

Bayesian Models for Non-Linear Autoregressions

PETER MÜLLER & MIKE WEST

Institute of Statistics & Decision Sciences,
Duke University, Durham NC 27708-0251

and

STEVEN MACEACHERN

Department of Statistics, Ohio State University, Columbus OH

Peter Müller was partially supported by NSF under grant DMS-9404151, and Mike West under grant DMS-9024793. Much of this work was developed while Steve MacEachern was visiting Duke University during 1993-4.

Abstract. We discuss classes of Bayesian mixture models for non-linear autoregressive times series, based on developments in semi-parametric Bayesian density estimation in recent years. The development involves formal classes of multivariate discrete mixture distributions, providing flexibility in modelling arbitrary non-linearities in time series structure and a formal inferential framework within which to address the problems of inference and prediction. The models relate naturally to existing kernel and related methods, threshold models and others, though offer major advances in terms of parameter estimation and predictive calculations. Theoretical and computational aspects are developed here, the latter involving efficient simulation of posterior and predictive distributions. Various examples illustrate our perspectives on identification and inference using this mixture approach.

Keywords: autoregressive time series; Bayesian computations; Mixture models; Non-linear time series

1 INTRODUCTION

Recent developments in Bayesian density estimation using mixture models have produced flexible and implementable methods for inferring features of multivariate distributions and estimating non-linear regression relationships (West, Müller and Escobar 1994). This paper parallels that development by introducing novel mixture models in the time series domain, presenting approaches to non-linear modelling and analysis of autoregressive time series.

In a very real sense the development leads to an encompassing framework for non-linear autoregressions, offering flexibility to represent arbitrary forms of non-linear conditional expectations and, more generally and practically critically, conditional distributions. Here, this is exclusively in the context of real-valued, univariate time series, though the conceptual basis anticipates possible future developments to discrete data models and multivariate series. Thus, for example, non-linear predictors based on threshold autoregressions, non-parametric kernel approaches, neural networks and so forth, fall within the ambit of the framework here. From a methodological viewpoint, the computational Bayesian development via stochastic simulation of posterior and predictive distributions permits routine implementation and model fitting and assessment, as the examples in the final section demonstrate.

The models are based on mixtures of normal distributions, obviously linked, in terms of point estimates of conditional AR predictions, with non-parametric kernel regression approaches. However, the framework here involves a complete specification of sequences of conditional distributions, rather than just ad-hoc point estimates, with the facility for automatic smoothing parameter estimation embedded in the general Bayesian learning framework, and delivers full probabilistic measures of relevant uncertainties as well as point estimates. In addition, a rather nice feature is that the model “homes in” on sub-models based on small numbers of mixture components (and hence, crudely speaking, the number of distinct kernel components), offering automatic reduction of model dimension and complexity in the con-

text of realised time series that support simpler model forms.

Conditional expectations take the form of locally weighted mixtures of linear (auto-)regressions. The model puts a hierarchical prior on the size and terms of the mixture. Similar models, albeit naturally without the hierarchical mixture model prior, have been proposed in recent non-Bayesian literature. These include threshold autoregressive models, hidden Markov chain autoregressive models, non-linear additive AR models, threshold and smooth threshold AR models, and the general class of state-dependent models; For a review of these models and basic concepts of non-linear, non-stationary time series analysis see, for example, Priestley (1988), Tong (1990) and the references given there.

Bayesian approaches for some of these models have been proposed. Geweke and Terui (1991) develop Bayesian inference on threshold autoregressive models. Albert and Chib (1992) discuss hidden Markov chain autoregressive models. Some recent papers dealing with the related class of non-normal, non-linear state space models are Carlin, Polson and Stoffer (1992), Jacquier, Polson and Rossi (1994) and Carter and Kohn (1994). Such models may be used to express specific cases of the non-linear autoregressive relations considered in this paper. However, they differ from the models considered in this paper by defining dynamic parameters which evolve over time rather than a semi-parametric framework based on introducing hierarchical prior models on static parameters.

Section 2 develops the basic mixture modelling framework, introducing Dirichlet process priors as one basis for defining mechanisms for model switching in classes of standard normal/linear autoregressions. Extensions to models including non-linear regressions on exogenous covariates are included. Section 3 deals with computational Bayesian analysis and provides algorithms for fitting the models based on stochastic simulation methods that have their roots in related algorithms in density estimation contexts. Section 4 presents a simulation study to explore long run performance of the proposed models and two examples based on the Canadian lynx data and the Old Faithful geyser data, with some general summary comments.

2 MIXTURE MODELLING FRAMEWORK

General background to Bayesian mixture modelling in density estimation and regression appears in West, Müller and Escobar (1994), with more specific regression based material in Müller, Erkanli and West (1996). Some of the basic theory of Dirchlet mixture models is relevant here, and we discuss only the relevant components of that theory.

2.1 Introduction in non-linear AR(1) contexts

Consider a univariate, real-valued time series y_t to be observed at equally spaced time points $t = 1, 2, \dots, T$ and suppose interest lies in developing models for the conditional distributions under an AR(1) assumption; thus focus rests on density functions $p(y_t|y_{t-1})$. Note that we make no stationarity assumptions, though data may ultimately turn out to be consistent with stationarity. Several practically viable approaches to modelling non-linear structure are based on mixtures of normal, linear AR(1) models. The same is true here.

We assume that there exist some number $n > 0$ of possible linear models characterised by parameters $\theta_i = (\beta_{0i}, \beta_{1i}, w_i, \mu_i)$, ($i = 1, \dots, n$), such that y_t may be generated from one of these models. Denote the model selected for y_t by $r_t = i$. In particular, we assume

$$p(y_t|y_{t-1}, r_t = i, \theta_i) = N(y_t; \beta_{0i} + \beta_{1i}y_{t-1}, w_i), \quad (1)$$

where r_t is selected from a set of possible models, $i = 1, \dots, n$, with selection probabilities,

$$Pr(r_t = i|\theta_1, \dots, \theta_n, y_{t-1}) \propto \exp(-(y_{t-1} - \mu_i)^2/(2V)). \quad (2)$$

Here V is an additional hyperparameter, and $N(x; m, s)$ indicates that the random variable x has a normal distribution with mean m and variance S . Also, we assume that y_0 is either known or an informative prior $p(y_0)$ is available. Model (1) and (2) defines a locally weighted mixture similar to the idea of threshold autoregression. While in threshold autoregression

for any given value of the lagged variables y_{t-1} only one linear submodel applies with sudden changes at the thresholds, the locally weighted mixture provides for a smooth change between linear submodels as in STAR models (Tong 1990). Submodel i receives maximum weight for y_{t-1} around μ_i with the relative weight falling off like a Gaussian kernel.

A mixture prior on the set of possible models $\theta = \{\theta_i, i = 1, \dots, n\}$ is induced by building on the hierarchical mixture model structure inherent in the Bayesian mixture models of West, Müller and Escobar (1994). Let δ_x denote a point mass at x . Specifically, we assume that

$$\theta_i \stackrel{\text{i.i.d.}}{\sim} \underbrace{\sum_{j=1}^{\infty} u_j \delta_{\theta_j^*}(\theta_i)}_{G(\theta_i)}, \quad i = 1, \dots, n, \quad (3)$$

$$G \sim DP(G; \alpha \mathcal{G}_\nu).$$

where G is a random discrete distribution, modeled via a Dirichlet process (DP) prior with base measure $\alpha \mathcal{G}_\nu$, denoted here by $DP(G; \alpha \mathcal{G}_\nu)$. Here \mathcal{G}_ν is a probability distribution with hyperparameters ν and α is the total mass parameter of the DP that describes the concentration of the DP prior around the prior expectation \mathcal{G}_ν . For full mathematical details of the DP model see, for example, Antoniak (1974). Some key features and results relevant to model (3) will be discussed below. Later we will complete the model by assuming hyperpriors on ν and α . Denote with F_G and $F_{\mathcal{G}_\nu}$ the c.d.f. corresponding to G and \mathcal{G}_ν , respectively. The DP prior samples a random discrete measure G by generating a step function F_G centered around $F_{\mathcal{G}_\nu}$. The step sizes u_j , i.e., how close the discrete c.d.f. F_G is to $F_{\mathcal{G}_\nu}$, is determined by the total mass parameter α . Large values for α lead to random c.d.f.'s F_G very close to $F_{\mathcal{G}_\nu}$. Smaller values of α correspond to more variation around $F_{\mathcal{G}_\nu}$. The assumption of the DP prior is not critical in our model, and alternative prior models on the weights u_j and the locations θ_j^* , possibly using a finite number k of point masses, would lead to very similar inference. We utilize the Dirichlet process prior formulation because of the straightforward interpretation of the parameters \mathcal{G}_ν and α in the DP, and the

fact that it has proven so useful in other applications (Bush and MacEachern 1996; Escobar and West 1995; MacEachern 1994; Müller, Erkanli and West 1996; West, Müller and Escobar 1994). The computational issues in the implementation are very similar to mixture models with other forms of prior distributions. Some key features and results relevant to using (3) as a prior model for the mixture are summarised here.

