley and Wong (1992) and the kernel approach by Hart (1994). Kohn, Ansley, and Wong (1992) estimate the smoothing parameter and the autoregressive parameters by two methods, marginal likelihood and generalised cross-validation. Hart (1994) estimates the bandwidth by what he calls time series cross validation. We refer the reader to Kohn, Ansley and Wong (1992), and Hart (1994) for a description of their methods.

One hundred observations were generated from the model (2.1) using the following three functions: the first function is $f_1(x) = 32x^2(1-x)^2$ which is used by Hart (1991) in his simulations; the other two functions $f_2(x)$ and $f_3(x)$ are described in (4.1). The independent variable was set to be $x_t = t/100, t = 1, \ldots, 100$. The error u_t is the first order autoregression

$u_t = 0.5u_{t-1} + e_t$, with e_t independent $N(0, \sigma^2)$. For each function, σ takes three values corresponding to the standard deviation of e_t being one eighth, one quarter, and one half the range of the function. These three values for the standard deviation of e_t represent low, medium, and high noise examples.

One hundred replications were run for the four estimators, the three functions and the three noise levels and the performance of the estimators compared using ISE. To ensure a fair comparison all four estimators had the order of the autocorrelation fixed to the true value, so that in the Bayesian case $\kappa_1 = 1$ and $\kappa_i = 0$ for $i > 1$. Figure 3 presents boxplots of $\log(ISE)$ and shows that the Bayesian regression spline estimator compares favourably to the other estimators over all three functions and all three noise levels. It 'breaks down' (that is, confuses the function estimate with the autoregressive process) far less frequently than the other three estimators.

In particular, the regression spline estimator performs much better for the function f_3. This function has differing curvature over the domain of x and requires an estimator with a degree of 'local adaptability', such as a variable bandwidth smoother. The smoothing spline and the kernel based estimators are single bandwidth smoothers and therefore are not locally adaptive in nature and perform poorly. This is illustrated in Figure 4 which plots, for each estimator, the function estimate corresponding to the 51st worst value of the ISE together with the corresponding data set. The plots are for the low noise case. The plot shows that the regression spline produces a relatively unbiased and smooth estimate for the entire function, despite the fact that the curvature is much greater on the left side than the right side of the domain of x. The estimate of θ_1 is 0.646, close to the true value of $\theta_1 = 0.5$.

The smoothing spline estimate using generalised cross-validation uses the same bandwidth throughout. This value is too small for the function on the right hand side of the x domain and results in under-smoothing of the function. The estimate of θ_1 is poor, with $\hat{\theta}_1 = -0.076$, because much of the autocorrelation process is identified as function curvature and incorporated into the function estimate. The smoothing spline estimate using marginal likelihood grossly over-smoothes as it incorporates the entire shape of the curve into the

autocorrelation process with the estimate $\hat{\theta}_1 = 0.942$. Hart's estimate tends to interpolate the data by taking too small a bandwidth. The estimate of θ_1 is a non-stationary -3.45, while the bandwidth estimate is 0.1, which is the minimum allowed. This simulation underscores the need for both the autocorrelation process and the underlying function to be considered simultaneously. If one is incorrectly identified then the other is also likely to be poorly estimated.

4.3 Smoothing issues

In some sense, the choice of $c = 100$ in the prior for β_γ found at assumption A3 is arbitrary. However, we would like to stress that in the nonparametric application examined in this paper, it is not interpretable as a traditional smoothing parameter analogous to that of smoothing splines or kernel smoothers. For example, notice that $c = 100$ was an appropriate value for the four very different functional forms found in the simulations of sections 4.1 and 4.2. It is also appropriate for the two multivariate data examples examined later in section 5.

To demonstrate the high degree of insensitivity of the procedure to even quite large alterations in the value of c we conducted a small simulation experiment. Again, ISE was used as a performance measure and the design was the same as used in Section 4.1 with AR(2) errors where the order is known *a priori*. The functions f_1, f_2 and f_3, three levels of σ, $1/8, 1/4$ and $1/2$ and three levels of c, $c = 100, 500$, and 1000 were considered. As before, a regression spline as described in Section 4.1 was fit. The results are plotted in Figure 5 which present the $\log_e(ISE)$ averaged over 50 replications for each of the functions, for the three levels of c and the three levels of σ.

The plots illustrate that the performance of the regression spline is relatively insensitive for a range of values of c. This is true regardless of function type and signal to noise ratio.

4.4 Related Work

Engle *et. al.* (1986) and Harvey and Koopman (1993) are two of the very few papers that discuss nonparametric regression when the errors are autocorrelated, but the independent variable is not in time order. Both papers only consider a first order autoregressive process for the errors, but do not estimate the autoregressive parameter simultaneously with the unknown function. The results in Section 4.1 show that, in general, substantially better function estimates are obtained if the estimation is simultaneous.

To make the computation tractable, Engle *et. al.* (1986) use a discretised version of spline smoothing by dividing the range of the independent variable into subintervals and assume the regression function is constant in each subinterval. They do not obtain a data driven subdivision of the range of the independent variable.

Harvey and Koopman (1993) use regression splines to estimate the unknown function assuming that the errors are independent. The autoregressive parameter is then estimated from the residuals. Harvey and Koopman do not have a data driven approach for determining the position of the knots of the regression spline.

Chib and Greenberg (1994) use the Gibbs sampler to provide a Bayesian analysis of a regression model with autoregressive-moving-average errors. They do not carry out model selection for either the regression variables or the autoregressive moving average parameters.

5 Additive semiparametric regression

5.1 Introduction

Consider the following additive regression model with independent variables x and z,

$$y_t = f(x_t) + g(z_t) + u_t \tag{5.1}$$

When the errors u_t are independent, the backfitting algorithm described by Hastie and Tibshirani (1990, p. 90) is a popular approach for estimating f and g. Starting from some initial estimate of g, the backfitting algorithm iteratively estimates f given g and then g given f until convergence is achieved. Any univariate smoothers can be used to carry out the iteration. Hastie and Tibshirani (1990, p. 122) show that the backfitting algorithm converges for a class of smoothers including spline smoothers; no proof is yet available for kernel based smoothers or supersmoother which is used by the ACE algorithm of Breiman and Friedman (1985).

However, at present no backfitting algorithm exists for an additive model with autocorrelated errors and the following example suggests that it will be difficult to obtain one. Consider (5.1) with u_t generated by the first order autoregression $u_t = \phi u_{t-1} + e_t$, with ϕ known. Equation (5.1) is equivalent to

$$y_t - \phi y_{t-1} = f(x_t) + g(z_t) - \phi f(x_{t-1}) - \phi g(z_{t-1}) + e_t \,. \tag{5.2}$$

Let $v_t = x_{t-1}$, $w_t = z_{t-1}$, $f_1(x) = f(x)$, and $g_1(z) = g(z)$. Then (5.2) can be written as

$$y_t - \phi y_{t-1} = f(x_t) + g(z_t) - \phi f_1(v_t) - \phi g_1(w_t) + e_t \,. \tag{5.3}$$

By treating f, g, f_1 and g_1 as four separate functions the backfitting algorithm can be applied to (5.3) to give estimators of f and g. However, in empirical work using smoothing splines we have found this approach often gives poor estimators of f and g as the constraints $f = f_1$ and $g = g_1$ are not enforced. Moreover, it seems difficult to enforce the constraints. A second shortcoming of using backfitting, even in the independent error case, is that it seems difficult to obtain high quality estimates of the smoothing parameters involved. Usually the estimates of the smoothing parameters are chosen as functions of the independent variables, but not the dependent variable and hence cannot adequately take account of any curvature in the unknown functions.

By contrast, it is straightforward to extend the regression spline approach to handle an additive nonparametric regression model. Each nonlinear component is modelled as a regression spline, but where only a single global intercept is included in the model. This gives rise to a linear regression model in which the knots are chosen by subset selection in

the same manner as in the univariate case.

