Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 4424–4437

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A consistent nonparametric Bayesian procedure for estimating autoregressive conditional densities

Yongqiang Tang[a,*], Subhashis Ghosal[b,1]

[a]Department of Psychiatry, SUNY Health Sciences Center, 450 Clarkson Avenue, Box 1203, Brooklyn, NY 11203, USA
[b]Department of Statistics, North Carolina State University, 220 Patterson Hall, 2501 Founders Drive, Raleigh, NC 27695-8203, USA

Received 24 October 2005; received in revised form 26 June 2006; accepted 29 June 2006
Available online 24 July 2006

Abstract

This article proposes a Bayesian infinite mixture model for the estimation of the conditional density of an ergodic time series. A nonparametric prior on the conditional density is described through the Dirichlet process. In the mixture model, a kernel is used leading to a dynamic nonlinear autoregressive model. This model can approximate any linear autoregressive model arbitrarily closely while imposing no constraint on parameters to ensure stationarity. We establish sufficient conditions for posterior consistency in two different topologies. The proposed method is compared with the mixture of autoregressive model [Wong and Li, 2000. On a mixture autoregressive model. J. Roy. Statist. Soc. Ser. B 62(1), 91–115] and the double-kernel local linear approach [Fan et al., 1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika 83, 189–206] by simulations and real examples. Our method shows excellent performances in these studies.

Keywords: Dirichlet process mixture models; Posterior consistency; No-gaps algorithm

---

1. Introduction

Conditional densities characterize the probabilistic aspects of a time series with Markovian structures. They determine, except for the initial distribution, the likelihood function of an observed time series (Tong, 1990). Estimating the conditional density of a time series is very useful to understand the phenomena underlying the data and to make future predictions. Let $\boldsymbol{X} = \{X_n : n \geqslant 1 - k\}$ be an ergodic time series with conditional density $f(X_i|Z_i)$, where $Z_i = (X_{i-1}, \ldots, X_{i-k})^{\mathrm{T}}$. We are interested in the problem of estimating f using a nonparametric approach.

Traditional nonparametric techniques such as kernel methods (Rosenblatt, 1969; Fan et al., 1996; Hyndman and Yao, 2002), spline smoothing (Gu and Wang, 2003) and wavelet (Clemencon, 2000) have been widely used for conditional density estimation. These methods have performed successfully for low-dimensional problems (i.e., $k \leqslant 3$, where k is the dimension of Z_i), however few have been applied to higher-dimensional data ($k > 3$). For high-dimensional data, the kernel method suffers from the sparse data problem (Hastie et al., 2001) while density estimation via spline

---

[*] Corresponding author. Tel.: +1 7182702231; fax: +1 718 2704081.
E-mail addresses: yongqiang_tang@yahoo.com (Y. Tang), ghosal@stat.ncsu.edu (S. Ghosal).
[1] Research of the second author is partially supported by NSF Grant number DMS-0349111.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.020

smoothing usually involves high-dimensional numerical integrations (Jeon and Lin, 2006). In addition, the estimator may be negative and their integrals with respect to x may not equal 1 (Hyndman and Yao, 2002).

In this paper, we consider a Bayesian nonparametric approach for estimating the autoregressive (AR) conditional density. The Bayesian nonparametric modelling involves specifying priors on an infinite-dimensional functional space. Many such priors have been developed since the pioneering work of Ferguson (1973) on the Dirichlet process (DP) prior; see Choudhuri et al. (2005) for a comprehensive review. In our approach, we induce a prior on the conditional density via a DP mixture model (DPMM), that is, the conditional density is modeled as a nonparametric mixture of normal kernels where the mixing distribution function follows a DP. Bayesian nonparametric approaches via DP have become increasingly popular as efficient Markov chain Monte Carlo (MCMC) schemes were developed (Escobar and West, 1995; MacEachern and Müller, 1998). The DPMMs have been widely applied to density estimation (Escobar and West, 1995; Müller et al., 1996). Consistency and rates of convergence issues for Dirichlet mixture approaches to density estimation have been studied by Ghosal et al. (1999, 2000) and Ghosal and van der Vaart (2001). Our approach can be considered as a natural extension of the well studied Dirichlet mixture approach to density estimation. Müller et al. (1997) proposed a finite locally weighted mixtures of AR models via the DP, where the parameters of the mixture components are assumed to be i.i.d variables from some random distribution drawn from a DP prior. Our model differs from that of Müller et al. (1997) in that the number of mixture components in DPMM is unlimited. Furthermore, ergodicity of the AR process is explicitly enforced in our model. We have studied posterior consistency issues for a wide class of order-one Markov process in Tang and Ghosal (2006) and Ghosal and Tang (2005). We will extend the consistency results to ergodic time series.