First, the set of model vectors $\theta_i = (\beta_{0i}, \beta_{1i}, w_i, \mu_i)$, $i = 1, \dots, n$, contains some $k \leq n$ distinct quadruples $\theta_j^* = (\beta_{0j}^*, \beta_{1j}^*, w_j^*, \mu_j^*)$, $j = 1, \dots, k$. It is useful to introduce configuration indicators $\mathcal{S} = \{s_1, \dots, s_n\}$ such that $s_i = j$ if and only if $\theta_i = \theta_j^*$. Thus the n models are arranged into some $k \leq n$ different groups, each group having its own regression line and selection probability parameter μ_j^* . Count the relative numbers in each distinct set by defining $n_j = \#\{i \in (1, \dots, n) | s_i = j\}$, noting that $n = n_1 + \dots + n_k$. There exists a prior distribution, implicit in the underlying Dirichlet process theory, generating both the number k of distinct model vectors and the mechanism by which the configuration indicators are chosen. The prior for k is Poisson-like, determined by the single precision parameter $\alpha > 0$ (and the sample size n), as discussed further below. Then, given k and marginalizing over θ_j^* , $j = 1, \dots, k$, the configuration indicators s_i follow the marginal prior $p(\mathcal{S} | k, n) \propto \prod_{j=1}^k (n_j - 1)!$ (see, for example, MacEachern and Müller, 1994).

Second, the θ_j^* are an i.i.d. sample from \mathcal{G}_ν , i.e. marginally each θ_i is sampled from \mathcal{G}_ν .

Third, the model (1) – (3) produces a predictive distribution for y_{T+1} which, conditional on y_T and the list of model vectors θ , takes the form of a locally weighted mixture of normal, linear regressions. Thus

$$p(y_{T+1} | y_T, \theta, V) = \sum_{j=1}^k q_j N(y_{T+1}; \beta_{0i}^* + \beta_{1i}^* y_T, w_i^*), \quad (4)$$

with $q_j \propto n_j \exp(-(y_T - \mu_j^*)^2 / (2V))$ for $j = 1, \dots, k$. In words, to generate y_{T+1} the model picks one of k normal, linear regressions, with respective probabilities depending on y_T via a Gaussian kernel weight function. Equation (4) obviously connects with traditional non-linear, non-parametric autoregression concepts (e.g. Priestley 1988) though it is here embedded in

a coherent framework allowing formal parameter estimation in a Bayesian context. In problems where the underlying autoregression of $p(y_t|y_{t-1})$ is very irregular, non-linear, perhaps non-normal and possibly multimodal, the value of k will tend to be large, and the slope and intercept parameters $\beta_{0j}^*, \beta_{1j}^*$ widely dispersed. The corresponding features of the local regression lines may be quite distinct, providing varying structure as well as differing degrees of smoothing in different regions of the sample space. Furthermore, the weights of components, determined by the n_j , adapt to varying concentrations of mass. Typically, k will be moderate, of the order $\alpha \log(n)$. It should be stressed, however, that k will grow towards n as context and observed data configurations demand.

Formally, the required Bayesian predictive distribution is obtained by averaging (4) with respect to the posterior distribution for θ and V , given the data $y = (y_1, \dots, y_T)$, i.e.

$$p(y_{T+1}|y_T,) = \int p(y_{T+1}|y_T, \theta, V) dP(\theta, V|y). \quad (5)$$

By including a hyperprior on (α, ν) the model will allow for learning about α and any uncertain features of \mathcal{G}_ν in addition to the primary uncertainties about k, S and the θ_j vectors. We return to this below after equation (8).

The computations required to evaluate the necessary posterior $P(\theta, V|y)$ and perform the integration in (5) are made possible via methods of stochastic simulation developed and exemplified in West, Müller and Escobar (1994), MacEachern (1994), Bush and MacEachern (1996), MacEachern and Müller (1994) and Müller, Erkanli and West (1996), based on original work in the univariate modelling context in Escobar and West (1995). An important aspect in the computation is the stage where the θ_j^* are moved, as introduced in Bush and MacEachern (1996) and detailed in steps (iii) and (iv) of the algorithm in Section 3.1. Simulation analysis delivers sequences of values for the parameters θ and V representing draws from the posterior $P(\theta, V|y)$. Hence Monte Carlo approximation of the central distribution (5) may be deduced by replacing the integral with a Monte Carlo summation. Specifically, given posterior samples $\theta^{(m)}, V^{(m)}, m = 1, \dots, M$, (5) is

approximated by

$$p(y_{T+1}|y_T, y) \approx M^{-1} \sum_{m=1}^M p(y_{T+1}|y_T, \theta^{(m)}, V^{(m)}). \quad (6)$$

The applications in the above references concern estimation of independent sampling models. For normal models with no covariates, this reduces to replacing (1) and (2) by $p(y_t|\theta) = N(y_t; \mu_t, \Sigma_t)$ with $\theta_t = (\mu_t, \Sigma_t)$, $t = 1, \dots, T$. In particular the additional mixture in (2) is not present. Although many details are different, the basic principle of using a Dirichlet process model to parameterize the mixture and the main issues in building the Markov chain Monte Carlo scheme to estimate the model are the same.

2.2 The general autoregressive mixture model

Model (1) – (3) is easily extended to include higher order autoregression and additional covariates. Let $x_t = (y_{t-1}, \dots, y_{t-p}, z_t)$ denote a $(p + q)$ dimensional vector of p lagged observations and an additional q -dimensional exogenous covariate vector z_t , and write $\phi(x; m, S)$ for the $(p + q)$ dimensional Gaussian kernel with moments m and S , evaluated at x . As before, we assume that initial values y_{1-j} , $j = 1, \dots, p$ are either known or an informative prior is available. Model (1) – (3) is replaced by

$$p(y_t|x_t, \theta, V) \propto \sum_{i=1}^n \phi(x_t; \mu_i, V) \cdot N(y_t; \beta_i' x_t, w_i), \quad (7)$$

or, equivalently,

$$(y_t|x_t, r_t = i, \theta_i) \sim N(y_t; \beta_i' x_t, w_i)$$

and

$$Pr(r_t = i|\theta, V, x_t) \propto \phi(x_t; \mu_i, V). \quad (8)$$

We complete the model description by specifying \mathcal{G}_ν and hyperpriors on V, ν and α . We assume that

$$\theta_i \stackrel{\text{i.i.d.}}{\sim} G, \quad i = 1, \dots, n,$$

$$\begin{aligned}
G &\sim DP(G; \alpha \mathcal{G}_\nu), \\
\mathcal{G}_\nu(\beta, w, \mu) &= N(\beta; m_\beta, B_\beta) \cdot \\
&\quad Ga(w^{-1}; a_w/2, b_w/2) \cdot N(\mu; m_\mu, B_\mu), \\
m_\beta &\sim N(m_\beta; a, A), \\
B_\beta^{-1} &\sim W(B_\beta^{-1}; c, (cC)^{-1}), \\
V^{-1} &\sim W(V^{-1}; q, (qR)^{-1}), \\
\alpha &\sim Ga(\alpha; a_\alpha, b_\alpha),
\end{aligned} \tag{9}$$

where $W(\cdot; r, a)$ denotes a Wishart distribution with scalar parameter r and matrix parameter A , and $Ga(\cdot; a, b)$ denotes a gamma distribution with shape a and scale b .

As in the special case of the non-linear AR(1) model, the predictive distributions take the form of a locally weighted mixture of linear regressions:

$$p(y_{T+1} | x_{T+1}, \theta, V) \propto \sum_{j=1}^k n_j \phi(x_t; \mu_j^*, V) N(y_{T+1}; \beta_j^* x_{T+1}, w_j^*). \tag{10}$$

The Dirichlet process model puts a prior probability model on the size k and the individual terms $\theta_j^* = (\beta_j^*, w_j^*, \mu_j^*)$ of the mixture, as well as the process of allocating parameters θ_i to common values θ_j^* .

3 COMPUTATIONS

3.1 A Markov chain Monte Carlo scheme

To estimate the autoregressive mixture model (7) with prior (9) we implement a Markov chain Monte Carlo (MCMC) scheme by an appropriate combination of Gibbs sampling, independence chain and Metropolis steps. For a review of Markov chain Monte Carlo schemes see, for example, Gelfand and Smith (1990), Tierney (1994), Smith and Roberts (1993) or Gilks et al. (1993) and the references listed there. The basic rationale of Markov chain Monte Carlo posterior inference is to simulate a Markov chain which is defined to have the desired posterior as its limiting distribution. By taking

ergodic averages, as in (6), we can evaluate (almost) any posterior inference. We describe an MCMC scheme suited for model (7).

