Additive nonparametric regression with autocorrelated errors

Michael Smith,
Monash University, Clayton, Australia

Chi-Ming Wong
Hong Kong University of Science and Technology, Hong Kong

and Robert Kohn†
University of New South Wales, Kensington, Australia

[Received November 1995. Revised May 1997]

Summary. A Bayesian approach is presented for nonparametric estimation of an additive regression model with autocorrelated errors. Each of the potentially non-linear components is modelled as a regression spline using many knots, while the errors are modelled by a high order stationary autoregressive process parameterized 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 by using the Markov chain Monte Carlo method. It is shown empirically that the approach proposed works well on several simulated and real examples.

Keywords: Autoregressive model; Bayesian analysis; Data transformation; Markov chain Monte Carlo method; Regression spline; Subset selection

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 et al. (1986) and Harvey and Koopman (1993). Few approaches are currently available for estimating a regression model nonparametrically when the errors are autocorrelated, despite the fact that a failure to take account of the autocorrelation can result in poor function estimates; see Section 4.1 and Altman (1990) for simulation evidence. Those researchers who 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 generalize 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 non-linear component is modelled as a regression spline with many knots and the errors are modelled as a stationary autoregression parameterized by its partial autocorrelations. The distribution of significant knots in the regression spline and significant partial autocorrelations is accounted

†Address for correspondence: Australian Graduate School of Management, University of New South Wales, PO Box 1, Kensington, New South Wales 2033, Australia.
E-mail: robertk@agsm.unsw.edu.au

© 1998 Royal Statistical Society
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 by using the Markov chain Monte Carlo method.

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 et al. (1996). It can also be extended to bivariate surface estimation, as is demonstrated in Smith and Kohn (1997).

This paper links two lines of research. The first is by Smith and Kohn (1996) who combined regression splines with Bayesian subset selection to estimate nonparametrically an additive regression model with independent errors. They showed 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. (1996) who proposed 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 (1996) builds on the Bayesian subset selection paper of George and McCulloch (1993).

The paper is organized 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 with previous estimators. Section 5 considers multiple-regression examples and Appendix A shows how to implement the sampler.

2. Model and prior assumptions

2.1. Autoregressive model for the errors

Suppose that

\[ y_t = f(x_t) + u_t, \quad t = 1, \ldots, n, \quad (2.1) \]

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 \( s \),

\[ u_t = \theta_1 u_{t-1} + \ldots + \theta_s u_{t-s} + e_t, \]

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 et al. (1997).

Let \( \psi_i \) be the \( i \)-th 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 Barnett et al. (1996) it is convenient to enforce stationarity by reparameterizing \( u_t \) 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. (1996), the following prior assumptions are made on \( \sigma^2 \) and \( \psi \).
Additive Nonparametric Regression 313

(a) log(σ²) has a flat prior on the line, so that \( p(\sigma^2) \propto 1/\sigma^2 \). This is a commonly used prior for \( \sigma^2 \).

(b) Let \( \kappa_i \) be a binary variable determining the status of \( \psi_i \). If \( \kappa_i = 0 \) then \( \psi_i \) is identically zero; if \( \kappa_i = 1 \) then \( \psi_i \) is uniformly distributed on \((-1, 1)\). We assume that, \textit{a priori}, the \( \psi_i \) are independent of each other. This ensures that the \( u_i \) are both stationary and parsimoniously parameterized and that the prior distribution of \( \psi|\kappa_i = 1 \) is non-informative. We also assume that \( \psi_i \) is \textit{a priori} independent of \( \sigma^2 \).

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. Modelling the unknown function

Following Smith and Kohn (1996) we model the unknown regression function as a linear combination of cubic regression spline basis terms