The Dirichlet process prior allows flexible nonparametric mixture modeling. It is more flexible than the finite mixture AR models introduced by Wong and Li (2000). One main difficulty with finite mixture models is the determination of the number of mixture components. However in a DPMM, the number of mixture components is not specified in advance and can grow when new observations come in. In addition, with no or few assumptions, the DPMM allows the data to drive the shape of the error distribution and thus provides more reliable inference. Similar to the kernel-based approaches, the estimates from the DPMM have local properties in the sense that they are heavily determined by points (X_i, Z_i) that are close in the sample space; see Müller et al. (1996) and the references therein. However, a Bayesian model-assisted approach, enabling exact inference based on the posterior distribution, seems promising to overcome the difficulties with the classical nonparametric approaches discussed above. In this paper, these methods are compared numerically.

The paper is organized as follows. Section 2 introduces the DPMM. Section 3 suggests an appropriate MCMC scheme for posterior inference. Sufficient conditions for posterior consistency are established in two different topologies described in Section 4. Section 5 illustrates the performance of the DPMM and compares it with mixture of autoregressive (MAR) models (Wong and Li, 2000) and the double-kernel local linear approach (Fan et al., 1996) through simulations. Section 6 applies DPMM to the analysis of Canadian lynx data.

2. Dirichlet process mixture model

The DPMM introduced in this section provides a Bayesian nonparametric framework for modeling ergodic time series via the DP. Section 7 discusses other Dirichlet mixture models using different AR link functions. Let $\boldsymbol{X}^{(n)} = \{X_i : i \leqslant n\}$ denote the observed time series. We assume that $Z_1 = (X_{1-k}, \ldots, X_0)^{\mathrm{T}}$ is fixed or has a known initial distribution. We model the conditional density as a nonparametric mixture of normal kernels,

$$f\left(X_i|Z_i\right) = \int \phi\left(X_i; g\left(u, \gamma, \beta, Z_i\right), \sigma\right) \mathrm{d}P\left(u, \gamma, \sigma\right), \tag{1}$$

where $g(u, \gamma, \beta, Z) = u + \beta \tanh(\gamma^{\mathrm{T}} (Z - u\mathbf{1}_k))$ is the AR link function, $\gamma = (\gamma_1, \ldots, \gamma_k)^{\mathrm{T}}$, $\mathbf{1}_k$ is a k-dimensional vector containing only 1, β is a scale parameter,

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} = \frac{2}{1 + \exp(-2x)} - 1$$

is the hyperbolic tangent function, $\phi\left(x; u_\eta, \sigma_\eta\right)$ denotes the normal probability density function with mean u_η and variance σ_η^2, and P is the mixing distribution randomly drawn from a DP. The shape of P is flexible and could be