5.2 Electricity consumption data

As an illustrative example we apply the methodology to the residential electricity data in Harris and Liu (1993). The data consists of 264 consecutive observations of monthly electricity consumption (y) and the following four independent variables: number of heating degree days (x_1), number of cooling degree days (x_2), average real electricity price (x_3) and real disposable income (x_4). Plots of the variables are given by Harris and Liu, along with a detailed explanation of the data.

We model the data as the additive nonparametric regression

$$y_{t,\lambda} = \alpha + D(t) + f_1(x_{1t}) + f_2(x_{2t}) + f_3(x_{3t}) + f_4(x_{4t}) + u_t\,. \tag{5.4}$$

Here, α is the intercept and $D(t)$ is a trend. Each of the functions $D(t), f_1(x_{1t}), \ldots, f_4(x_{4t})$ is modelled as a regression spline with zero intercept and with a potential knot site every fifteenth observation of the ranked independent variable. Fewer potential knot sites are used for each function in the multivariate case than in the univariate case to prevent the matrix $X_\gamma'\Omega_\psi^{-1}X_\gamma$ becoming singular or nearly singular. Nevertheless, the number of knots used appears more than sufficient to capture any potential nonlinearity in the functions.

To allow for seasonality, the errors u_t are modelled as an autoregression with maximum order $s = 20$; the descending prior for the indicator variables κ_i is given in Section 2.1. Alternatively, to capture the seasonality, the errors u_t could be modelled as a multiplicative autoregressive model containing both seasonal and non-seasonal terms as in Barnett *et. al.* (1996a).

To obtain additivity of the regression function and normality of the intrinsic errors e, a normalised power transformation of the dependent variable is used as outlined in Section 2.3; i.e. $y_{t,\lambda} = a_\lambda + b_\lambda T_\lambda(y_t)$ with $T_\lambda(y) = y^\lambda$ and with λ restricted to the nine values $\Lambda = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$. Inclusion of a trend term was suggested by looking at

the residuals of the nonparametric fit of the model at (5.4) with the trend term omitted; see also Figure 7(b) for a plot of the original data.

The Bayesian approach was applied with a warmup period of 200 iterations and a further 100 iterations were used to estimate the posterior distribution of λ. The posterior probability of $\lambda = 0$ was 0.99 so the normalised logarithmic transformation was selected. Using this transformation of the dependent variable, the sampler was then run for a further 400 iterations to estimate the regression function and the autoregressive parameters.

Estimates of $p(\kappa_i = 1|y)$ and $E(\theta_i|y)$ are given in Table 1 and suggest that u_t is a twelfth order autoregression, which is consistent with monthly data. The estimate of the intercept $\hat{\alpha} = 0.4209$ while the estimates of $f_1, \ldots, f_4$ are plotted in Figures 6(a)–6(d). These suggest $\log(y)$ is linear in heating degree days and cooling degree days, but nonlinear in real electricity price and real disposable income. They also suggest that the two environmental predictors explain more of the variability in the dependent variable than the financial variables because f_1 and f_2 both have a greater range than either f_3 or f_4. The estimate of the time trend $D(t)$ is plotted in Figure 7(a) and is consistent with the plot of the dependent variable in Figure 7(b).

To explain the scatter plots in Figures 6(a)–6(d) and 7(a) we need some additional notation. For $t = 1, \ldots, n$, let $\xi_t = u_t - E(u_t|u_1, \ldots, u_{t-1})$, $\sigma^2 R_t = \mathrm{var}(\xi_t)$ and $\zeta_t = \xi_t/\sqrt{R_t}$. Then ζ_t is independent $N(0, \sigma^2)$ and $\zeta_t = e_t$ for $t \geq s + 1$. Let $\hat{\xi}_t$ and $\hat{R}_t$ be the estimates of ξ_t and R_t based on the estimate of ψ. If the fitted model is correct then $\hat{\zeta}_t = \hat{\xi}_t/\sqrt{\hat{R}_t}$ is approximately $N(0, \sigma^2)$ and independent. The scatter plot in Figure 6(a) is $\hat{f}_1(x_{1t}) + \hat{\zeta}_t$ against x_{1t} and suggests that the effect of x_{1t} on $\log(y_t)$ is captured correctly. The scatter plots in Figures 6(b)–6(d) and Figure 7(a) are interpreted similarly and suggest that the effects of x_{2t}, x_{3t}, x_{4t} and the time trend are also estimated correctly. Figure 6(e) is a plot of the sample autocorrelations of $\hat{\zeta}_t$ and Figure 6(f) is a normal probability plot of $\hat{\zeta}_t$. These plots indicate that ζ_t is independent and normal and thus validate the regression assumptions.

Figure 7(b) is a time plot of the dependent variable and the Bayesian fit showing that

the fitted values $\exp\{(\hat{y}_{t,\hat{\lambda}} - a_{\hat{\lambda}})/b_{\hat{\lambda}}\}$ track the data well on the original scale. Figure 7(c) plots the logarithm of the posterior density $\log p(\gamma^{[k]}, \psi^{[k]}, \kappa^{[k]}, \beta^{[k]}, \lambda^{[k]}|y)$, (up to an additive constant), for the first 200 iterations and suggests that the sampler converges quickly.

5.3 Toothpaste data

As a second example we consider the toothpaste data previously analysed by Wichern and Jones (1977). This data includes the market share and price of both Crest and Colgate over 276 consecutive weeks during the years 1958-1963. During that period the American Dental Association (ADA) publicly endorsed Crest between weeks 135 and 136. This intervention is modelled by the dummy variable $X_t = 1$ if $t > 135$ and $X_t = 0$ otherwise. We model the market share of Crest in terms of X_t and P_t = price of Colgate /price of Crest at time t. Because the market share of Crest is a fraction in the interval $(0, 1)$ we transform the dependent variable to help ensure that the regression assumption are satisfied. Nine candidate transformations $T_\lambda(y), \lambda = 1, \ldots, 9$, are considered and listed in Table 2 together with the normalisation constants a_λ and b_λ which ensure that the median and the interquartile range are similar for each transformation. Thus the model we attempt to estimate is

$$y_{t,\lambda} = \alpha_1(1 - X_t) + \alpha_2 X_t + f(P_t) + u_t\,,$$

with $y_{t,\lambda}$ the transformed market share of Crest, and α_1 and α_2 are the pre-endorsement and the post-endorsement intercepts. The errors u_t are modelled as a stationary autoregressive process of maximum order $s = 8$. The function $f(P_t)$ is modelled as a regression spline without intercept and with potential knot sites placed every fifteenth value of the ranked independent variable P_t.

The sampling scheme in Section 3.1 was run for a warmup period of 100 iterations with a further 100 iterations used to estimate λ and a final 400 iterations to estimate $\alpha_1, \alpha_2, f, \kappa$ and θ. Sequence plots of the parameter values and posterior probability indicate that the sampling scheme converged within a handful of iterations in the same way as occurred with the fit to the residential electricity data. Table 2 presents the estimate of the marginal

posterior distribution of λ and shows that the mode occurs for $T_\lambda(y_t) = \Phi^{-1}(y_t^{0.1})$, where $\Phi(\cdot)$ is the standard normal cdf. The rest of the estimates were calculated conditional on $\hat{\lambda}_M$ as outlined in Section 3.2.

The estimate of $E(\kappa|y)$ is $\hat{p}(\kappa = 1|y) = (1, 1, 0.886, 0.361, 0.149, 0.041, 0.014, 0.104)$, suggesting that either an AR(3) or an AR(4) model for the errors is adequate. The estimates of α_1 and α_2 are $(\hat{\alpha}_1, \hat{\alpha}_2) = (0.3063, 0.5113)$, while the estimate of θ is

$$\hat{\theta} = (0.2722, 0.2133, 0.1655, 0.0524, 0.0233, 0.0030, -0.0013, 0.0006).$$

Figure 8(a) is a plot of the estimated curve $\hat{f}(P_t)$ together with the added residual scatter plot $\hat{\zeta}_t + \hat{f}(P_t)$, where $\hat{\zeta}_t$ is defined in Section 5.2. From this plot, we conclude that $\hat{y}_{t,\lambda}$ is roughly linear in P_t and that P_t is a relatively insignificant determinant of the market share of Crest.