Given currently imputed values for the unknown parameters $(\mathcal{S}, \theta^*, V, \alpha, m_\beta, B_\beta)$ we move to the next iteration of the Markov chain by replacing each of the parameters via steps (i) through (ix) described below. We will write $\theta^{(i)}$ for $\theta^{(i)} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$, $k^{(i)}$ for the number of distinct elements in $\theta^{(i)}$, $\mathcal{S}^{(i)}$ for $\mathcal{S} \setminus \{s_i\}$, and $n_j^{(i)}$ for the number of indices $l \neq i$ with $s_l = j$. Also, $y = (y_1, \dots, y_T, z_1, \dots, z_T)$ denotes the full data set, and we write $p(y|\theta, V) = \prod_{t=1}^T p(y_t|x_t, \theta, V)$ for the likelihood (7). Remember that $\theta^* = \{\theta_1^*, \dots, \theta_k^*\}$ is the set of unique elements in $\{\theta_1, \dots, \theta_n\}$. The indicators s_i map θ^* into θ by $\theta_i = \theta_j^*$ if $s_i = j$. Each vector θ_j^* is partitioned into $\theta_j^* = (\beta_j^*, w_j^*, \mu_j^*)$. The list $\beta^* = (\beta_1^*, \dots, \beta_k^*)$ collects all β_j^* , and similarly for μ^* and w^* . In step (i) we will define an additional latent parameter vector $\eta = \{\theta_{k+1}^*, \dots, \theta_n^*\}$. Finally, remember that ν denotes the set of hyperparameters (m_β, B_β) .

Before explicitly describing the transition probability we give an outline by schematically listing the updating sequence. Each item of the form $X|Y, Z$ indicates that parameter X is being updated with parameters Y, Z kept fixed. Absence of a parameter (or data) W in the conditioning set indicates that either X and W are conditionally independent given Y, Z , or that we marginalize over W when updating X . The detailed description below will point out when we are using marginalization. Fixed hyperparameters such as m_μ or covariates z_t are not listed in the conditioning set. Wherever possible updating is done by a draw from the conditional posterior distribution. For some parameters, however, we will have to resort to a Metropolis or independence chain step, as indicated below. Invariance of the posterior distribution under transitions of the defined Markov chain is guaranteed since we only use Gibbs, independence chain and Metropolis steps. See, for example, Tierney (1994). The last step (ix) concerns resampling a latent parameter vector η which will be defined in (i). The set of model vectors $\{\theta_i, i = 1, \dots, n\}$ can alternatively be parameterized by $\{\theta_1^*, \dots, \theta_k^*, s_1, \dots, s_n\}$. In the following we will use whichever parameteri-

zation makes notation clearer in a given expression. The sequence is:

- (i) $s_i | \theta^*, \mathcal{S}^{(i)}, V, \eta, y, \quad i = 1, \dots, n$
- (ii) $r_t | \theta^*, \mathcal{S}, V, y, \quad t = 1, \dots, T$
- (iii) $\beta_j^*, w_j^* | \mathcal{S}, r, \nu, y, \quad j = 1, \dots, k$
- (iv) $\mu_j^* | \beta^*, \mu_1^*, \dots, \mu_{j-1}^*, \mu_{j+1}^*, \dots, \mu_k^*, w^*, \mathcal{S}, V, y, \quad j = 1, \dots, k$
- (v) $V | \theta^*, \mathcal{S}, y$
- (vi) $\alpha | \theta^*, \mathcal{S}$
- (vii) $m_\beta | \beta^*, B_\beta$
- (viii) $B_\beta | \beta^*, m_\beta$
- (ix) $\eta | \nu, k$

Note that in step (ii) the parameter vector is augmented by the latent indicator variables r_t . Sampling β^* and w^* we condition on r . Starting with step (iv) we again marginalize over r . In step (ix) the parameter vector is augmented by $\eta = \{\theta_{k+1}^*, \dots, \theta_n^*\}$ on which we condition in (i). In (ii)–(viii) we again marginalize over η .

Steps (i), (ii), (iii), (vii), (viii) and (ix) will be draws from the complete conditional posterior, i.e. straightforward Gibbs sampling steps. Step (iv) will be a Metropolis step. Step (v) will be an independence chain step. Step (vi) will be draws from the complete conditional posterior under an appropriate data augmentation. Each step is now described in detail.

(i) Sampling $s_i \sim p(s_i | \theta^{*(i)}, \mathcal{S}^{(i)}, V, \eta, y)$. To motivate the latent variable scheme which we will introduce below, we start by considering resampling θ_i , and thus implicitly s_i , from the conditional posterior $p(\theta_i | \theta^{(i)}, V, \alpha, \nu, y)$. The new value of θ_i is either set equal to $\theta_j, j \neq i$ with probability $q_j = c p(y | \theta^{(i)}, \theta_i = \theta_j, V)$, or is equal to a new draw from

$$G_i(\tilde{\theta}) \propto \mathcal{G}_\nu(\tilde{\theta}) p(y | \theta^{(i)}, \theta_i = \tilde{\theta}, V)$$

with probability $q_0 = c \alpha \int p(y | \theta^{(i)}, \theta_i = \tilde{\theta}) d\mathcal{G}_\nu(\tilde{\theta})$. Here c is the appropriate normalizing constant. Note that the distribution G_i is simply the posterior on θ_i under prior \mathcal{G}_ν . Combining identical θ_i 's and defining

$q_j^* = c p(y|\theta^{(i)}, \theta_i = \theta_j^*, V)$ this can be written as:

$$p(\theta_i|\theta^{(i)}, \mathcal{S}^{(i)}, k^{(i)}, \alpha, \nu, V, y) = \sum_{j=1}^{k^{(i)}} n_j^{(i)} q_j^* \delta_{\theta_j^*} + q_0 G_i(\theta_i). \quad (11)$$

Direct simulation from (11) is unfortunately hindered by the integral expression for q_0 . However, MacEachern & Müller (1994) (MM) show that the following scheme is possible: Extend the list of cluster locations to include $\eta = (\theta_{k+1}^*, \dots, \theta_n^*)$. These are “empty” clusters, i.e., $n_j = 0$, $j > k$. MM specify a prior distribution on the augmented parameter vector $(\theta_1^*, \dots, \theta_n^*, \mathcal{S})$ which induces the same prior distribution as model (3) on the original parameterization. Note that the definition includes a constraint that the cluster locations θ_i^* be indexed such that the non-empty clusters come before the empty ones. MM derive the full posterior conditionals required for Gibbs sampling in this model. Rearrange the indices of θ_j^* and correspondingly redefine \mathcal{S} such that $s_i = k$. Drawing from the conditional (11) is replaced by simulating from the following conditional posterior distribution. Note that resampling θ_i conditional on $(\theta_1^*, \dots, \theta_n^*)$ will change s_i only.

$$s_i | (\theta_1^*, \dots, \theta_n^*, \mathcal{S}^{(i)}, V, y) = \begin{cases} j, & j = 1, \dots, k^{(i)} & \text{w.p. } q_j, \\ k^{(i)} + 1, & & \text{w.p. } q_{k^{(i)}+1}, \end{cases} \quad (12)$$

with multinomial probabilities q_j given as follows. If $n_{s_i}^{(i)} > 0$

$$q_j \propto \begin{cases} n_j^{(i)} p(y|\theta^{(i)}, \theta_i = \theta_j^*, V), & \text{for } j = 1, \dots, k^{(i)}, \\ \alpha / (k^{(i)} + 1) p(y|\theta^{(i)}, \theta_i = \theta_{k^{(i)}+1}^*, V), & \text{for } j = k^{(i)} + 1. \end{cases} \quad (13)$$

If $n_{s_i}^{(i)} = 0$, then with probability $1/k^{(i)}$ use the multinomial probabilities q_j , otherwise $s_i = k$. Notice that the integral expression for q_0 is replaced by a sum of simple $(p + q)$ -dimensional normal density evaluations required for (13).

- (ii) Sampling $r_t \sim p(r_t|\theta^*, \mathcal{S}, V)$, $t = 1, \dots, T$. Generating the latent variables r_t is straightforward multinomial sampling with probabilities $Pr(r_t = i) \propto \phi(x_t; \mu_i, V) N(y_t; \beta_i' x_t, w_i)$.