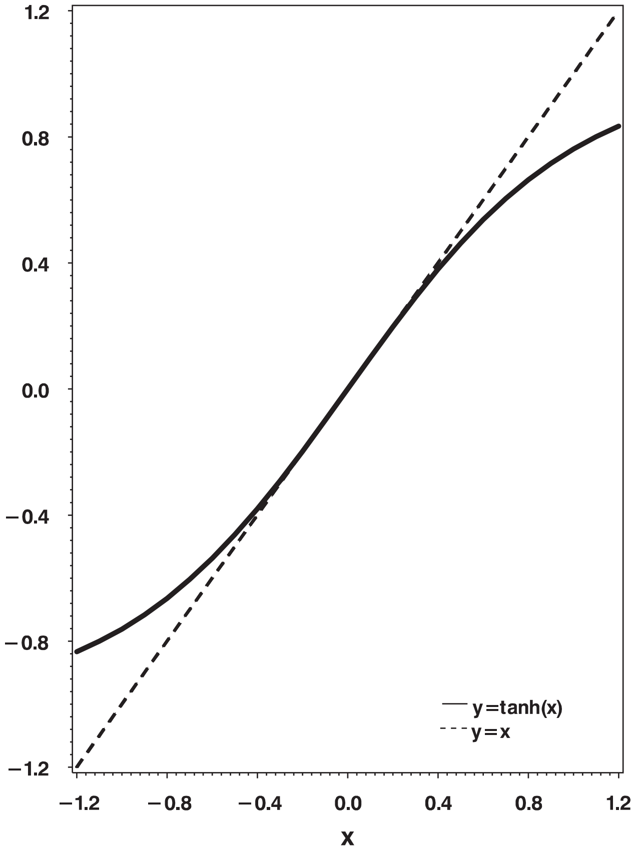


Fig. 1. $y=\tanh(x)$ versus $y=x$.

multi-modal, heavy-tailed, skewed, etc. If P is degenerate, (1) becomes a parametric model. If $k=0$ or $\beta=0$, (1) reduces to the Dirichlet mixture models for i.i.d. observations (Lo, 1984). Model (1) is ergodic for any P because the hyperbolic tangent function is bounded (Tang, 2006).

One particular motivation for choosing the kernel $\phi(X;g(u,\gamma,\beta,Z),\sigma)$ in (1) is that it can approximate any linear AR conditional density arbitrarily closely while imposing no constraint on parameters to ensure stationarity. Fig. 1 shows that $\tanh(x)\approx x$ when x is near 0. If $\max\left\{|\gamma_1|,\ldots,|\gamma_k|\right\}$ is small enough,

$$\phi(X;u+\beta\tanh(\gamma^{\mathrm{T}}(Z-u)),\sigma)\approx\phi(X;u+\beta\gamma^{\mathrm{T}}(Z-u),\sigma). \tag{2}$$

The right-hand side (RHS) of (2) is the conditional density of a linear AR model with mean u, autoregressive coefficients $\beta\gamma$, and i.i.d. normal error terms. Thus, we can use the kernel to approximate any linear AR model by choosing sufficiently small γ, and then adjusting β. In practice, this may be achieved through the specification of the prior distribution.

We put a DP prior on P. The parameters of a DP consist of a base measure G_0, and a scalar precision parameter $\alpha>0$. The base distribution G_0 is the best guess of what the true P is believed to be while the larger the value of α the closer a realization of the process is to G_0 (Ferguson, 1973). The DPMM model is equivalent to the following random effect autoregressive model, where associated with each (X_i,Z_i) is a latent variable $\theta_i=\left(u_i,\gamma_i^{\mathrm{T}},\sigma_i\right)$:

$$\begin{aligned}&X_i|Z_i,\theta_i,\beta\sim N\left(X_i;u_i+\beta\tanh\left(\gamma_i^{\mathrm{T}}(Z_i-u_i)\right),\sigma_i^2\right),\\&\theta_1,\ldots,\theta_n|P\overset{\text{i.i.d.}}{\sim}P,\end{aligned} \tag{3}$$

where $N\left(\cdot;u_\eta,\sigma_\eta^2\right)$ denotes the normal distribution with mean u_η and variance σ_η^2. Unlike that in a parametric model, the unknown parameters are varying along with the index of the observation, and are actually drawn as i.i.d. samples from an unknown distribution. West and Harrison (1997) considered similar models as convenient forecasting tools, where the random effects can have dependence.

The hierarchical representation (3) in terms of the latent variables θ_i's is very useful for developing the MCMC algorithm, as described in the next section. The random P could be further integrated out of the prior distribution, and the joint distribution of θ_i's is described by the generalized Polya Urn scheme (Blackwell and MacQueen, 1973):

$$\begin{aligned}&\theta_1 \sim G_0(\theta_1),\\ &\theta_{i+1}|\theta_1,\ldots,\theta_i \sim \frac{\alpha}{\alpha+i}G_0(\theta_{i+1}) + \sum_{j=1}^{i}\frac{1}{\alpha+i}\delta_{\theta_j}(\theta_{i+1}), \quad i \geqslant 1,\end{aligned} \tag{4}$$

where $\delta_{\theta_j}(\cdot)$ denotes a point mass at θ_j. The distribution (4) implies that θ_i's tend to share values in common. When a new observation arrives, it either takes the same value as some θ_j that has been drawn, or uses a new one generated from G_0.