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

where \( \tilde{x}_1, \ldots, \tilde{x}_m \) are \( m \) knots placed along the domain of the independent variable \( x \), such that \( \min_j(x_j) < \tilde{x}_1 < \tilde{x}_2 < \ldots < \tilde{x}_m < \max_i(x_i) \) and \( z_+ = \max(0, z) \). However, the framework can easily accommodate other univariate bases. When \( f(x) \) is observed with noise, the two most important problems in approximating \( f(x) \) by expression (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 (1996) solved 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 equation (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, \( x = (x_1, \ldots, x_n)' \) and \( 1 \) be an \( n \times 1 \) vector of 1s. Let \( x_j' \) and \( (x - 1 \tilde{x}_j)_+^3 \) be the \( n \times 1 \) vectors with \( j \)th elements \( x_j' \) and \( (x_j - \tilde{x}_j)_+^3 \) respectively, and define the \( n \times (r + 1) \) design matrix \( X = (1, x, x^2, x^3, (x - 1 \tilde{x}_1)_+^3, \ldots, (x - 1 \tilde{x}_m)_+^3) \). By replacing \( f(x) \) by expression (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 \( \text{var}(u) = \sigma^2 \Omega_u \) 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 \( \beta_\gamma \) and \( X_\gamma \) be the subvector of \( \beta \) and the submatrix of \( X \) respectively, corresponding to the non-zero elements of \( \gamma \). Given \( \sigma^2, \psi \) and \( \kappa \), we place the following prior on \( \gamma \) and \( \beta \).
The $\gamma_i$ are independent \textit{a priori} with $p(\gamma_i = 1) = \frac{1}{2}$, whereas the conditional prior for $\beta_i$ is $\beta_i|\sigma^2, \gamma, \psi \sim N(0, c\sigma^2(X'_i\Omega_\psi^{-1}X_i)^{-1})$.

The prior for $\gamma$ means that we have no prior information on which columns of $X$ to include. By taking $c$ large, the prior for $\beta_i$ is almost diffuse relative to the information in the likelihood. In particular, in this paper we take $c = n$ which results in a joint posterior probability for the parameters that corresponds closely to the Schwarz (1978) criterion. A detailed discussion of the possible priors for $\beta_i$ can be found in Smith (1996). In addition, we have checked using comprehensive simulations that estimates based on this prior are relatively insensitive to a wide range of values for $c$ (see Smith \textit{et al.} (1996)).

2.3. \textbf{Data transformation}

For a linear regression model with independent errors, Box and Cox (1964) showed 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 the scale of $T_\lambda(y)$ different for each $\lambda$. Therefore, in a linear regression model Box and Cox (1964) advocated making the prior for $\beta$ and $\sigma^2$ depend on $\lambda$. We take a different approach. Instead of working directly with $T_\lambda(y)$ we follow Smith and Kohn (1996) and work with the normalized transformation $y_\lambda = a_\lambda + b_\lambda T_\lambda(y)$. For each $\lambda$, the scalars $a_\lambda$ and $b_\lambda$ are chosen as follows. Let $y_{(i)}$ and $y_{(0),\lambda}$ be the $i$th ordered values of $y_i$ and $y_{i,\lambda} = a_\lambda + b_\lambda T_\lambda(y_i)$. The scalars $a_\lambda$ and $b_\lambda$ 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 $a_\lambda$ and $b_\lambda$ is motivated by, but is a little different from, the transformation approach of Emerson and Stoto (1983). Because the centre and the scale of the $y_{i,\lambda}$ are the same for all $\lambda$ we make the following assumption.

The parameters $\sigma^2, \kappa, \psi, \gamma$ and $\beta$ are \textit{a priori} independent of $\lambda$.

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

3. \textbf{Sampling scheme and parameter estimation}

3.1. \textbf{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. See Tierney (1994) for a discussion of Markov chain Monte Carlo sampling methods. To describe the sampling scheme it is convenient to use the notation $\alpha|\beta$ to mean that $\alpha$ is generated conditionally on $\beta$. 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. Details of implementation are given in Appendix A. Let $\psi = (\psi_1, \ldots, \psi_s)$, $\kappa = (\kappa_1, \ldots, \kappa_s)$ and $\gamma = (\gamma_0, \ldots, \gamma_s)$.

### 3.1. 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 by using the following conditioning:

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

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 $\sigma^2$ is integrated out of the sampling scheme. In step (a), $\beta_i$ is generated from its conditional distribution $p(\beta_i|\kappa, \psi, \gamma, \lambda, y)$ which is multivariate $t$. In step (b), $\psi_i$ and $\kappa_i$ are generated as a block. The binary indicator $\kappa_i$ is generated first from $p(\kappa_i|\kappa_{j\neq i}, \psi_{j\neq i}, \gamma, \beta, \lambda, y)$ by using numerical integration over $\psi_i$ to determine the conditional probability that $\kappa_i = 1$. The partial autocorrelation $\psi_i$ 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 $\psi_i$. The generation of $\psi_i$ is completed using a Metropolis–Hastings step. It is necessary to generate $\psi_i$ and $\kappa_i$ simultaneously as generating them one at a time produces a reducible sampling scheme which does not converge. In step (c), 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 $\gamma_i = 1$ up to a scale factor and then normalizing. The transformation parameter $\lambda$ is generated from its multinomial conditional probability as for each $\lambda \in \Lambda$ the conditional probability of $\lambda$ can be evaluated up to a scale factor and then normalized.