- (iii) Sampling (β_j^*, w_j^*) , $j = 1, \dots, k$ from $p(\beta_j^*, w_j^* | \mathcal{S}, r, \nu, y)$. Note that the conditioning set includes the indicators introduced in (8). The indicators r_t allow us to associate each θ_j^* with a set of indices $\Gamma_j = \{t : s_{r_t} = j\}$. Conditional on $r = (r_1, \dots, r_T)$, the posterior on (β_j^*, w_j^*) is a simple autoregression of y_t on x_t , $t \in \Gamma_j$, giving an normal/inverse gamma conditional posterior for (β_j^*, w_j^*) .
- (iv) Updating μ_j^* , $j = 1, \dots, k$. Implementation of a straightforward Gibbs sampling step for μ_j^* is precluded by the complicated form of $p(\mu_j^* | \mu_1^*, \dots, \mu_{j-1}^*, \mu_{j+1}^*, \dots, \mu_k^*, \beta^*, w^*, \mathcal{S}, V, y)$ which does not allow efficient random variate generation. Instead we use a Metropolis step (Tierney 1994) to update μ_j^* . First we generate a “candidate” $\tilde{\mu}_j^* \sim N(\tilde{\mu}_j^*; \mu_j^*, \Sigma)$ and compute

$$a(\mu_j^*, \tilde{\mu}_j^*) = \min \left(1, \frac{p(y | \tilde{\theta}, V) p(\tilde{\mu}_j^*)}{p(y | \theta, V) p(\mu_j^*)} \right),$$

where $\tilde{\theta}$ denotes the currently imputed θ vector with μ_j^* replaced by $\tilde{\mu}_j^*$, and $p(\mu_j^*)$ is the normal $N(\mu_j^*; m_\mu, B_\mu)$ hyperprior in (9). Also, $p(y | \theta, V) = \prod_{t=1}^T p(y_t | x_t, \theta, V)$ denotes the likelihood. In the current implementation we used the initial value of V for the covariance matrix Σ in the normal proposal distribution. Second, with probability $a(\mu_j, \tilde{\mu}_j)$ we replace μ_j by $\tilde{\mu}_j$; otherwise we leave μ_j unchanged.

- (v) Updating V . Again, the complete conditional posterior does not allow efficient random variate generation to implement a Gibbs sampling step. Instead we realize an independence chain step (Tierney 1994). Generate a candidate $\tilde{V}^{-1} \sim g(V^{-1}) = W(V^{-1}; q, (qR)^{-1})$ and compute

$$a(V, \tilde{V}) = \min \left(1, \frac{p(\tilde{V} | \theta^*, \mathcal{S}, y) g(V^{-1})}{p(V | \theta^*, \mathcal{S}, y) g(\tilde{V}^{-1})} \right) = \min \left(1, \frac{p(y | \theta, \tilde{V})}{p(y | \theta, V)} \right).$$

With probability $a(S, \tilde{V})$ replace S by \tilde{S} , otherwise keep S unchanged. The choice of the candidate generating distribution $g(V^{-1})$ is motivated by the simple form of the acceptance probabilities $a(V, \tilde{V})$.

- (vi)–(viii) Updating α , μ_β and B_β . The conditionals to resample μ_β and B_β are straightforward multivariate normal and Wishart distributions.

Resampling α is done by introducing a latent beta distributed variable as described in Escobar and West (1995), based on West (1992).

- (ix) Resampling $\theta_j^* \sim p(\theta_j^*|\nu, k)$, $j = k + 1, \dots, n$ is straightforward. Note that θ_j^* , $j > k$ is independent of y and all other parameters given k . Therefore the conditional posterior $p(\beta_j^*, w_j^*, \mu_j^*|\nu, k)$ is identical to the normal/inverse gamma/normal prior.

3.2 Convergence

In this section we discuss convergence issues for the Markov chain Monte Carlo scheme defined earlier. For a more detailed discussion of parameterization and convergence issues in mixture of Dirichlet process models like (9) we refer to MacEachern and Müller (1994). Here we only summarize the main result.

Let $P^n(\omega, \cdot)$ denote the transition probability defined by n iterations of the Gibbs sampler if the current state of the chain is ω , and write π for the posterior distribution. Following Tierney (1994) a sufficient condition for convergence is that for each subset A of the parameter space Θ with $\pi(A) > 0$, and each parameter vector $\omega \in \Theta$ there exists an integer $n = n(\omega, A) \geq 1$ such that $P^n(\omega, A) > 0$ (π -irreducibility).

The configuration vector \mathcal{S} introduces a finite partition of the parameter space into subspaces Θ_s with equal configurations. There exists a configuration s with $\pi_s(A) := \pi(A \cap \Theta_s) > 0$. Since at each iteration the Gibbs sampler allows a move to any other configuration with positive probability, we can, for any initial ω , move to any any configuration s with positive probability in one iteration. In sub-steps (ii)–(ix) we only generate from distributions that are mutually absolutely continuous with respect to π_s . These arguments suffice to show that $n(\phi, A) = 1$.

To actually decide termination of the simulated Markov chain we use a diagnostic proposed by Geweke (1992) together with informal graphical methods. The diagnostic indicated practical convergence after 5000 and 10000 iterations for the examples reported in Sections 4.2. and 4.3., respectively. Actual CPU time on a DEC alpha 3000 was 139 and 132 minutes,

respectively.

4 EXAMPLES

4.1 A Simulation Experiment

For a simulated example we generated data from three known models: (i) An AR(1) model with an exogenous covariate: $y_t = a_0 + a_1 y_{t-1} + a_2 z_t + \epsilon_t$ with $a = (1.0, 0.9, 0.25)$; (ii) a threshold autoregression TAR(1) with

$$y_{t+1} = \begin{cases} 1.0 + 0.6y_t + \epsilon_t, & \text{if } y_t \leq 0, \\ -1.0 + 0.4y_t + \epsilon_t, & \text{if } y_t > 0; \end{cases}$$

(iii) an AR(2) model: $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \epsilon_t$ with $a = (0, 0.2, 0.1)$. In all three examples we assumed $\epsilon_t \sim N(0, 0.2^2)$ and simulated $T = 190$ observations y_t , $t = 1, \dots, 190$. Additional data points y_{191}, \dots, y_{200} which were not used in fitting the models were generated to compute predictive mean squared error.

In simulation example (i) we estimated the AR coefficients a_j in the true AR(1) model and the autoregressive mixture model (7). In example (ii) we estimated the coefficients a_j in the true TAR(1) model using the true threshold, an AR(1) model without threshold, and model (7) with $n = 10$, $n = 5$ and $n = 2$. For example (iii) we fitted the true AR(2) model and model (7) with $n = 5$. For each example we simulated $M = 100$ experiments. In each simulated experiment the AR and the threshold AR models were estimated by maximum likelihood, and the non-linear AR model was estimated by posterior simulation as described in Section 3.1. Table 1 reports the mean squared errors over the M simulations for each of the fitted models. Let $\hat{y}_{t,m}$ denote the fitted value for the t -th observation in the m -th experiment, and let $y_{t,m}$ denote the simulated observation at time t in the m -th experiment. The third column reports the mean squared error

$$\text{MSE}(\text{fit}) = \sum_{t=t_0}^{190} \left(\frac{1}{M} \sum_{m=1}^M (y_{t,m} - \hat{y}_{t,m})^2 \right).$$

Here $t_0 = 2$ for Examples (i) and (ii), and $t_0 = 3$ for Example (iii). The last three columns report the mean squared prediction error for $j = 1, 5$ and 10 step ahead forecasts:

$$\text{MSE}(\text{forec}) = \frac{1}{M} \sum_{m=1}^M (y_{T+j,m} - \hat{y}_{T+j,m})^2.$$

Under the true model $\text{MSE}(\text{fit}) = (T - t_0) \cdot 0.04$, and $\text{MSE}(\text{forec}) = 0.04$. Smaller values indicate overfit, larger values indicate lack of fit. For comparison the footnote reports the marginal variance $V(y_t)$. All MSE's are conditional on stationary, i.e., the fitted values $\hat{y}_{t,m}$ are posterior means conditional on stationarity. We compute these from the Markov chain Monte Carlo simulation described in Section 3.1 by dropping posterior simulations corresponding to non-stationary solutions from the MCMC averages.

Table 1 suggests that there is little loss in efficiency for using the mixture model when the data is actually generated by a linear AR model (Examples i and iii), but potential gains when the true model violates the linear AR assumption (Example ii). Also, in Example (ii) the loss in efficiency when using the mixture model instead of the threshold AR is rather small, even if the true threshold is used in fitting the threshold AR. The role of n , the maximum number of terms in the mixture of linear AR's, is not critical as long as n is large enough to explain the non-linearity ($n = 10, 5, 2$ in Example ii).

From the simulations in Example (ii), some guidelines emerge for the choice of n . The parameter n is the maximum number of distinct linear submodels in the locally weighted mixture (7). Thus the choice of n is related to the expected non-linearity of the autoregressive function. For example, $n = 2$ allows for one change of regimen, $n = 3$ allows for S-shaped functions etc. As a general guideline we suggest that one choose larger values of n if in doubt. The simulations reported in Table 1 give evidence that too large a value for n (for example $n = 10$ in Example ii) is not as critical as too small a choice for n (for example, $n = 1$, i.e., the linear AR(1) in Example ii).

4.2 An AR(1) model with covariate

Azzalini and Bowman (1990) analyzed a data set concerning eruptions of the Old Faithful geyser in Yellowstone National Park in Wyoming. The data set records eruption durations and intervals between subsequent eruptions, collected continuously from August 1st until August 15th, 1985. Of the original 299 observations we removed 78 observations which were taken at night and only recorded durations as “short”, “medium” or “long”. Figure 1 plots the data.