The Figure 8(b) is a time plot of the fitted values $[\Phi\{(\hat{y}_{t,\hat{\lambda}} - a_{\hat{\lambda}})/a_{\hat{\lambda}}\}]^{10}$ and the actual observations on the original scale. The fitted values appear to track the data well and capture the discontinuity of the intervention, suggesting that the model provides a good fit to the data. Figure 9(a) is a plot of the autocorrelations of $\hat{\zeta}_t$ and Figure 9(b) is a normal probability plot of $\hat{\zeta}_t$. These plots suggest that the errors u_t are stationary and normally distributed and follow a fourth order autoregressive process.

6 Summary

A Bayesian approach is proposed for nonparametric regression with autocorrelated errors. This approach compares favourably with other nonparametric approaches in the univariate case with time as the independent variable. More importantly, the procedure works well when the independent variable is not time and is also capable of handling a more general additive nonparametric regression model. It seems difficult to do so with other approaches to nonparametric regression. Furthermore, it should be straightforward to adapt the Bayesian approach to handle nonparametric regression with other correlated error structures.

Acknowledgement Robert Kohn's research was partially supported by an ARC grant. We would like to thank Jeff Hart for permission to use his program, Steve Marron for comments and Denzil Fiebig for suggesting the electricity consumption data set.

Appendix Implementation of the sampler

This appendix outlines how to implement the sampling scheme in Section 3.1. Let

$$\begin{aligned} S(\gamma,\psi,\lambda) &= y_\lambda'\Omega_\psi^{-1}y_\lambda - \frac{c}{1+c}y_\lambda'\Omega_\psi^{-1}X_\gamma(X_\gamma'\Omega_\psi^{-1}X_\gamma)^{-1}X_\gamma'\Omega_\psi^{-1}y_\lambda \\ \tilde{\beta}_\gamma &= \frac{c}{1+c}(X_\gamma'\Omega_\psi^{-1}X_\gamma)^{-1}X_\gamma'\Omega_\psi^{-1}y_\lambda\,, \end{aligned}$$

let q_γ be the number of elements of β_γ, and let J_λ be the absolute value of the determinant of the Jacobian matrix for the transformation $y \to y_\lambda$. From the assumptions in Section 2,

$$p(y|\beta_\gamma,\gamma,\psi,\sigma^2,\lambda)p(\beta_\gamma|\psi,\gamma,\sigma^2)p(\sigma^2) \propto (2\pi)^{-(n+q_\gamma)/2}(\sigma^2)^{-(n+q_\gamma)/2-1}c^{-q_\gamma/2}|X_\gamma'\Omega_\psi^{-1}X_\gamma|^{1/2}J_\lambda \times \exp\left\{-\frac{1}{2\sigma^2}[S(\gamma,\psi,\lambda)+\frac{c}{1+c}(\beta_\gamma-\tilde{\beta}_\gamma)'X_\gamma'\Omega_\psi^{-1}X_\gamma(\beta_\gamma-\tilde{\beta}_\gamma)]\right\} \quad (A.1)$$

Generating β_γ

$$p(\beta_\gamma,\sigma^2|y,\gamma,\psi,\lambda) \propto p(y|\beta_\gamma,\gamma,\psi,\sigma^2,\lambda)p(\beta_\gamma|\psi,\gamma,\sigma^2)p(\sigma^2)\,. \quad (A.2)$$

Integrating σ^2 out of (A.1) and (A.2) using an inverse gamma integral,

$$p(\beta_\gamma|y,\gamma,\psi,\lambda) \propto \left\{S(\gamma,\psi,\lambda)+\frac{c}{1+c}(\beta_\gamma-\tilde{\beta}_\gamma)'X_\gamma'\Omega_\psi^{-1}X_\gamma(\beta_\gamma-\tilde{\beta}_\gamma)\right\}^{-(n+q_\gamma)/2}.$$

That is, the conditional distribution of β_γ is multivariate t with degrees of freedom n, which is centred at $\tilde{\beta}_\gamma$ and has scale matrix $(X_\gamma'\Omega_\psi^{-1}X_\gamma)^{-1}cS(\gamma,\psi,\lambda)/n(1+c)$; for n large, this conditional distribution is effectively multivariate normal.

Generating ψ_i and κ_i These variables are generated as a block with σ^2 integrated out. Note that

$$p(\psi_i,\kappa_i|y,\beta_\gamma,\gamma,\psi_{j\neq i},\lambda) \propto p(\psi_i,\kappa_i|u,\psi_{j\neq i})\,.$$

We sketch out the necessary steps to generate ψ_i, κ_i; the details are in Barnett *et. al.* (1996a).

$$\begin{aligned} p(\psi_i, \kappa_i = 1, \sigma^2 | u, \psi_{j \neq i}) &\propto p(u|\psi, \kappa_i = 1, \sigma^2) p(\psi_i | \kappa_i = 1) p(\sigma^2) \\ &\propto \frac{1}{2} p(\kappa_i = 1)(\sigma^2)^{-n/2+1} |\Omega_\psi|^{-1/2} \exp(-u' \Omega_\psi^{-1} u/2) \,. \end{aligned}$$

Integrating σ^2 out using an inverse gamma integral,

$$p(\psi_i, \kappa_i = 1 | u, \psi_{j \neq i}) \propto \frac{1}{2} p(\kappa_i = 1) |\Omega_\psi|^{-1/2} (u' \Omega_\psi^{-1} u)^{-n/2} \,.$$

The binary variable κ_i is generated first without conditioning on ψ_i. To integrate out ψ_i we proceed as follows. Let ξ_t and R_t be defined as in Section 5.2. Then,

$$u' \Omega_\psi^{-1} u = \sum_{t=1}^{n} \xi_t^2 / R_t \,,$$

the ξ_t are independent and, for $t > s$, $\xi_t = u_t - \theta_1 u_{t-1} \ldots - \theta_s u_{t-s}$ and $R_t = 1$. Barnett *et. al.* (1996a) show that, for $t > s$, ξ_t is linear in ψ_i so it is possible to write $\sum_{t=s+1}^{n} \xi_t^2 = A(\psi_i - B)^2 + C$, with A, B and C independent of ψ_i, A and C positive, and all three terms computed as in Barnett *et. al.* (1995). Let $D(\psi_i) = \sum_{t=1}^{s} \xi_t^2 / C$. Then $D << 1$ as its numerator is the sum of s terms whereas its denominator is the sum of $n - s$ terms. Thus,

$$(u' \Omega_\psi^{-1} u)^{-n/2} \approx C^{-n/2} \{1 + A(\psi_i - B)^2 / C\}^{-n/2} \approx C^{-n/2} \exp\{-nA(\psi_i - B)^2 / 2C\} \,.$$

Let $g(\psi_i) = |\Omega_\psi|^{-1/2} \{D + 1 + A(\psi_i - B)^2 / C\}^{-n/2} (C/nA)^{1/2} \phi(\psi_i; B, C/nA)^{-1}$, where $\phi(x, \mu, \sigma^2)$ is the normal density evaluated at x with mean μ and variance σ^2. Then,

$$p(\kappa_i = 1 | u, \psi_{j \neq i}) \propto \frac{1}{2} p(\kappa_i = 1) \int_{-1}^{1} g(\psi_i) \phi(\psi_i, B, C/nA) d\psi_i \,. \tag{A.3}$$

The integral in (A.3) is readily evaluated by approximating $\log g(\psi_i)$ by a quadratic in each of the intervals $B + (j-1)(C/nA)^{1/2}/2, B + j(C/nA)^{1/2}/2$, $j = -7, \ldots, 8$. Similarly,

$$p(\kappa_i = 0 | u, \psi_{j \neq i}) \propto p(\kappa_i = 0) |\Omega_\psi|^{-1/2} \{D + 1 + A(\psi_i - B)^2 / C\}^{-n/2} \,, \tag{A.4}$$

with (A.4) evaluated at $\psi_i = 0$. The conditional probability of κ_i is obtained by normalising (A.3) and (A.4). Once κ_i is generated, $\psi_i = 0$ if $\kappa_i = 0$. If $\kappa_i = 1$, we generate ψ_i from $q(\psi_i) = \phi(\psi_i, B, c/nA)$ and use a Metropolis-Hastings step.