The family of transformations $\Lambda$ is taken discrete and small to make it easy to integrate $\lambda$ out when generating $\gamma_i$ and also to generate $\lambda$. It is necessary to integrate $\lambda$ out when generating $\gamma$ as the sampler which conditions $\gamma$ on $\lambda$ converges too slowly to be practical.

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

### 3.2. Estimation

Estimation is done in two stages. The parameter $\lambda$ 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 conditionally on $\lambda_M$. 

**Additive Nonparametric Regression**

315
This two-stage approach was used by Box and Cox (1964) who advocated estimating the regression on a given scale (i.e. by conditioning on the estimate of \( \lambda \)) rather than averaging over the distribution of \( \lambda \).

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

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

Let \( \hat{\lambda}_M \) be 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 (d) of the sampling scheme. New iterates \( \{ \beta^{[k]}, \gamma^{[k]}, \kappa^{[k]}, \psi^{[k]} \} \) are obtained. The following estimates of \( E(\beta | y), E(\gamma | y) \) and \( E(\kappa_i | y) = p(\kappa_i = 1 | y) \) are used to estimate the regression function and the autoregressive parameters:

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

The estimates \( \hat{\beta} \) and \( \hat{\kappa} \) are called mixture estimates as they are based on the conditional distributions of \( \beta \) and \( \kappa \). The estimate \( \hat{\gamma} \) 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 expression (2.2). The estimate \( \hat{\theta} \) of the autoregressive parameter \( \theta \) 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) dealt only with the univariate case where time is the independent variable. They showed 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 model (2.1) with \( u_t = \phi u_{t-1} + e_t \), with \( e_t \) independent \( N(0, \sigma^2) \) and \( \phi \) 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 \( \sigma^2 \). 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 \( \phi \) close to 1.
This gain in efficiency is illustrated by simulation. 100 observations were generated from model (2.1) for each of the three functions

\[
\begin{align*}
  f_1(x) &= 2x - 1, \\
  f_2(x) &= \sin(10\pi x), \\
  f_3(x) &= \{\phi(x; 0.15, 0.05) + \phi(x; 0.6, 0.2)\}/4;
\end{align*}
\]

(4.1)

\(\phi(x; \mu, \sigma)\) is the Gaussian density, with mean \(\mu\) and standard deviation \(\sigma\), evaluated at \(x\). The independent variable \(x_i\) was generated uniformly on \((0, 1)\), and the errors \(u_i\) were generated from the second-order autoregression \(u_i = 0.9u_{i-1} - 0.9u_{i-2} + e_i\), with \(e_i \sim N(0, 0.5^2)\). This was replicated 100 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 \(\kappa_1\) and \(\kappa_2\) fixed at 1 and \(s = 2\), i.e. the error structure is assumed to be known except for the values of the parameters \(\theta_1\), \(\theta_2\) and \(\sigma^2\). 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 us, some of which appear in Smith and Kohn (1996), 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 that we use for judging the quality of an estimate \(\hat{f}\) of an unknown function \(f\) is the integrated squared error ISE — the integral of \((f - \hat{f})^2\) over the domain of \(f\). We approximate this integral by taking a grid of 400 equally spaced points \(z_i = i/400, i = 1, \ldots, 400\), and compute ISE as

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

Fig. 1 presents the box plots of \(\log_{10}(\text{ISE})\) 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 \(u_i\) is 2.6 times the standard deviation of \(e_i\). 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 assess the performance of the full estimator visually (i.e. 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 Fig. 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 \(p(\gamma_1, \ldots, \gamma_4, \phi^{[2]}, \beta^{[2]} | y)\) 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 Fig. 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 et al. (1992). 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 et al. (1992) and the kernel approach by Hart (1994). Kohn et al. (1992) estimated the smoothing parameter and the autoregressive parameters by two methods: marginal likelihood and generalized cross-validation. Hart (1994) estimated the bandwidth by what he called time series cross-validation. We refer the reader to Kohn et al. (1992) and Hart (1994) for a description of their methods.