We fit the non-linear AR(1) model (7) to the times between eruptions (y_t), using the duration of the previous eruption as an additional covariate (z_t). The likelihood (8) is specified as:

$$\begin{aligned} (y_t|y_{t-1}, z_t, \theta, V, r_t = i) &= \beta_{0i} + \beta_{1i}y_{t-1} + \beta_{2i}z_t + \epsilon_t, \quad \epsilon_t \sim N(0, w_i), \\ Pr(r_t = i|\theta, V, y_{t-1}, z_t) &\propto \phi(x_t; \mu_i, V), \end{aligned}$$

where, now, $x_t = (y_{t-1}, z_t)$.

Figures 2 through 4 show some elements of the estimated non-linear autoregression. Model (7) allows non-linearity as well as non-normality. Both are crucial for this data set. Figure 3a illustrates how the conditional modal trace can sometimes provide a better summary of important features of the conditional distribution $p(y_{t+1}|y_t, z_{t+1})$ than conditional expectations. Notice the lack of normality in Figure 3. For the lagged variable y_{t-1} in the range 70–90 the data suggest a bimodal conditional for $p(y_t|y_{t-1}, z_t)$ exemplified in Figure 3b. Figure 5 plots one-step ahead forecasts. See section 4.4 for details of implementation and prior choice.

4.3 A Harmonic Process Model

In this example we use the non-linear autoregression (7) to estimate a generalization of harmonic process models following West (1995), who estimates the unknown frequencies in harmonic models by fitting corresponding unit root AR(2) models. Let $\epsilon_t \sim N(0, \sigma^2)$. The AR(2) model

$$p(y_t|y_{t-1}, y_{t-2}, \beta) = N(y_t; \beta y_{t-1} - y_{t-2}, w) \tag{14}$$

defines an harmonic model with one fixed frequency $\arccos(\beta/2)$ and time varying amplitude and phase, subject to $|\beta| < 2$.

A generalization of the harmonic process model allows frequencies to be selected conditional on covariates. Clearly this can be formalized by modifying (14) to allow a locally weighted mixture of auto-regressions in the form of (7). Specifically, the likelihood (8) takes the form:

$$\begin{aligned} p(y_t|y_{t-1}, y_{t-2}, \theta, V, r_t = j) &= N(y_t; \beta_j y_{t-1} - y_{t-2}, w_j), \\ Pr(r_t = j|\theta, V, y_{t-1}, y_{t-2}) &\propto \phi(x_t; \mu_j, V), \end{aligned}$$

where, now, $x_t = (y_{t-1}, y_{t-2})$.

We illustrate the model by estimating data giving the annual number of lynx trappings in the Mackenzie River District of North-West Canada for the period 1821 to 1934 (Priestley 1988). Figure 6a shows the time series, on a log scale and detrended. Figure 6b plots histograms of imputed periods (i.e. $2\pi/\arccos(\beta_j/2)$) for all posterior samples of regression coefficients β_j . Priestley (1988) observes an asymmetry in the behaviour of the series. The time spent on the rising side of a period (i.e., rising from “trough” to “peak”) seems slightly longer than the time spent on the falling side (i.e. from “peak” to “trough”). The histograms in Figure 6b are separated for points y_t on the rising side of a period ($y_{t-2} < y_{t-1}$) and points on the falling side ($y_{t-2} > y_{t-1}$), confirming this observation of asymmetry.

Using Markov chain Monte Carlo simulation makes it easy to consider additional model elaborations by simply adding appropriate layers to the simulation scheme. For example, a measurement error model, assuming the autoregressive process on the lynx population rather than the trappings, would be a straightforward extension of model (15), requiring but an additional step in the simulation scheme to impute the latent variables representing the true population numbers.

4.4 Implementation and prior choice

In the examples we used the following choices for the hyperparameters in the model. The hyperprior on m_β is chosen noninformative with $A^{-1} = 0$; the

matrices R , C and B_μ are diagonal matrices simply reflecting the scale of the problem. All variables $(y_t, z_t, t = 1, \dots, T)$ are centered, therefore $m_\mu = 0$. The remaining hyperparameters in the example reported in Section 4.2. are chosen as $q = c = 10$, $a_w = 10$, $b_w = 100$, $a_0 = 5$, $b_0 = 1$. And $q = 5$, $c = 5$, $a_w = 10$, $b_w = 5$, $a_\alpha = 5$, $b_\alpha = 1$. In the example of Section 4.3. the β -vector is one-dimensional only. Hence we replaced the Wishart hyperprior on B_β^{-1} by a gamma distribution $Ga(B_\beta^{-1}; c, cC)$, i.e. $c = 5$ is a gamma shape parameter. Also, in both examples we used $n = 10$ for the maximum number of terms in the mixture.

We initialized the Gibbs samplers by $k = 10$, $\mu_i = 0$ and $\beta_i = m_\beta = a$, $i = 1, \dots, 10$, with $a = (-4, -0.5, 0)$ and $a = 2.0$ in the examples reported in Section 4.2. and 4.3., respectively. These initial values were computed by least squares fits of the regression $y_t = a'x_t + \epsilon_t$. The total mass parameter α is initialized at its hyperprior mean $\alpha = 5$.

Interest in our analysis focuses on prediction and estimation of the full conditional distribution $p(y_{t+1}|y_t, \dots)$. Empirical evidence suggests that this inference is only little sensitive to the particular hyperprior parameter choice within a wide range. Also, larger values for n , the maximum number of terms in the weighted mixture, tends to only marginally increase the number k of a posteriori estimated distinct sub-models.

REFERENCES

- ALBERT, J. and CHIB, S. (1993) Bayes inference via Gibbs sampling of autoregressive time series subject to markov mean and variance shifts. *Journal of Business and Economic Statistics* 11, pp. 1–15.
- ANTONIAK, C.E. (1974) Mixtures of Dirichlet processes with applications to non-parametric problems. *Annals of Statistics* 2, pp. 1152-1174.
- AZZALINI, A. and BOWMAN, A. W. (1990) A look at some data on the Old Faithful geyser. *Applied Statistics* 39, pp. 357-365.
- BUSH, C.A. and MACEACHERN, S.N. (1996) A semi-parametric Bayesian model for randomised block designs. *Biometrika* 83, 275-285.
- CARLIN, B.P, POLSON, N.G., and STOFFER, D.S. (1992) A Monte Carlo

- approach to non-normal and non-linear state-space modelling. *Journal of the American Statistical Association* 87, pp. 493-500.
- CARTER, C.K. and KOHN, R. (1994) Bayesian methods for conditionally Gaussian state space models. Technical Report, Australian Graduate School of Management, UNSW.
- ESCOBAR, M.D. and WEST, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90, 577-588.
- GELFAND, A.E. and SMITH A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, pp. 398-409.
- GEWEKE, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, ed. J.O. Berger, J.M. Bernardo, A.P. Dawid and Smith, A.F.M., pp. 169-194.
- GEWEKE, J. and TERUI, N. (1993) Bayesian threshold auto-regressive models for nonlinear time series. *Journal of Time Series Analysis* 14 (5), p. 441.
- GILKS, W.R., CLAYTON, D.G., SPIEGELHALTER, D.J., BEST, N.G., MCNEIL, A.J., SHARPLES, L.D., and KIRBY, A.J. (1993) Modelling complexity: applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society Ser. B*, 55, pp. 39-52.
- JACQUIER, E., POLSON, N.G., and ROSSI, P.E. (1994) Bayesian analysis of stochastic volatility models. *Journal of Business and Economics Statistics*, 12, pp. 371-389.
- MACEachern, S.N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* 23 (3), pp. 727-741.
- MACEachern, S.N. and MUELLER, P. (1994) Estimating mixture of Dirichlet process models. Discussion Paper 94-A11, ISDS, Duke University.
- MUELLER, P., ERKANLI, A., and WEST, M. (1996) Bayesian curve fitting

- using multivariate normal mixtures. *Biometrika*, 83, pp. 67-79.
- PRIESTLEY, M.B. (1988) *Non-linear and Non-stationary Time Series Analysis* Academic Press, London.
- SMITH, A.F.M. and ROBERTS, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 55, pp. 3-23.
- TIERNEY, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics*, 22, pp. 1701-1728.
- TONG, H. (1990) *Nonlinear time series*, Clarendon Press, Oxford.
- WEST, M. (1992) Hyperparameter estimation in Dirichlet process mixture models. Discussion paper 92-A03, ISDS, Duke University.
- WEST, M. and CAO, G. (1993) Assessing mechanism of neural synaptic activity. In *Bayesian Statistics in Science and Technology: Case Studies*, ed. Gatsonis, C., Hodges, J, Kass, R. and Singpurwalla, N., New-York.
- WEST, M., MUELLER, P., and ESCOBAR, M.D. (1994) Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A tribute to D. V. Lindley*, eds. A.F.M. Smith and P. Freeman, Wiley, New-York.
- WEST, M. and TURNER, D.A. (1994) Deconvolution of mixtures in analysis of neural synaptic transmission. *The Statistician* 43, pp. 31-43.
- WEST, M. (1995) Bayesian inference in cyclical component dynamic linear models. *Journal of the American Statistical Association*, 90, pp. 1301-1312.