Generating γ_i

$$p(\gamma_i, \beta_\gamma, \sigma^2, \lambda | y, \gamma_{j\neq i}, \psi) \propto p(y|\beta_\gamma, \gamma, \psi, \sigma^2, \lambda) p(\beta_\gamma|\psi, \gamma, \sigma^2) p(\sigma^2) p(\lambda) p(\psi_i) \,. \tag{A.5}$$

Integrating β_γ out of (A.1) and (A.5) using a normal integral, then integrating σ^2 out using an inverse gamma integral, and finally summing over $\lambda \in \Lambda$, we obtain

$$p(\gamma_i|y, \gamma_{j\neq i}, \psi) \propto \sum_{\lambda\in\Lambda} J_\lambda p(\lambda)(1+c)^{-q_\gamma/2} S(\gamma, \psi, \lambda)^{-n/2} \tag{A.6}$$

The conditional density of γ_i is obtained by evaluating (A.6) for $\gamma_i = 0$ and 1, and normalising.

Generating λ

$$p(\lambda|y, \gamma, \psi) \propto p(y|\gamma, \psi, \lambda) p(\lambda) \propto S(\gamma, \psi, \lambda)^{-n/2} J_\lambda p(\lambda) \tag{A.7}$$

The posterior probability of λ is obtained by evaluating (A.7) for all $\lambda \in \Lambda$ and normalising.

We conclude by showing how to efficiently compute $y'_\lambda \Omega_\psi^{-1} y_\lambda, y'_\lambda \Omega_\psi^{-1} X_\gamma, X'_\gamma \Omega_\psi^{-1} X_\gamma$, and $|\Omega_\psi|$. As in Monahan (1984), for $t = 1, \ldots, s$,

$$\xi_t = u_t - \theta_{t,1} u_{t-1} - \cdots - \theta_{t,t-1} u_1 \quad \text{and} \quad R_t = \{(1-\psi_t^2)\cdots(1-\psi_s^2)\}^{-1} \,. \tag{A.8}$$

Let M be a lower triangular matrix with ones on the diagonal; for $t = 1, \ldots, s$, $M_{t,t-j} = -\theta_{t,j}, j = 1, \ldots, t-1$; for $t > s$, $M_{t,t-j} = -\theta_j, j = 1, \ldots, s$, and $M_{t,j} = 0$ for $j > s$. Thus M is a lower triangular band matrix with bandwidth at most s. From (A.8) and the definition of ξ_t and R_t, $Mu = \xi, M\Omega_\psi M' = R$, where R is a diagonal matrix with tth diagonal element R_t. Let $\tilde{y} = My_\lambda, \tilde{X} = MX_\gamma$; then, $y'_\lambda \Omega_\psi^{-1} y_\lambda = \tilde{y}' R^{-1} \tilde{y}, y'_\lambda \Omega_\psi^{-1} X_\gamma = \tilde{y}' R^{-1} \tilde{X}, X'_\gamma \Omega_\psi^{-1} X_\gamma = \tilde{X}' R^{-1} \tilde{X}$, and $|\Omega_\psi| = \prod_{t=1}^s R_t$.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Am. Statist. Ass.*, **85**, 749–759.

Barnett, G., Kohn, R. and Sheather, S. (1996a). Bayesian estimation of an autoregressive model using Markov Chain Monte Carlo. *J. Econometrics.*, **74**, 237–254.

Barnett, G., Kohn, R. and Sheather, S. (1996b). Robust Bayesian estimation of autoregressive-moving-average models. *J. Time Series Analysis*, in press.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J. R. Statist. Soc.*, B **26**, 211–52.

Box, G.E.P. and Cox, D.R. (1982). An analysis of transformations revisited, rebutted. *J. Am. Statist. Ass.*, **77**, 209–210.

Breiman, L., and Friedman J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Ass.*, **80**, 580–598.

Chib, S. and E. Greenberg. (1994). Bayes inference in regression models with ARMA(p, q) errors. *J. Econometrics*, **64**, 183-206.

Chu, C.K. and Marron, J.S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statistics*, **19**, 1906-1918.

Diggle P., and Hutchinson, M. (1989). On Spline Smoothing with Autocorrelated errors. *Australian J. Statist.*, **31**, 166–182.

Emerson, J.D. and Stoto, M.A. (1983). Transforming data. In *Understanding robust and exploratory data analysis* (eds. Hoaglin, D.C., Mosteller, F. and Tukey, J.W.), pp. 97-128. New York: John Wiley and Sons.

Engle, R., Granger, W., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales *J. Am. Statist. Ass.*, **81**, 310–320.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398-409.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

Harris, J., and Liu, L. (1993). Dynamic Structural analysis and forecasting of residential

electricity consumption. *International J. Forecasting*, **2**, 437–455.

Hart, J.D. (1991). Kernel regression estimation with time series errors. *J. R. Statist. Soc.* B, **53**, 173–187.

Hart, J.D. (1994). Automated Kernel smoothing of dependent data by using time series cross-validation. *J. R. Statist. Soc.* B **56**, 529–42.

Harvey, A. and Koopman, S. (1993). Forecasting hourly electricity demand using time-varying splines. *J. Am. Statist. Ass.*, **88**, 1228–1236.

Hastie, T.J. and R.J. Tibshirani, R.J. (1990). *Generalized additive models.* New York: Chapman Hall.

Kohn, R., Ansley, C. and Wong, C. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, **79**, 335–346.

McCulloch R. and Tsay, R. (1994). Bayesian analysis of autoregressive time series via the Gibbs sampler. *J. Time Series Analysis*, **51**, 235–250.

Monahan, J. (1984). Full Bayesian analysis of ARMA time series models. *J. Econometrics*, **21**, 307–31.

Smith, M. and Kohn, R. (1996a). Nonparametric regression using Bayesian variable Selection. *J. Econometrics*, **75**, 317-344.

Smith, M., and Kohn, R. (1996b). Nonparametric Bivariate Regression. preprint

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals Statist.*, **22**, 1701–1762.

Wichern, D.W. and Jones, R.H. (1977). Assessing the impact of market disturbances using intervention analysis. *Management Science*, **24**, 329-337.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with

g-prior distributions. In *Bayesian Inference and Decision Techniques-Essays in Honor of Bruno de Finetti*, (eds. Goel, P.K. and Zellner, A.), pp. 233-243. Amsterdam: North-Holland.