100 observations were generated from model (2.1) using the following three functions: the first function is \( f_1(x) = 32x^2(1-x)^2 \) which was used by Hart (1991) in his simulations; the other two functions \( f_2 \) and \( f_3 \) are described in equations (4.1). The independent variable was

---

**Fig. 1.** Box plots of log(ISE) for each of the three test functions and the three estimators: (a) \( f_1 \); (b) \( f_2 \); (c) \( f_3 \) (in each part, the box plot on the left corresponds to the estimator which assumes that the order of the autoregression is known to be 2, the middle box plot corresponds to the estimator which does not assume that the order of the autoregression is known and the right-hand box plot corresponds to the estimator which takes the errors to be independent)
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, \( \sigma \) takes three values corresponding to the standard deviation of \( e_t \) being an eighth, a quarter and a half the range of the function. These three values for the standard deviation of \( e_t \) represent low, medium and high noise examples.

100 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 at the true value, so that in the Bayesian case \( \kappa_1 = 1 \) and \( \kappa_i = 0 \) for \( i > 1 \). Fig. 3 presents box plots of \( \log(\text{ISE}) \) and shows that the Bayesian regression spline estimator compares favourably with the other
Fig. 3. Box plots of log(ISE) for each of the three test functions, three different levels and four different estimators: (a) $f_1$, high noise; (b) $f_2$, high noise; (c) $f_3$, high noise; (d) $f_1$, medium noise; (e) $f_2$, medium noise; (f) $f_3$, medium noise; (g) $f_1$, low noise; (h) $f_2$, low noise; (i) $f_3$, low noise (in each part, the estimators from left to right are regression splines, smoothing splines using generalized cross-validation, smoothing splines using marginal likelihood and Hart’s estimate).
estimators over all three functions and all three noise levels. It 'breaks down' (i.e. 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 Fig. 4 which plots, for each estimator, the function estimate corresponding to the 51st worst value of 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-hand side than on the right-hand side of the domain of $x$. The estimate of $\theta_1$ is 0.646, which is close to the true value of $\theta_1 = 0.5$.

The smoothing spline estimate using generalized 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 undersmoothing of the function. The estimate of $\theta_1$ is poor, with $\hat{\theta}_1 = -0.076$.

---

**Fig. 4.** Plots of the estimate of $f_5$ 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): (a) regression spline estimate; (b) smoothing spline estimate using generalized cross-validation; (c) smoothing spline estimate using marginal likelihood; (d) Hart's estimate.
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 oversmooths 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 $\theta_1$ is $-3.45$ (giving an explosive autoregression), 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. It should also be noted that estimating a function $f$ when time is the independent variable is more difficult than for a general regressor because there is more potential for confusing $f(t)$ and the time-based autocorrelation process.

4.3. Some 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 only considered a first-order autoregressive process for the errors but did 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) used a discretized version of spline smoothing by dividing the range of the independent variable into subintervals and assumed that the regression function is constant in each subinterval. They did not obtain a data-driven subdivision of the range of the independent variable. Harvey and Koopman (1993) used 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 did not have a data-driven approach for determining the position of the knots of the regression spline.

5. Additive semiparametric regression

5.1. Introduction

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

$$ y_t = f(x_t) + g(z_t) + u_t. $$

(5.1)

When the errors $u_t$ are independent, the back-fitting 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 back-fitting 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.

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

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

(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 equation (5.2) can be written as
By treating $f, g, f_1$ and $g_1$ as four separate functions the back-fitting algorithm can be applied to equation (5.3) to give estimators of $f$ and $g$. However, in empirical work using smoothing splines we have found that 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 back-fitting, 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 non-linear component is modelled as a regression spline with zero intercept and with a potential knot site every 15th 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 $X'\Omega_k^{-1}X$ from becoming singular or nearly singular. Nevertheless, the number of knots used appears more than sufficient to capture any potential non-linearity 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 $\kappa_i$ 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. (1996).

To obtain additivity of the regression function and normality of the intrinsic errors $e$, a normalized power transformation of the dependent variable is used as outlined in Section 2.3, i.e. $y_{i,\lambda} = a_\lambda + b_\lambda T_\lambda(y_i)$ with $T_\lambda(y) = y^\lambda$ and with $\lambda$ restricted to the nine values $\Lambda = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$. The inclusion of a trend term was suggested by looking at the residuals of the nonparametric fit of model (5.4) with the trend term omitted; see also Fig. 5(b) for a plot of the original data.