Table 1: Mean squared errors in the simulation experiments (i)-(iii).

	fitted model	MSE(fit)	MSE(forec) $T + 1$	MSE(forec) $T + 5$	MSE(forec) $T + 10$
(i)	AR(1)*	7.46 (0.08)	0.040 (0.004)	0.053 (0.006)	0.059 (0.01)
	mixture	7.90 (0.08)	0.041 (0.006)	0.041 (0.006)	0.067 (0.008)
(ii)	TAR(1)*	7.71 (0.08)	0.043 (0.005)	0.061 (0.011)	0.11 (0.03)
	AR(1)	28.24 (.030)	0.15 (0.02)	0.24 (0.02)	0.41 (0.03)
	mixture n=10	11.80 (0.17)	0.051 (0.013)	0.090 (0.013)	0.23 (0.05)
	mixture n=5	10.90 (0.16)	0.052 (0.013)	0.15 (0.04)	0.20 (0.04)
	mixture n=2	9.80 (0.13)	0.045 (0.009)	0.15 (0.04)	0.19 (0.04)
(iii)	AR(2)*	7.33 (0.08)	0.040 (0.005)	0.053 (0.007)	0.070 (0.009)
	mixture	7.59 (0.08)	0.040 (0.005)	0.088 (0.013)	0.10 (0.01)

NOTE: The true sampling models are marked with *. The theoretically optimal values for MSE(fit) are 7.56 ($= 189 \cdot 0.04$), 7.56 and 7.52 ($= 188 \cdot 0.04$) for Examples (i), (ii), and (iii), respectively. The optimal value for MSE(forec) is 0.04. Larger values indicate lack of fit, smaller values indicate overfit. The values in parentheses are numerical standard deviations (i.e., accuracy of the reported value). For comparison, the marginal variances $V(y_t)$ are 0.86, 0.49 and 0.057 in Examples (i), (ii) and (iii), respectively.

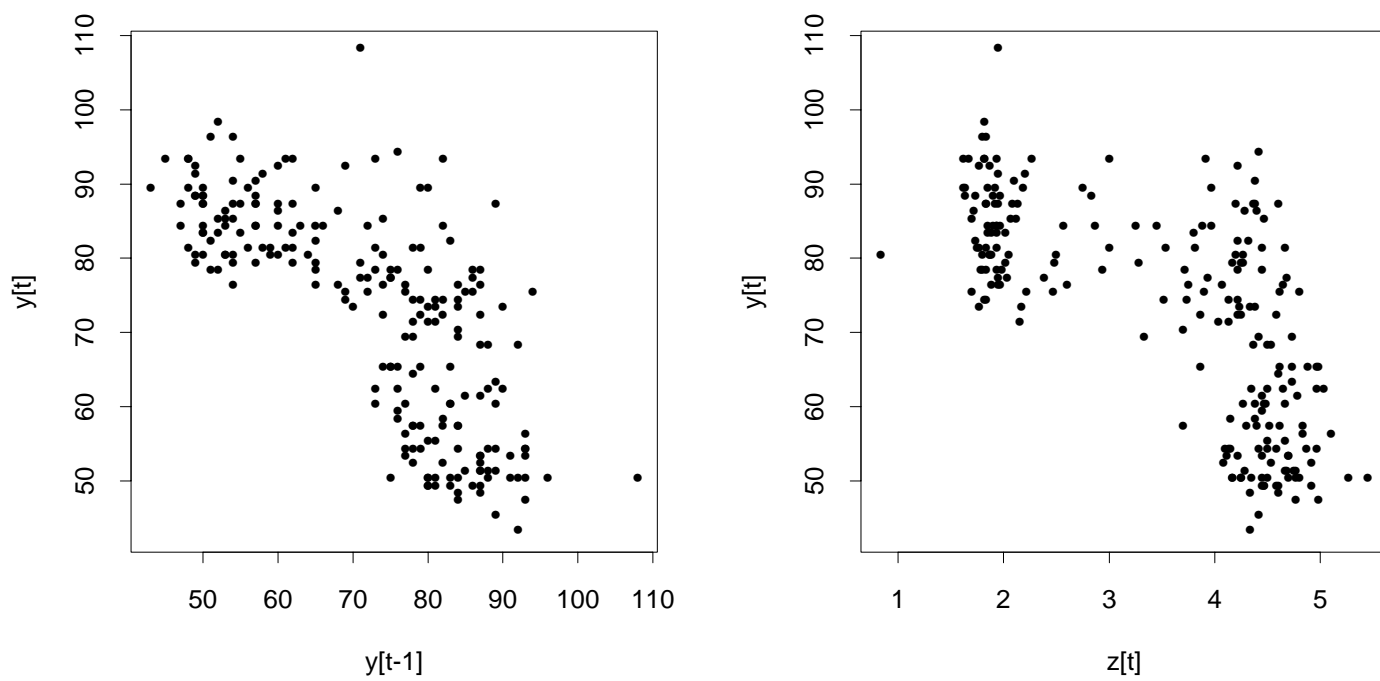
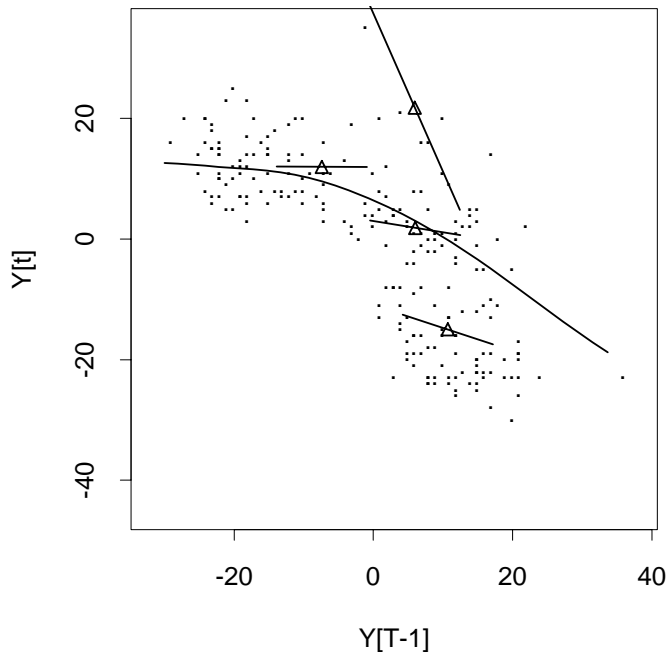
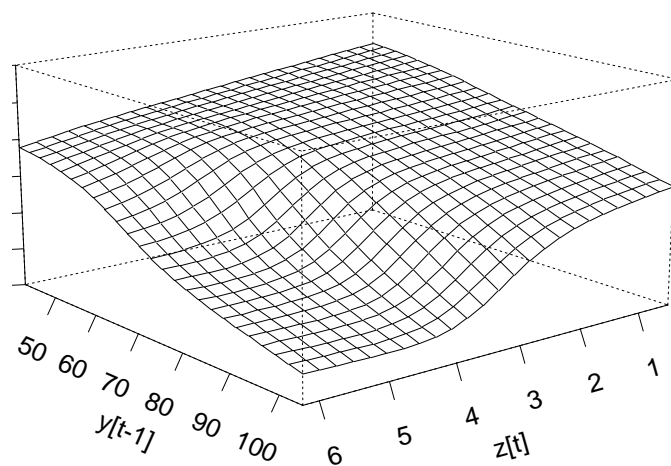


Figure 1: Old Faithful geyser data set: Waiting time until next eruption (y_t) versus lagged y_{t-1} and y_t versus duration of last eruption (z_t).