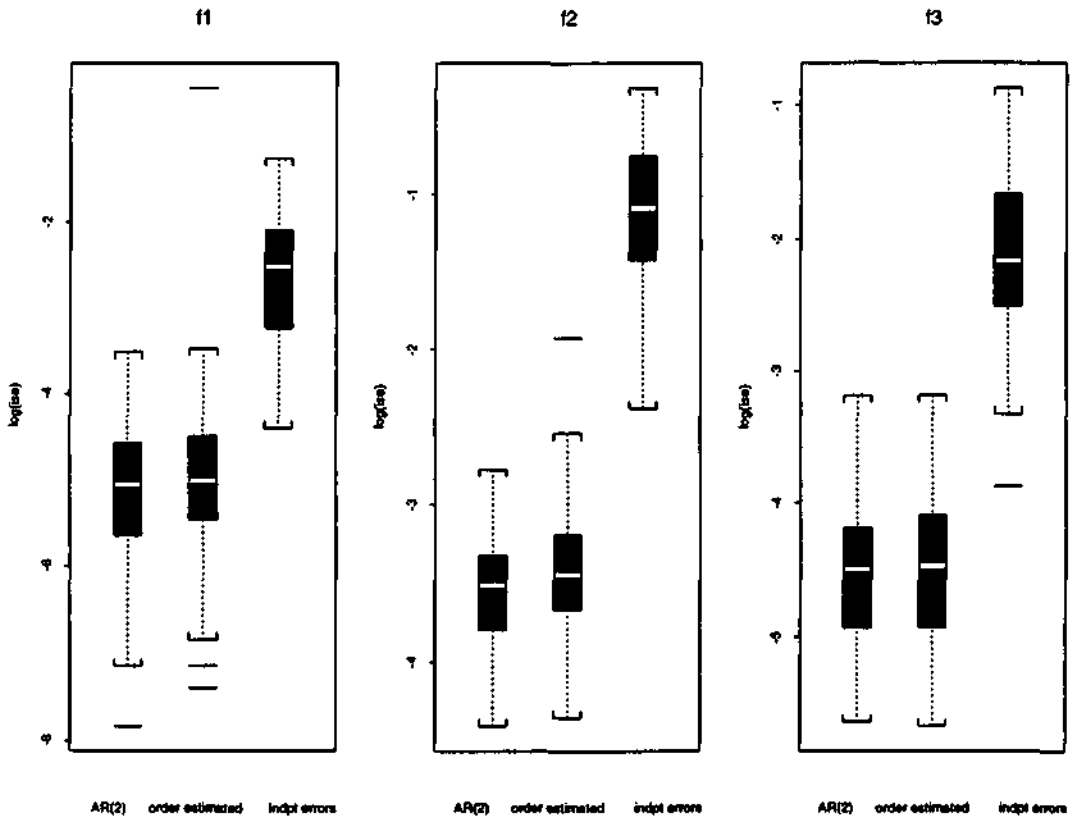

Figure 1: Boxplots of $\log(ISE)$ for each of the three test functions and the three estimators. The left, middle and right panels correspond to f_1, f_2 and f_3. In each panel the boxplot on the left corresponds to the estimator which assumes that the order of the autoregression is known to be 2, the middle boxplot corresponds to the estimator which does not assume that the order of the autoregression is known, and the right boxplot corresponds to the estimator which takes the errors to be independent.

i	$\hat{\kappa}_i$	$\hat{\theta}_i$	i	$\hat{\kappa}_i$	$\hat{\theta}_i$	i	$\hat{\kappa}_i$	$\hat{\theta}_i$	i	$\hat{\kappa}_i$	$\hat{\theta}_i$
1	1	0.398	6	0.089	0.011	11	1	0.158	16	0.055	-0.002
2	0.992	-0.094	7	0.068	-0.131	12	1	0.544	17	0.047	-0.001
3	0.232	0.003	8	0.741	0.206	13	0.171	-0.024	18	0.055	0.005
4	0.202	-0.058	9	0.093	-0.126	14	0.071	-0.002	19	0.053	0
5	0.117	0.062	10	1	0.013	15	0.059	0.002	20	0.082	-0.009

Table 1: Estimates of $p(\kappa_i = 1|y)$ $(\hat{\kappa}_i)$ and ϕ for the residential electricity data.

λ	$T_\lambda(y)$	a_λ	b_λ	$\hat{p}(\lambda\|y)$
1	$\Phi^{-1}(y^{0.1})$	-0.4061	0.5809	0.5016
2	$\Phi^{-1}(y^{0.25})$	-0.0039	0.4656	0.2363
3	$\Phi^{-1}(y^{0.5})$	0.2498	0.3814	0.1097
4	$\Phi^{-1}(y^{0.75})$	0.3756	0.3348	0.0655
5	$\Phi^{-1}(y)$	0.4544	0.3033	0.0444
6	$\Phi^{-1}(y^{1.5})$	0.5510	0.2618	0.0253
7	$\Phi^{-1}(y^2)$	0.6095	0.2345	0.0170
8	$\log(\frac{y}{1-y})$	0.4434	0.1763	0.0001
9	$\log(-\log(1-y))$	0.5007	0.2014	0

Table 2: Candidate transformations $T_\lambda(y)$, normalising constants a_λ and b_λ, and the posterior probability estimate of each transformation $\hat{p}(\lambda|y)$ for the toothpaste data. Φ is the cumulative distribution function of the standard normal.

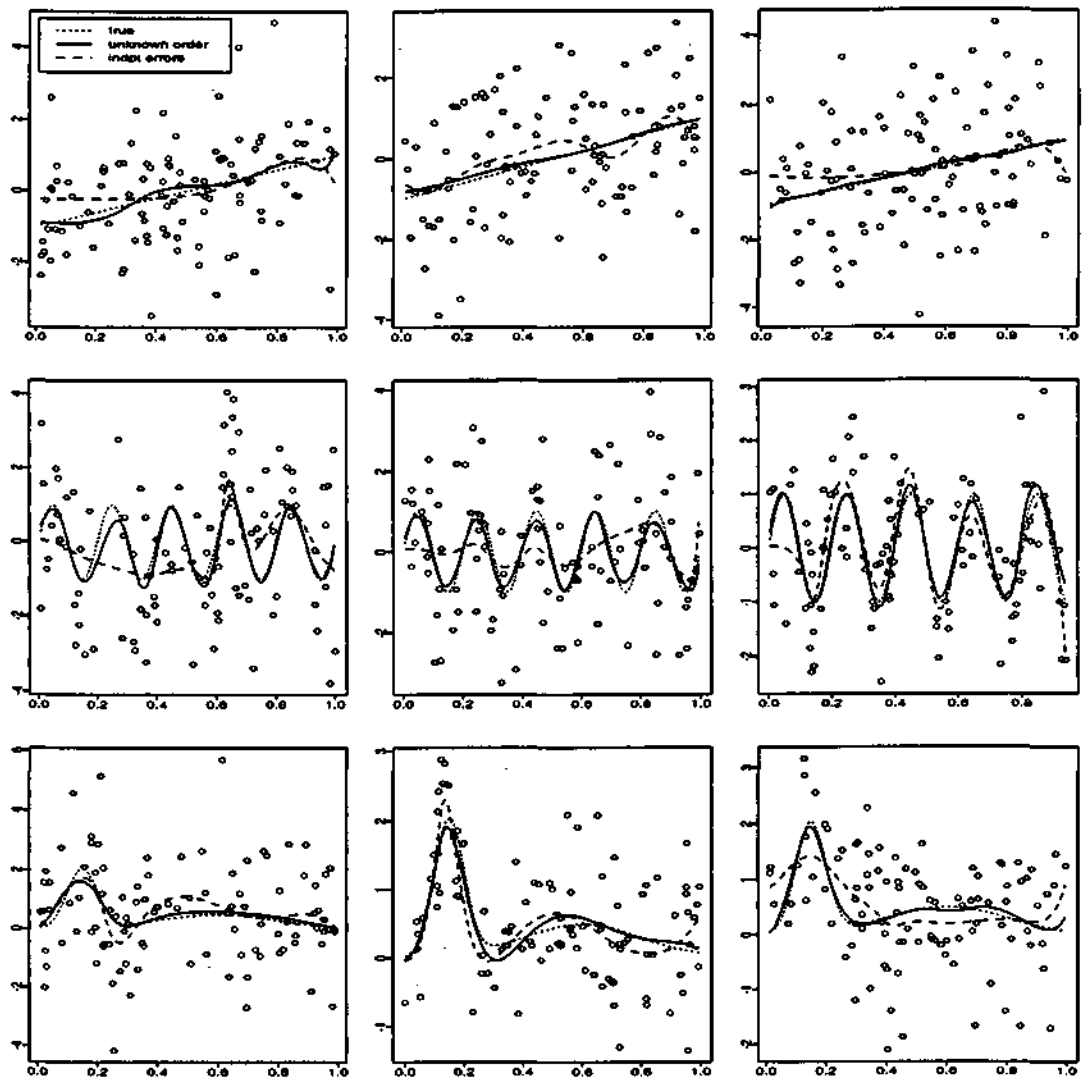

Figure 2: Plots of f_1, f_2 and f_3 (dotted line) together with estimate based on unknown order of the autoregression (solid line) and independent errors (dashed line). For each function, the left, middle and the right panels correspond to the 10th, 51st and 90th worst fits as judged by ISE.