The Bayesian approach was applied with a warm-up period of 200 iterations and a further 100 iterations were used to estimate the posterior distribution of $\lambda$. The posterior probability of $\lambda = 0$ was 0.99 so the normalized 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 \( p \hat{i} \) and \( E \hat{j} y \) are given in Table 1 and suggest that \( ut \) is a 12th-order autoregression, which is consistent with monthly data. The estimate of the intercept \( \hat{0} = 4209 \) and the estimates of \( f_1, \ldots, f_4 \) are plotted in Figs 6(a)–6(d). These suggest that \( log y \) is linear in heating degree days and cooling degree days, but non-linear 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 Fig. 5(a) and is consistent with the plot of the dependent variable in Fig. 5(b).

To explain the scatterplots in Figs 5(a) and 6(a)–6(d) we need some additional notation. For \( t = 1, \ldots, n \), let \( \xi_t = u_t - E(u_t|u_1, \ldots, u_{t-1}) \), \( \sigma^2 R_t = \text{var}(\xi_t) \) and \( \zeta_t = \xi_t/\sqrt{R_t} \). Then \( \zeta_t \) 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 \( \xi_t \) and \( R_t \).
Table 1. Estimates of $p(\kappa_i = 1|y)$ ($\hat{\kappa}_i$) and $\phi$ for the residential electricity data

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\hat{\kappa}_i$</th>
<th>$\hat{\theta}_i$</th>
<th>$i$</th>
<th>$\hat{\kappa}_i$</th>
<th>$\hat{\theta}_i$</th>
<th>$i$</th>
<th>$\hat{\kappa}_i$</th>
<th>$\hat{\theta}_i$</th>
<th>$i$</th>
<th>$\hat{\kappa}_i$</th>
<th>$\hat{\theta}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.398</td>
<td>0.089</td>
<td>6</td>
<td>0.111</td>
<td>0.034</td>
<td>11</td>
<td>0.158</td>
<td>0.011</td>
<td>16</td>
<td>0.055</td>
<td>-0.002</td>
</tr>
<tr>
<td>2</td>
<td>0.992</td>
<td>-0.094</td>
<td>7</td>
<td>0.068</td>
<td>-0.131</td>
<td>12</td>
<td>0.544</td>
<td>0.047</td>
<td>17</td>
<td>0.047</td>
<td>-0.001</td>
</tr>
<tr>
<td>3</td>
<td>0.232</td>
<td>0.003</td>
<td>8</td>
<td>0.741</td>
<td>0.206</td>
<td>13</td>
<td>0.171</td>
<td>-0.024</td>
<td>18</td>
<td>0.055</td>
<td>0.005</td>
</tr>
<tr>
<td>4</td>
<td>0.202</td>
<td>-0.058</td>
<td>9</td>
<td>0.093</td>
<td>-0.126</td>
<td>14</td>
<td>0.071</td>
<td>-0.002</td>
<td>19</td>
<td>0.053</td>
<td>0.005</td>
</tr>
<tr>
<td>5</td>
<td>0.117</td>
<td>0.062</td>
<td>10</td>
<td>0.013</td>
<td>0.013</td>
<td>15</td>
<td>0.059</td>
<td>0.002</td>
<td>20</td>
<td>0.082</td>
<td>-0.009</td>
</tr>
</tbody>
</table>

Fig. 6. (a)–(d) Plots of the function estimates $\hat{f}_i(x_{i1}), \ldots, \hat{f}_i(x_{i4})$ (—) together with the added residual scatterplots $\hat{f}(x_{i1}) + \zeta_i, \ldots, \hat{f}(x_{i4}) + \zeta_i$; (e) autocorrelations of $\zeta_i$; (f) normal probability plot of $\zeta_i$
based on the estimate of $\psi$. If the fitted model is correct then $\hat{\zeta}_t = \hat{\xi}_t / \sqrt{R_t}$ is approximately $N(0, \sigma^2)$ and independent. The scatterplot in Fig. 6(a) is $\hat{f}_t(x_{1t}) + \hat{\zeta}_t$ against $x_{1t}$ and suggests that the effect of $x_{1t}$ on $\log(y_t)$ is captured correctly. The scatterplots in Figs 6(b)–6(d) and Fig. 5(a) are interpreted similarly and suggest that the effects of $x_{2t}$, $x_{3t}$, $x_{4t}$ and the time trend are also estimated correctly. Fig. 6(e) is a plot of the sample autocorrelations of $\hat{\zeta}_t$ and Fig. 6(f) is a normal probability plot of $\hat{\zeta}_t$. These plots indicate that $\hat{\zeta}_t$ is independent and normal and thus validate the regression assumptions.