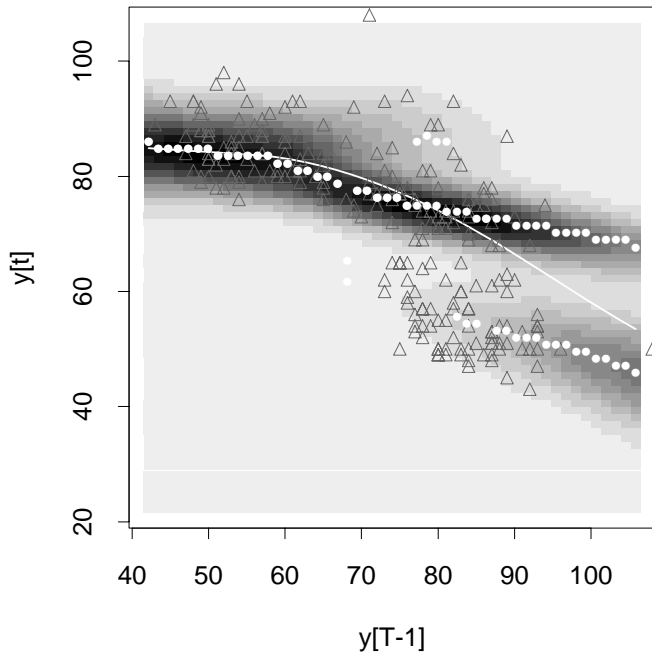


(a) $E(y_{t+1}|y_t, z = \bar{z}, y)$

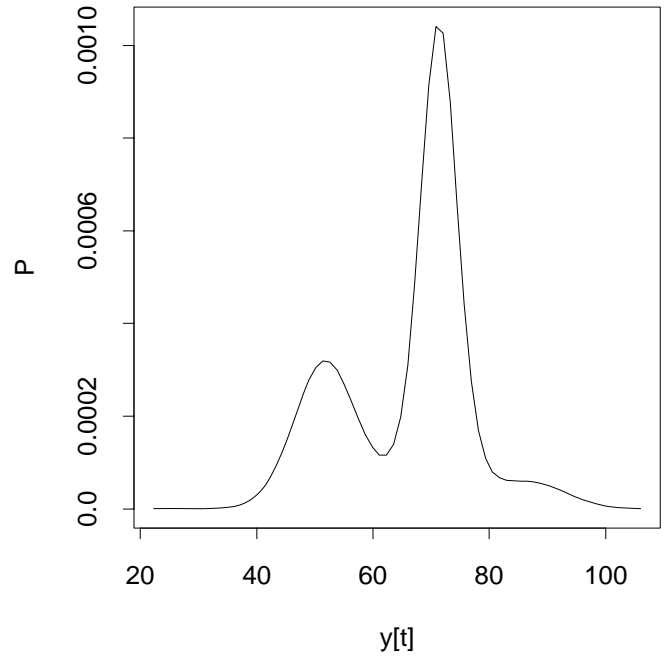


(b) $E(y_{t+1}|y_t, z_{t+1}, y)$

Figure 2: The estimated non-linear autoregression function. Panel (a) shows $E(y_{t+1}|y_t, z = \bar{z}, y)$ (solid curve). For comparison the dots plot the observed data. The short line segments show some elements of the $k = 6$ distinct regression lines which are imputed at iteration 10,000. The lines have intercept $\beta_{j_0}^*$ and slope $\beta_{j_1}^*$. The center of each of the segments corresponds to the $\mu_{j_2}^*$. Panel (b) shows the three-dimensional AR(2) regression surface as estimated by $E(y_{t+1}|y_t, z_t, y)$.



(a) $p(y_{t+1}|y_t, z_{t+1} = \bar{z})$



(b) $p(y_{t+1}|y_t = 90, z_{t+1} = \bar{z})$

Figure 3: In addition to the point estimate for the unknown autoregression shown in figure 2a the model allows inference about the uncertainty for this line. Panel (a) shows $p(y_{t+1}|y_t, z_{t+1} = \bar{z}, y)$ as a grey-shade contour plot. Note that this does not show a joint bivariate distribution, but only shows conditional distributions. The autoregressive relationship is summarized by the conditional expectations (solid white line) or, alternatively, by the conditional modes (white stars). For comparison the triangles show the actual data points. Panel (b) shows $p(y_{t+1}|y_t = 90, z_{t+1} = \bar{z}, y)$, evidently displaying clear non-normality.

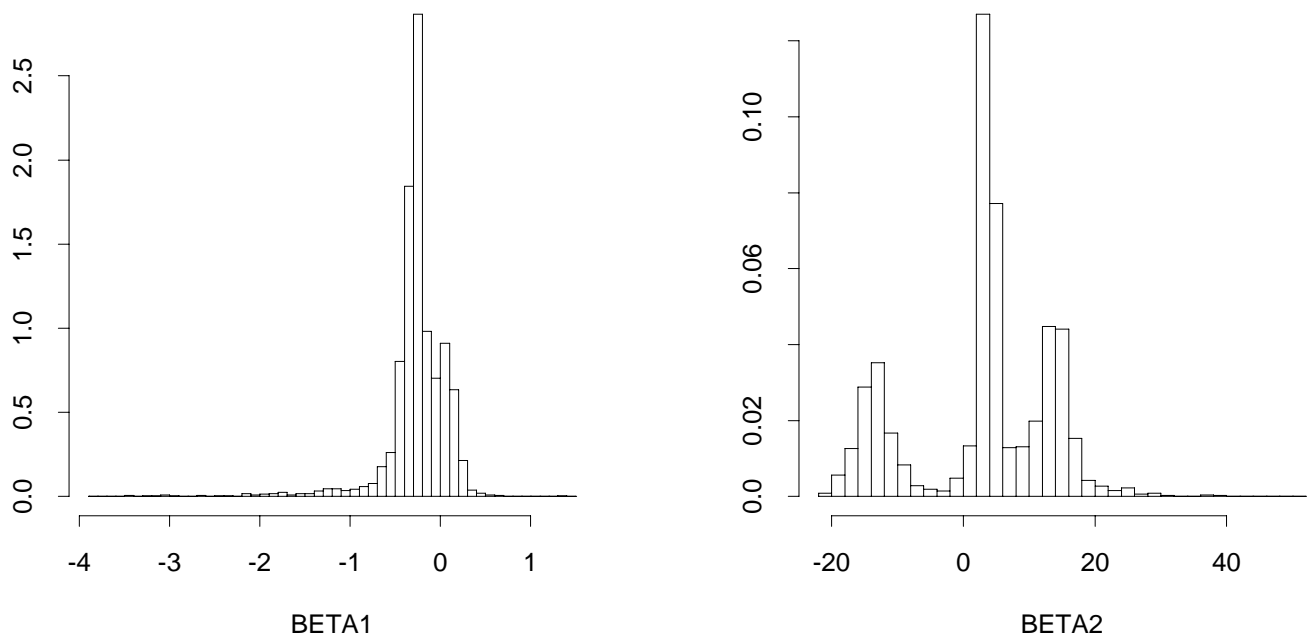


Figure 4: The posterior distributions $p(\beta_1|y)$ and $p(\beta_2|y)$ in the Old Faithful geyser data. Note the multimodality of $p(\beta_2|y)$ reflecting the clustering in figure 1b.

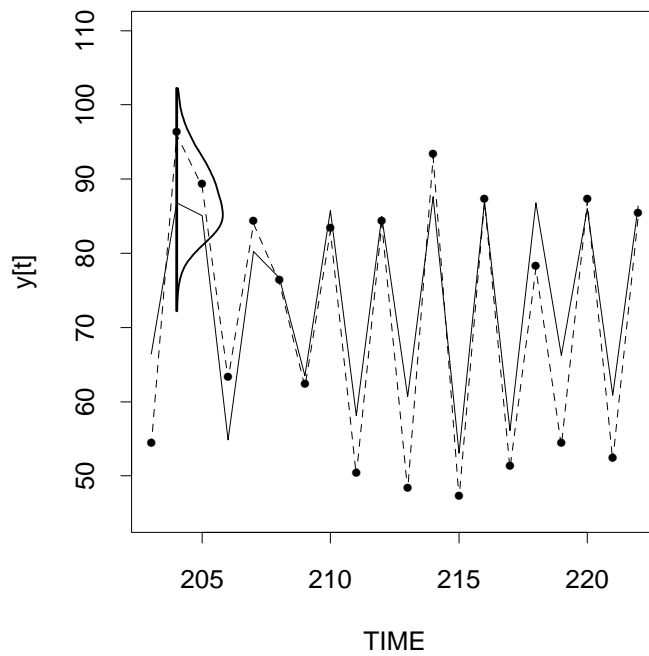


Figure 5: One step ahead forecasts $\hat{y}_t = E(y_t|y_1, \dots, y_{t-1}, z_t, y)$ (solid line). The actual time series (dashed line and dots) is shown for comparison. The noticeably non-bell shaped p.d.f. overlaid at $t = 204$ shows $p(y_{204}|y_1, \dots, y_{203}, z_{204}, y)$.