The base distribution G_0 is typically constructed according to mathematical convenience while the parameters of G_0 are chosen to reflect prior beliefs. As the autoregressive function of the kernel in (1) is nonlinear, no G_0 is conjugate to the normal kernel. We specify G_0 as the following product measure:

$$G_0(u,\gamma,\sigma) = N\left(u; u_0, \sigma_u^2\right) Ga\left(\sigma^{-2}; v_1, v_2\right)\prod_{j=1}^{k} N\left(\gamma_j; 0, \sigma_\gamma^2\right),$$

where $Ga(\cdot; v_1, v_2)$ denotes the Gamma distribution with mean v_1/v_2 and variance v_1/v_2^2. The prior for the scale parameter β is $N\left(\beta; 0, \sigma_\beta^2\right)$.

3. Posterior inference

As G_0 is not conjugate to the normal kernel in (1), the standard MCMC scheme based on the generalized Polya urn scheme (Escobar and West, 1995) is not efficient because it involves numerical integration. Instead, we use the "no-gaps" MCMC scheme (MacEachern and Müller, 1998), which by-passes the numerical integration problem through parameter augmentation.

Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ denote the set of distinct θ_i's, where p is the number of distinct elements of $\theta_1, \ldots, \theta_n$ and $\phi_j = \left(\tilde{u}_j, \tilde{\gamma}_j^{\mathrm{T}}, \tilde{\sigma}_j\right)$. Let $\boldsymbol{s} = (s_1, \ldots, s_n)$ denote the configuration vector, that is, $s_i = j$ if and only if $\theta_i = \phi_j$. Thus, $\boldsymbol{\theta} = \{\theta_i : i = 1, \ldots, n\}$ is reparameterized as $\{\phi_1, \ldots, \phi_p, s_1, \ldots, s_n\}$. Let n_i $(i = 1, \ldots, p)$ be the number of elements $s_j = i$. In the no-gaps algorithm, the vector $\boldsymbol{\phi}$ is augmented with an additional set of variables as follows:

$$\Bigg\{\underbrace{\phi_1, \ldots, \phi_p}_{\phi_{\mathrm{F}}}, \underbrace{\phi_{p+1}, \ldots, \phi_n}_{\phi_{\mathrm{E}}}\Bigg\}.$$

The vectors ϕ_{F} and ϕ_{E} are, respectively, referred to as the full and the potential clusters. The augmentation relies upon the constraint that there is no gap in the values of s_i, that is, $n_j > 0$ for $j = 1, \ldots, p$ and $n_j = 0$ for $j = p+1, \ldots, n$. Let a subscript "$-i$" stand for all the variables except the ith one.

To implement the algorithm, initialize β, s and ϕ_{F} and repeat the following steps until the algorithm converges. Empirical evidence indicates that initializing $p = 1$ often leads to quick convergence, where β and ϕ_1 may be initialized according to the prior expectations or the least-squares fit of linear AR models.

(1) For $i = 1, \ldots, n$, repeat (ia) and (ib).

(ia) If $n_{s_i} > 1$, resample s_i from the multinomial distribution

$$\Pr\left(s_i = j|\boldsymbol{\phi}, s_{-i}, \beta, \boldsymbol{X}^{(n)}\right) = \begin{cases} c n_{-i,j} q_{ij}, & j = 1, \ldots, p_{-i}, \\ c\dfrac{\alpha}{p_{-i}+1} q_{ij}, & j = p_{-i}+1, \end{cases} \tag{5}$$

where $q_{ij}=\phi\left(X_i; g\left(\tilde{u}_j, \tilde{\gamma}_j, \beta, Z_i\right), \tilde{\sigma}_j\right)$ and c is a normalizing constant such that $\sum_{j=1}^{p_{-i}+1} \Pr\left(s_i = j|\boldsymbol{\phi}, s_{-i}, \beta, \boldsymbol{X}^{(n)}\right) = 1$.

(ib) If $n_{s_i} = 1$, with probability $1 - p^{-1}$, leave s_i unchanged. Otherwise rearrange the indices of ϕ_j, update s and n_j's correspondingly such that $s_i = p$, and then resample s_i according to (5).