Fig. 5(b) is a time plot of the dependent variable and the Bayesian fit showing that the effects of $\hat{\zeta}_t$ and $\hat{\eta}_t$ track the data well on the original scale. Fig. 5(c) plots the track of the data well on the original scale. Fig. 5(c) plots the

5.3. Tooth-paste data

As a second example we consider the tooth-paste data previously analysed by Wichern and Jones (1977). These data include 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 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 to ensure that the regression assumptions are satisfied. Nine candidate transformations $T_\lambda(y)$, $\lambda = 1, \ldots, 9$, are considered and listed in Table 2 together with the normalization constants $a_\lambda$ and $b_\lambda$ which ensure that the median and the inter-quartile range are similar for each transformation. Thus the model that we 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 $\alpha_1$ and $\alpha_2$ 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 15th value of the ranked independent variable $P_t$.

The sampling scheme in Section 3.1 was run for a warm-up period of 100 iterations with a

Table 2. Candidate transformations $T_\lambda(y)$, normalizing constants $a_\lambda$ and $b_\lambda$, and the posterior probability estimate of each transformation $\hat{p}(\lambda | y)$ for the tooth-paste data

| $\lambda$ | $T_\lambda(y)$ | $a_\lambda$ | $b_\lambda$ | $\hat{p}(\lambda | y)$ |
|-----------|----------------|-------------|-------------|---------------------|
| 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         | $\log(y/(1-y))$  | 0.4434      | 0.1763      | 0.0001              |
| 9         | $\log(-\log(1-y))$ | 0.5007      | 0.2014      | 0.0                     |

$\Phi$ is the cumulative distribution function of the standard normal.
further 100 iterations used to estimate λ and a final 400 iterations to estimate α₁, α₂, f, κ and θ. Sequence plots of the parameter values and posterior probability indicate that the sampling scheme converged within a handful of iterations. Table 2 presents the estimate of the marginal posterior distribution of λ and shows that the mode occurs for Tₙ(yₙ) = Φ⁻¹(yₙ0.1), where Φ(·) is the standard normal cumulative distribution function. The rest of the estimates were calculated conditionally on λM as outlined in Section 3.2.

The estimate of E(κ|y) is

\[ \hat{\phi}(κ = 1|y) = (1, 1, 0.886, 0.361, 0.149, 0.041, 0.014, 0.0104), \]

suggesting that either an autoregressive AR(3) or an AR(4) model for the errors is adequate. The estimates of α₁ and α₂ are (α₁, α₂) = (0.3063, 0.5113), whereas the estimate of θ is

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

Fig. 7(a) is a plot of the estimated curve \( \hat{f}(P_t) \) together with the added residual scatterplot \( \hat{y}_t \hat{f}(P_t) \), where \( \hat{y}_t \) is defined in Section 5.2. From this plot, we conclude that \( \hat{y}_t \) is roughly linear in \( P_t \) and that \( P_t \) is a relatively insignificant determinant of the market share of Crest.

Fig. 7(b) is a time plot of the fitted values \( \Phi(\hat{y}_{t,\lambda} - a_\lambda) / a_\lambda \) and the actual observations on the original scale. The fitted values track the data well and capture the discontinuity of the intervention, suggesting that the model provides a good fit to the data. Fig. 8(a) is a plot of the autocorrelations of \( \hat{y}_t \) and Fig. 8(b) is a normal probability plot of \( \hat{y}_t \). These plots suggest that the errors \( u_t \) are stationary and normally distributed and follow a fourth-order autoregressive process.
6. Summarizing remarks

A Bayesian approach was proposed for nonparametric regression with autocorrelated errors. The framework can be employed with a variety of bases, but we have shown that with a simple cubic regression spline basis it compares favourably with other nonparametric approaches in the univariate case with time as the independent variable. 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 is 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.

Acknowledgements

Robert Kohn’s research was partially supported by an Australian Research Council grant. We thank Jeff Hart for permission to use his program, Steve Marron for comments and Denzil Fiebig for suggesting the electricity consumption data set.