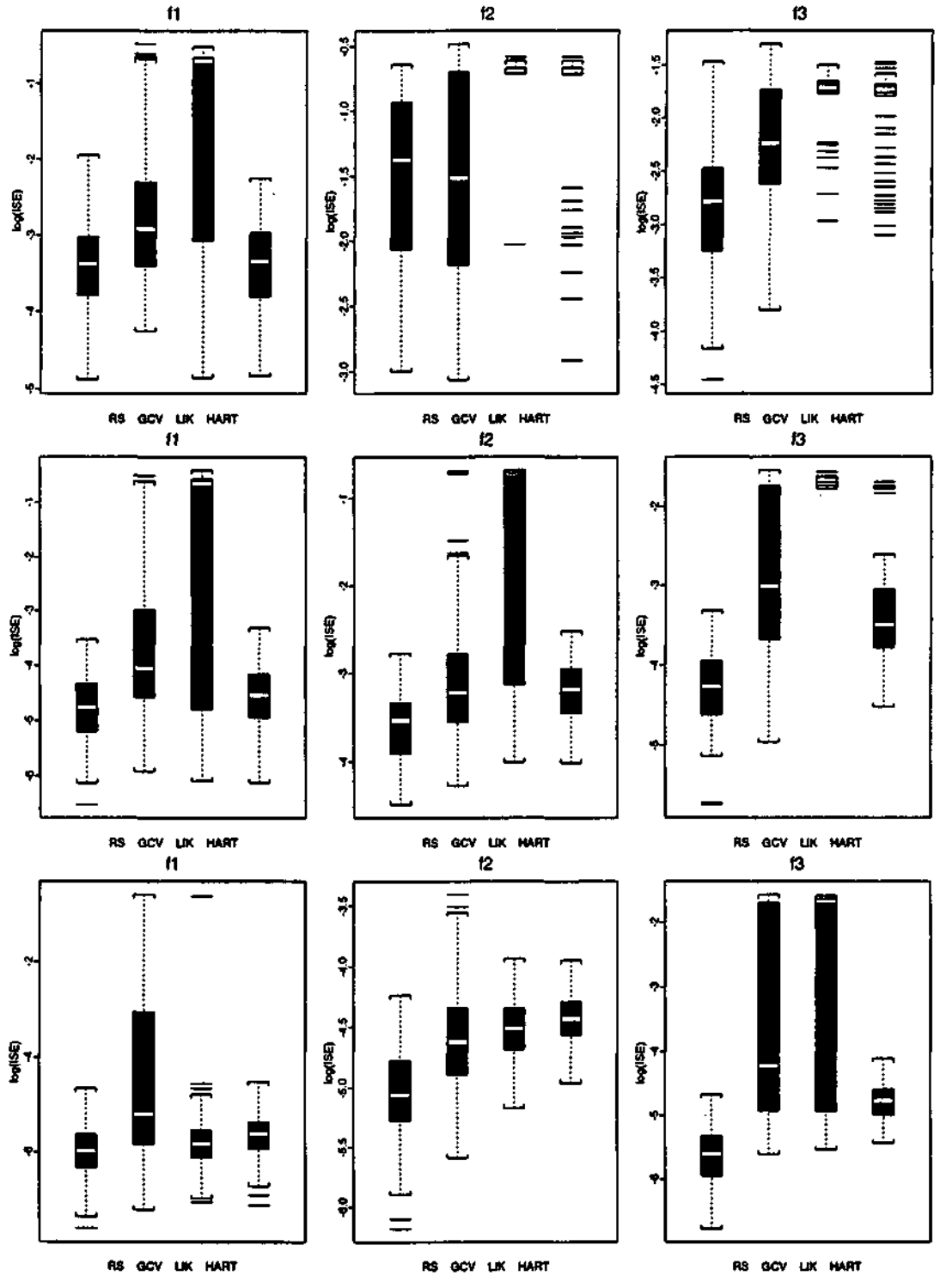

Figure 3: Boxplots of $\log(ISE)$ for each of the three test functions, three different noise levels and four different estimators. The bottom three panels correspond to low noise, the middle three to medium noise, and the top three to high noise. In each panel, the estimators from left to right are: regression spline, smoothing spline using generalised cross-validation, smoothing spline using marginal likelihood and Hart's estimate.

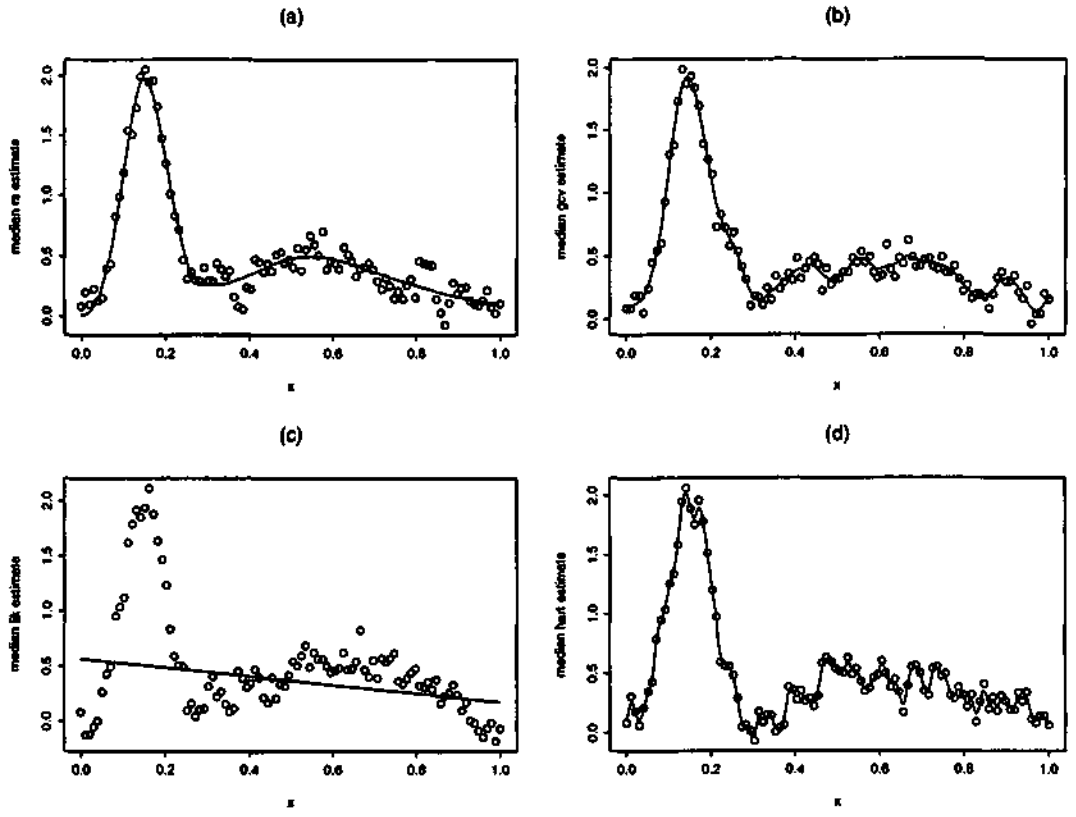

Figure 4: Plot of the estimate of f_3 corresponding to the median fit as judged by ISE together with the corresponding data set. The independent variable is in time order, the errors are a first order autoregression, and the noise level is low. Panels (a)–(d) correspond respectively to the regression spline estimate, the smoothing spline estimate using generalised cross-validation, smoothing spline estimate using marginal likelihood and Hart's estimate.