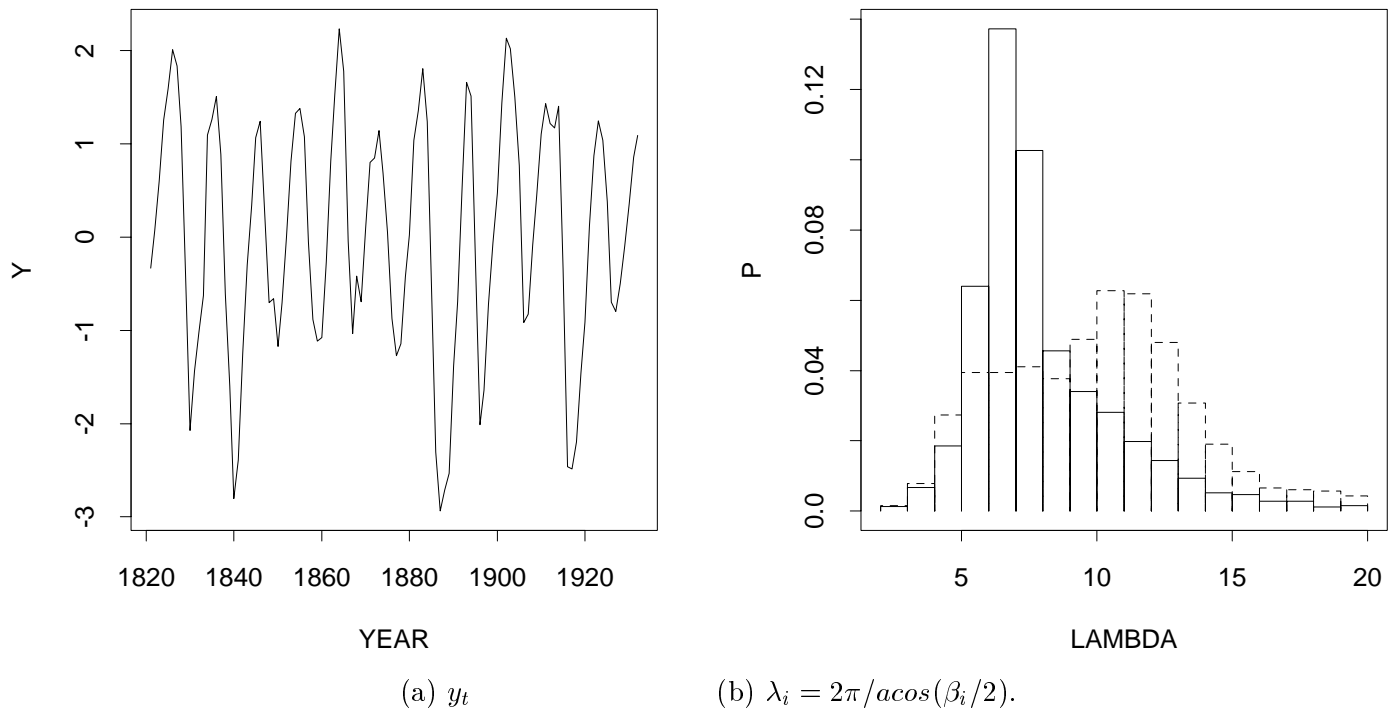


Figure 6: Annual number of lynx trappings in the Mackenzie River District of North-West Canada for the period 1821 to 1934. Panel (a) shows the time series (logarithms and detrended). Panel (b) shows histograms of the imputed periods $\lambda_t = 2\pi/\text{acos}(\beta_{r_t}/2)$. The histogram is separated for points on the falling side (i.e. $y_{t-2} > y_{t-1}$) (solid line) and points on the rising side (i.e. $y_{t-2} < y_{t-1}$) (dashed line).

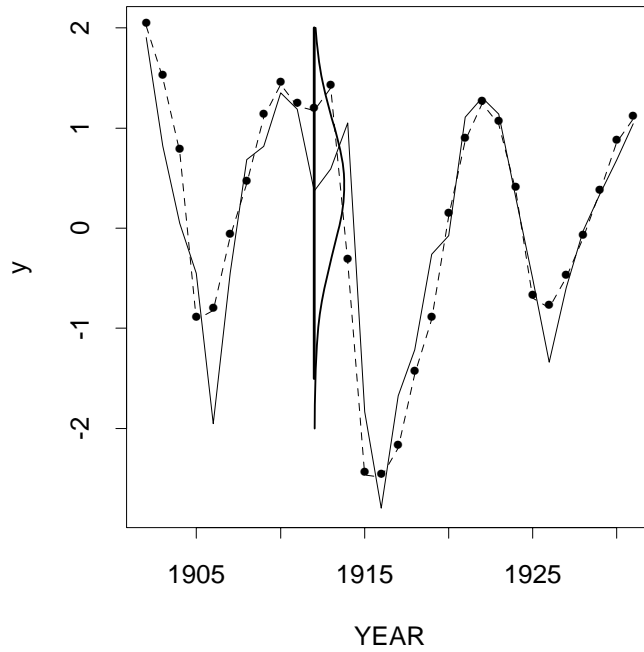


Figure 7: One step ahead forecasts $\hat{y}_t = E(y_t|y_1, \dots, y_{t-1})$ (solid line). The actual time series (dashed line and dots) is shown for comparison. Besides point forecasts \hat{y}_t , the model also delivers a full description of the uncertainty of y_t . This is exemplified in the figure by plotting the predictive distribution $p(y_{1912}|y_{1820}, \dots, y_{1911})$ (bell shaped curve overlayed at $t = 1912$).

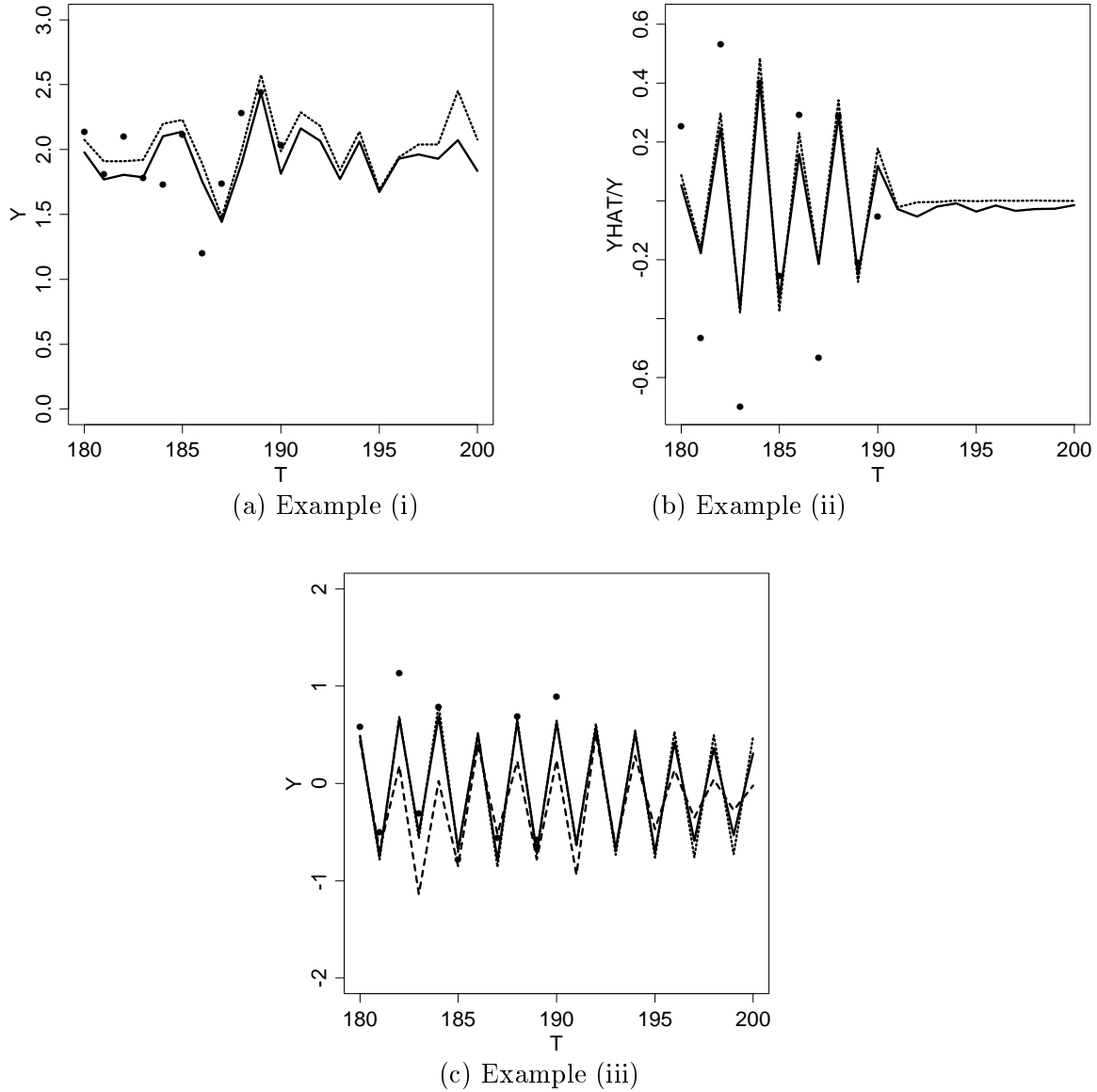


Figure 8: Simulation example: Forecasts under the true model (solid line), under the mixture model (dashed line), and under an AR(1) model (panel (c) only; dotted line). Forecasts are for one specific simulation in each of the three examples. For $t = 180, \dots, 190$ the curves show the fitted values under the respective model. For example, the solid line in Example (i) plots $E(\beta_0 + \beta_1 y_{t-1} + \beta_2 z_t | y_1, \dots, y_{190})$ against $t = 180, \dots, 190$. Here $(\beta_0, \beta_1, \beta_2)$ denote the coefficients in the AR(1) model with covariate z_t . For $t = 191, \dots, 200$ the curves show the posterior predictive mean under the respective model. For example, the solid line in Example (i) plots $E(\beta_0 + \beta_1 y_{t-1} + \beta_2 z_t | y_1, \dots, y_{190})$ against $t = 191, \dots, 200$. The dots plot for comparison the simulated data.