Appendix A: Implementation of the sampler

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

\[ S(\gamma, \psi, \lambda) = y_\lambda' \Omega_\psi^{-1} y_\lambda - \frac{c}{1+c} y_\lambda' \Omega_\psi^{-1} X_\gamma (X_\gamma' \Omega_\psi^{-1} X_\gamma)^{-1} X_\gamma' \Omega_\psi^{-1} y_\lambda, \]

\[ \beta_\gamma = \frac{c}{1+c} (X_\gamma' \Omega_\psi^{-1} X_\gamma)^{-1} X_\gamma' \Omega_\psi^{-1} y_\lambda, \]

let \( q \) be the number of elements of \( \beta_\gamma \) and let \( J_\lambda \) be the absolute value of the determinant of the Jacobian matrix for the transformation \( y \rightarrow y_\lambda \). From the assumptions in Section 2,

\[ p(y|\beta_\gamma, \gamma, \psi, \sigma^2) p(\beta_\gamma|\psi, \gamma, \sigma^2) p(\sigma^2) \propto (2\pi)^{-(n+q)/2} (\sigma^2)^{-(n+q)/2-1} e^{-\psi/2} |X_\gamma' \Omega_\psi^{-1} X_\gamma|^{1/2} J_\lambda \]

\[ \times \exp \left[ -\frac{1}{2\sigma^2} \left\{ S(\gamma, \psi, \lambda) + \frac{c}{1+c} (\beta_\gamma - \beta_\gamma') X_\gamma' \Omega_\psi^{-1} X_\gamma (\beta_\gamma - \beta_\gamma') \right\} \right]. \]

(A.1)
A.1. Generating $\beta_i$

\[ p(\beta_i, \sigma^2 | y, \gamma, \psi, \lambda) \propto p(y | \beta_i, \gamma, \psi, \sigma^2, \lambda) p(\beta_i | \gamma, \psi, \sigma^2) p(\sigma^2). \]  

(A.2)

Integrating $\sigma^2$ out of expressions (A.1) and (A.2) using an inverse gamma integral,

\[ p(\beta_i | y, \gamma, \psi, \lambda) \propto \left\{ S(\gamma, \psi, \lambda) + \frac{c}{1+c} (\beta_i - \bar{\beta}_i)^T \Sigma_i^{-1} X_i (\beta_i - \bar{\beta}_i) \right\}^{-(n+q_i)/2}, \]

i.e. the conditional distribution of $\beta_i$ is multivariate $t$ with degrees of freedom $n$, which is centred at $\bar{\beta}_i$ and has scale matrix $(X_i^T \Sigma_i^{-1} X_i)^{-1}(c+1) S(\gamma, \psi, \lambda)/nc$;

for $n$ large, this conditional distribution is effectively multivariate normal.

A.2. Generating $\psi_i$ and $\kappa_i$

The variables $\psi_i$ and $\kappa_i$ are generated as a block with $\sigma^2$ integrated out. Note that

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

We sketch out the necessary steps to generate $\psi_i$ and $\kappa_i$; the details are in Barnett et al. (1996).

\[ p(\psi_i, \kappa_i = 1, \sigma^2 | u, \psi_{j\neq i}) \propto p(u | \psi_i, \kappa_i = 1, \sigma^2) p(\psi_i | \kappa_i = 1) p(\sigma^2) \]

\[ \propto \frac{1}{2} p(\kappa_i = 1)(\sigma^2)^{-n/2} |\Omega_v|^{-1/2} \exp(-u^T \Omega_v^{-1} u/2). \]

Integrating $\sigma^2$ out using an inverse gamma integral,

\[ p(\psi_i, \kappa_i = 1 | u, \psi_{j\neq i}) \propto \frac{1}{2} p(\kappa_i = 1)|\Omega_v|^{-1/2} (u^T \Omega_v^{-1} u)^{-n/2}. \]

The binary variable $\kappa_i$ is generated first without conditioning on $\psi_i$. To integrate out $\psi_i$ we proceed as follows. Let $\xi_i$ and $R_i$ be defined as in Section 5.2. Then,

\[ u^T \Omega_v^{-1} u = \sum_{i=1}^{n} \xi_i^2 / R_i, \]

the $\xi_i$ are independent and, for $t > s$, $\xi_s = u_t - \theta_s u_{t-s} \ldots - \theta_0 u_{t-s}$ and $R_i = 1$. Barnett et al. (1996) showed that, for $t > s$, $\xi_t$ is linear in $\psi_i$ so it is possible to write $\Sigma_{i=1}^{n} \xi_i^2 = A(\psi_i - B)^2 + C$, with $A$, $B$ and $C$ independent of $\psi_i$, $A$ and $C$ positive, and all three terms computed as in Barnett et al. (1996). Let $D(\psi_i) = \Sigma_{i=1}^{n} \xi_i^2 / C$. Then $D \ll 1$ as its numerator is the sum of $s$ terms whereas its denominator is the sum of $n-s$ terms. Thus,

\[ (u^T \Omega_v^{-1} u)^{-n/2} \approx C^{-n/2} \left\{ 1 + A(\psi_i - B)^2 / C \right\}^{-n/2} \approx C^{-n/2} \exp(-nA(\psi_i - B)^2 / 2C). \]