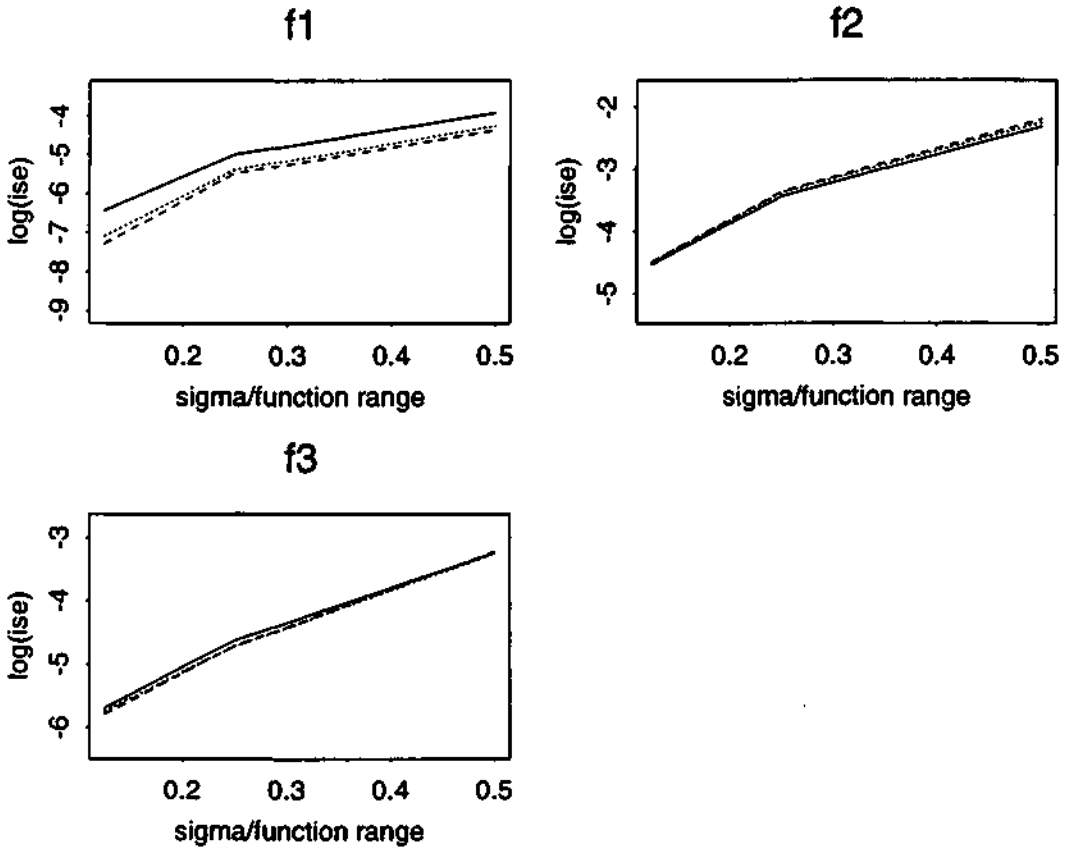

Figure 5: $\log_e(ISE)$ (vertical axes) for each of the three functions averaged over 50 replications. In each panel the bold line represents the results when $c = 100$, the dotted line $c = 500$ and the dashed line $c = 1000$. The horizontal axis is σ divided by the range of the function

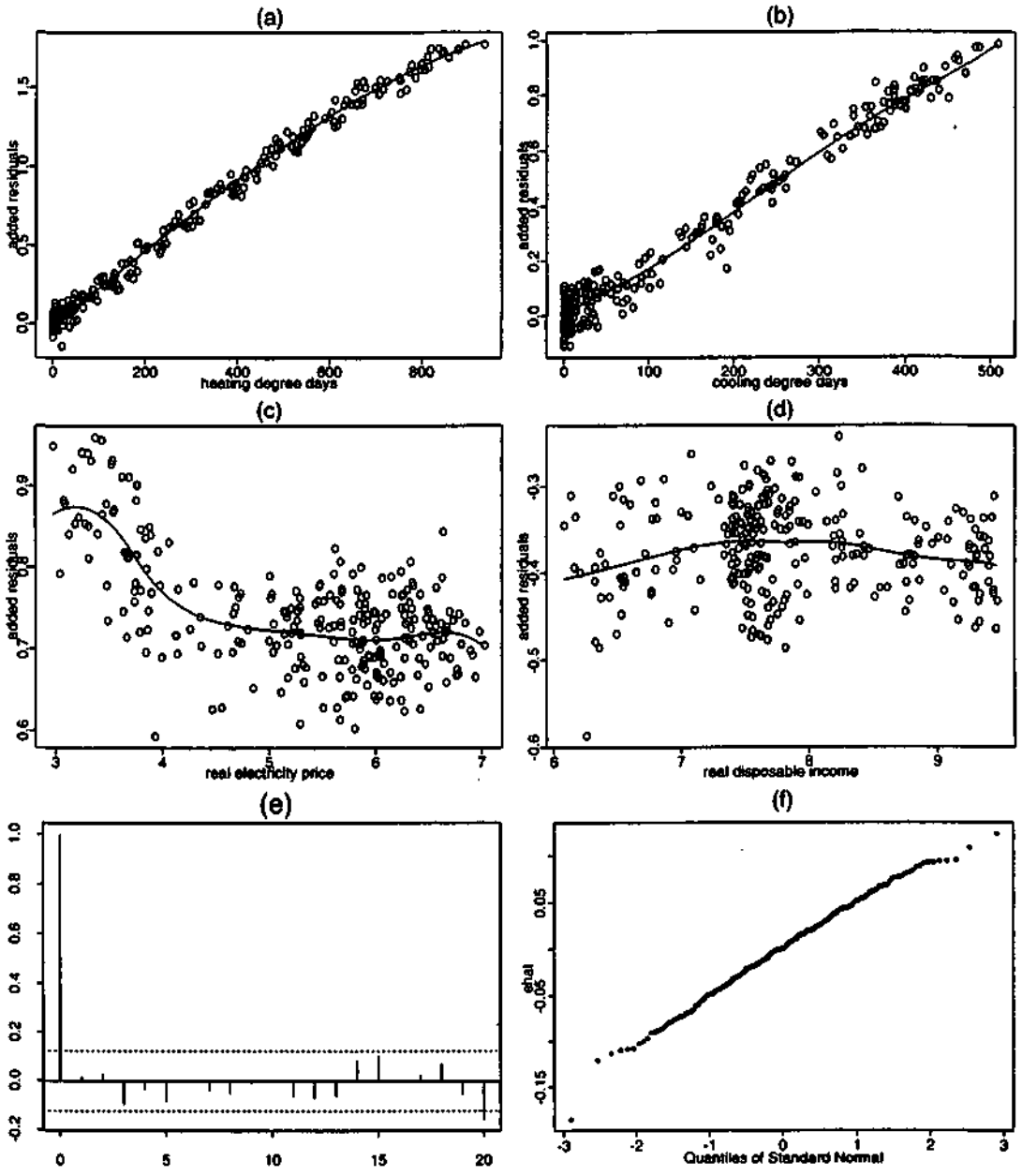

Figure 6: Parts (a)–(d). Plots of the function estimates $\hat{f}_1(x_{1t}), \ldots, \hat{f}_4(x_{4t})$ (solid line) together with the added residual scatter plots $\hat{f}(x_{1t}) + \hat{\zeta}_t, \ldots, \hat{f}(x_{4t}) + \hat{\zeta}_t$. (e) Autocorrelations of $\hat{\zeta}_t$. (f) Normal probability plot of $\hat{\zeta}_t$.