Let

\[ g(\psi_i) = |\Omega_v|^{-1/2} \left\{ D + 1 + A(\psi_i - B)^2 / C \right\}^{-n/2} (C/nA)^{1/2} \phi(\psi_i; B, C/nA)^{-1}, \]

where $\phi(x, \mu, \sigma^2)$ is the normal density evaluated at $x$ with mean $\mu$ and variance $\sigma^2$. Then,

\[ p(\kappa_i = 1 | u, \psi_{j\neq i}) \propto \frac{1}{2} p(\kappa_i = 1) \int_{-\infty}^{\infty} g(\psi_i) \phi(\psi_i; B, C/nA) \, d\psi_i. \]  

(A.3)

The integral in expression (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,

\[ p(\kappa_i = 0 | u, \psi_{j\neq i}) \propto p(\kappa_i = 0)|\Omega_v|^{-1/2} \left\{ D + 1 + A(\psi_i - B)^2 / C \right\}^{-n/2}. \]  

(A.4)

with expression (A.4) evaluated at $\psi_i = 0$. The conditional probability of $\kappa_i$ is obtained by normalizing
expressions (A.3) and (A.4). Once \( \kappa \) has been generated, \( \psi_i = 0 \) if \( \kappa_i = 0 \). If \( \kappa_i = 1 \), we generate \( \psi_i \) from

\[ q(\psi_i) = \phi(\psi_i, B, C/\Lambda) \]

and use a Metropolis–Hastings step.

### A.3. Generating \( \gamma_i \)

\[
p(\gamma_i, \beta_i, \sigma^2, \lambda|y, \gamma_{j \neq i}, \psi) \propto p(y|\beta_i, \gamma_i, \sigma^2, \lambda) p(\beta_i|\psi, \gamma_i, \sigma^2) p(\sigma^2) p(\lambda) p(\psi_i).
\]

(A.5)

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

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

(A.6)

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

### A.4. Generating \( \lambda \)

\[
p(\lambda|y, \gamma_i, \psi) \propto p(y|\gamma_i, \psi, \lambda) p(\lambda) \propto S(\gamma, \psi, \lambda)^{-n/2} J_\lambda p(\lambda).
\]

(A.7)

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

We conclude by showing how to compute \( y_j' \Omega^{-1}_v y_j, y_j' \Omega^{-1}_v X_j, X_j' \Omega^{-1}_v X_j \) and \( |\Omega_v| \) efficiently. As in Barnett et al. (1996), for \( t = 1, \ldots, s \),

\[
\xi_t = u_t - \theta_{t,1} u_{t-1} - \cdots - \theta_{t,t-1} u_1,
\]

\[
R_t = [(1 - \psi_{j_t}^2) \cdots (1 - \psi_{j_t}^2)]^{-1}.
\]

(A.8)

Let \( M \) be a lower triangular matrix with 1s on the diagonal; for \( t = 1, \ldots, s \), \( M_{t,t-1} = -\theta_{t,j} \), \( j = 1, \ldots, t-1 \); for \( t > s \), \( M_{t,t-1} = -\theta_{t,j} \), \( j = 1, \ldots, s \), and \( M_{t,t} = 0 \) for \( j > s \). Thus \( M \) is a lower triangular band matrix with bandwidth at most \( s \). From expression (A.8) and the definition of \( \xi_t \) and \( R_t \), \( M\xi = \xi \) and \( M\Omega_v M' = R \), where \( R \) is a diagonal matrix with \( t \)th diagonal element \( R_t \). Let \( \tilde{y} = My \) and \( \tilde{X} = MX \); then, \( y_j' \Omega^{-1}_v y_j = \tilde{y}' R^{-1} \tilde{y}, y_j' \Omega^{-1}_v X_j = \tilde{y}' R^{-1} \tilde{X}, X_j' \Omega^{-1}_v X_j = \tilde{X}' R^{-1} \tilde{X} \) and \( |\Omega_v| = \Pi_{t=1}^s R_t \).

### References