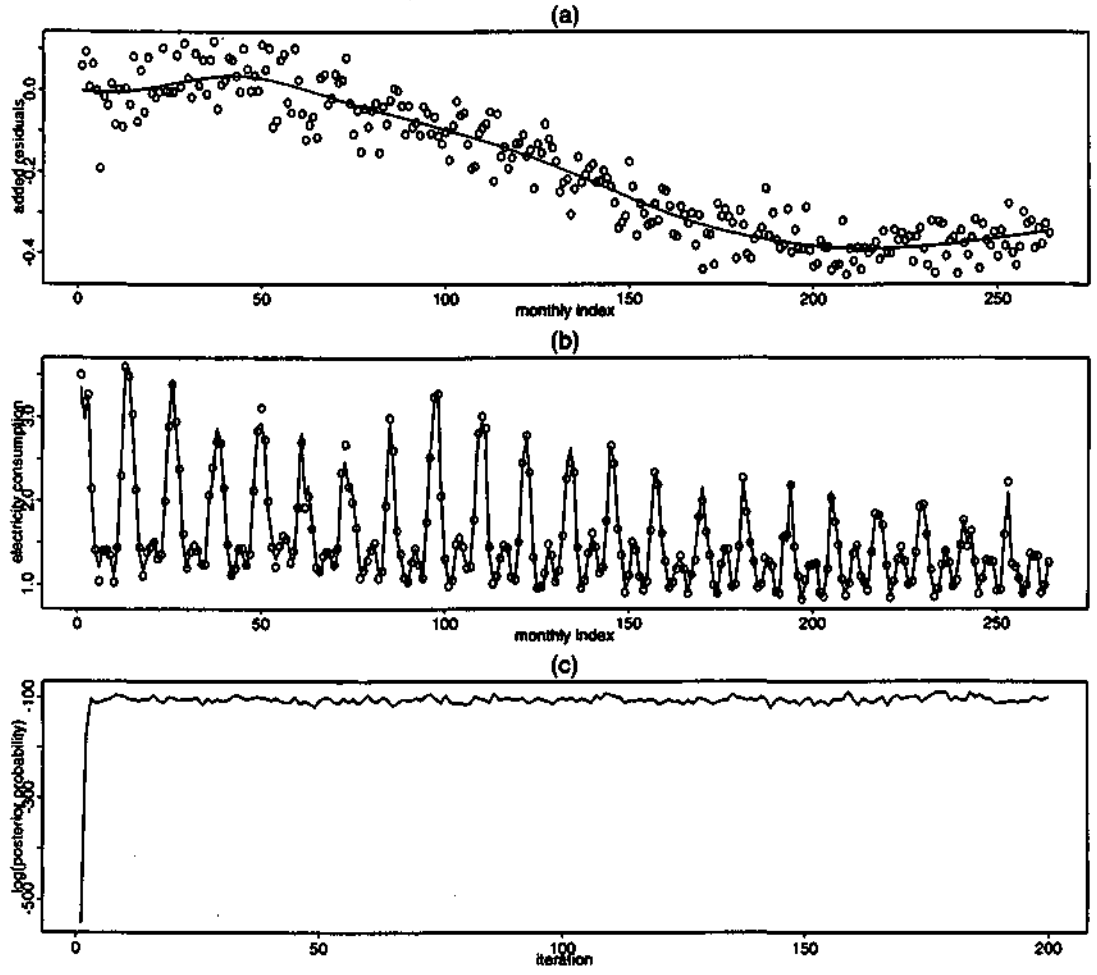

Figure 7: (a) Plot of the trend estimate $\hat{D}(t)$ (solid line) together with the added residual plot $\hat{D}(t) + \hat{\zeta}_t$. (b) Residential electricity consumption over time (scatter plot) together with the fitted values (solid line). (c) Trace of the log posterior density $\log p(\gamma^{[k]}, \kappa^{[k]}, \beta^{[k]}, \lambda^{[k]}, \psi^{[k]}|y)$ (up to an additive constant) for the first 200 iterations of the sampler.

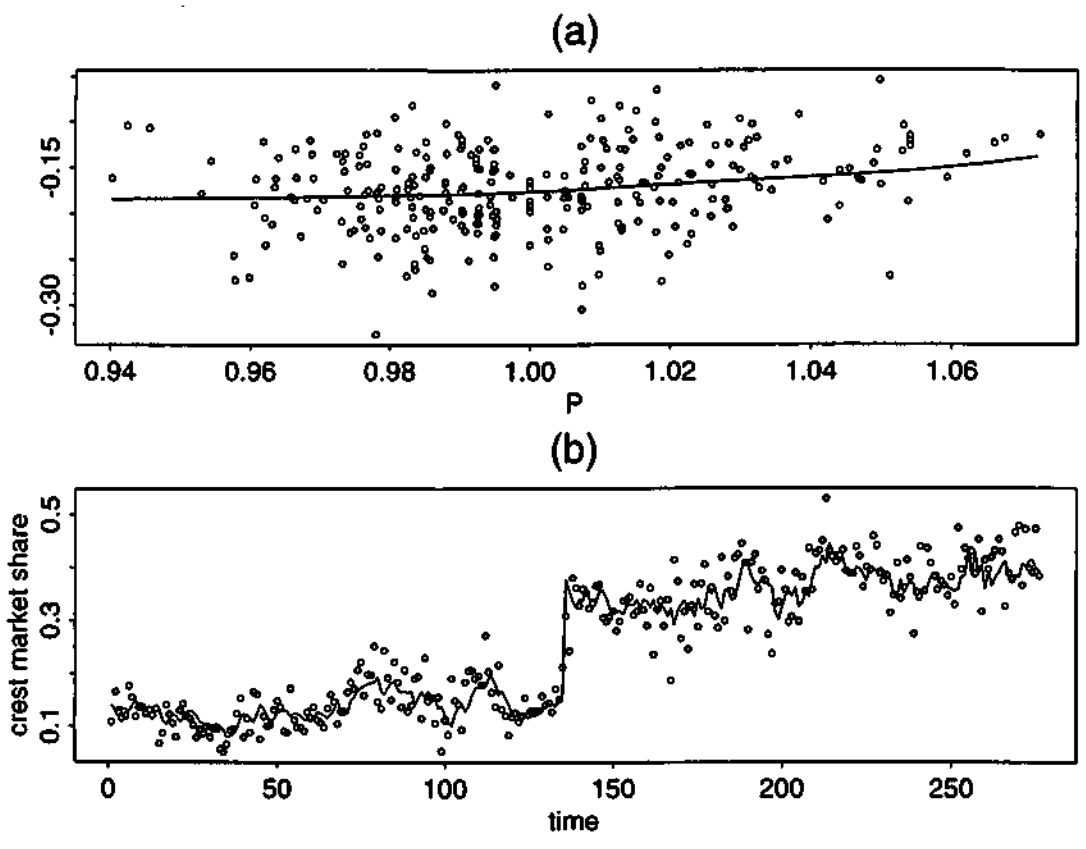

Figure 8: (a) Plot of $\hat{f}(P_t)$ (solid line) and the added residual scatter plot $\hat{f}(P_t)$ against P_t. (b) Crest market share (scatter plot) and the fitted values on the original scale (plotted as a solid line) from the model estimate.

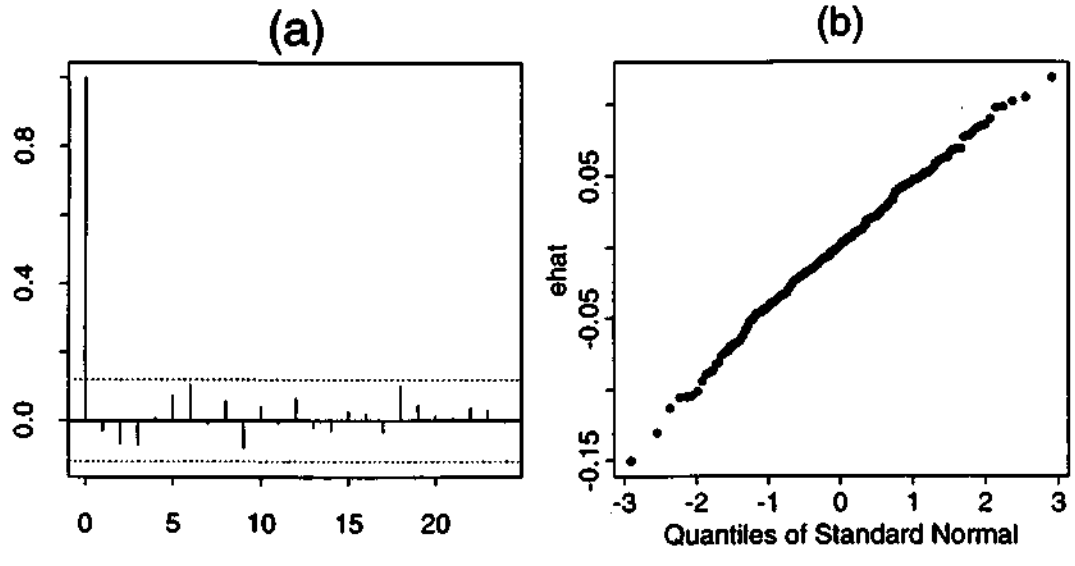

Figure 9: Crest data. (a) plot of the sample autocorrelation function of $\hat{\zeta}_t$. (b) Normal probability plot of $\hat{\zeta}_t$.