(2) Update $\phi_i = \left(\tilde{u}_i, \tilde{\gamma}_i, \tilde{\sigma}_i^{-2}\right)$ for $i = 1, \ldots, n$. The ϕ_i's are conditionally independent given the cluster structure s and scale parameter β under the posterior distribution. The posterior distribution of $\phi_i \in \phi_{\mathrm{F}}$ is proportional to the product of the base measure G_0

$$\phi\left(\tilde{u}_i; u_0, \sigma_u\right)\left[\prod_{l=1}^{k} \phi\left(\tilde{\gamma}_{il}; 0, \sigma_\gamma\right)\right] Ga\left(\tilde{\sigma}_i^{-2}; v_1, v_2\right)$$

and the likelihood of the data associated with ϕ_i

$$\prod_{j:s_j=i} \phi\left(X_j; g\left(\tilde{u}_i, \tilde{\gamma}_i, \beta, Z_j\right), \tilde{\sigma}_i\right).$$

We update $\tilde{\sigma}_i$ with a Gibbs step, $\tilde{u}_i$ with a random walk Metropolis step and $\tilde{\gamma}_i$ with a random walk Metropolis step. The candidate $\tilde{u}_i^*$ is generated from $N\left(\tilde{u}_i^*; \tilde{u}_i, V_u\right)$ and the candidate $\tilde{\gamma}_i^* = \left(\tilde{\gamma}_{l1}^*, \ldots, \tilde{\gamma}_{lk}^*\right)^{\mathrm{T}}$ is drawn from the product measure $\prod_{l=1}^{k} N\left(\tilde{\gamma}_{il}^*; \tilde{\gamma}_{il}, V_\gamma\right)$. The values of V_u and V_γ are determined automatically by the computer program in the early stage of the burn-in period such that the average acceptance probabilities of $\tilde{u}_i$'s and $\tilde{\gamma}_i$'s roughly lie in [0.20, 0.65] and then remain unchanged. The choice of V_u and V_γ makes a compromise between the jump distance in the parameter space and the acceptance frequency, both of them ensure the efficiency of the MCMC algorithm (Tierney, 1994). As no data are associated with $\phi_i \in \phi_{\mathrm{E}}$, the posterior distribution for $\phi_i \in \phi_{\mathrm{E}}$ is simply G_0. Since the number of distinct components p is typically small, $\phi_i \in \phi_{\mathrm{E}}$ is generated only if a new component is needed.

(3) Update β with a Gibbs step. The posterior distribution of β is normal

$$\beta|\boldsymbol{\phi}, s, \boldsymbol{X}^{(n)} \propto \phi\left(\beta; 0, \sigma_\beta\right) \prod_{i=1}^{n} \phi\left(X_i; g\left(u_i, \gamma_i, \beta, Z_i\right), \sigma_i\right) \propto N\left(\beta; u_{\beta^*}, \sigma_{\beta^*}^2\right),$$

where $\sigma_{\beta^*}^2 = \left[\sum_{j=1}^{n} \sigma_i^{-2} \tanh^2\left(\gamma_i^{\mathrm{T}}\left(Z_i - u_i \mathbf{1}_k\right)\right) + \sigma_\beta^{-2}\right]^{-1}$ and $u_{\beta^*} = \sigma_{\beta^*}^2 \left[\sum_{j=1}^{n} \sigma_i^{-2}\left(y_i - u_i\right) \tanh\left(\gamma_i^{\mathrm{T}}\left(Z_i - u_i \mathbf{1}_k\right)\right)\right]$.

MacEachern and Müller (1998) showed that the no-gaps algorithm converges almost surely under a very mild sufficient condition. In our empirical studies, convergence of the no-gaps algorithm, assessed through multiple chain together with informal graphical methods, was found to be quick. A burn-in period of 10, 000 was adequate for all examples.

Estimation of the conditional density is based on the evaluation of

$$\hat{f}(x|z) = N^{-1} \sum_{t=1}^{N} E\left[\int \phi(x; g(u, \gamma, \beta, z), \sigma)\, \mathrm{d}P(u, \gamma, \sigma)|\beta^{(t)}, \theta_1^{(t)}, \ldots, \theta_n^{(t)}\right],$$

where $\left(\beta^{(t)}, \theta_1^{(t)}, \ldots, \theta_n^{(t)}\right)$ denotes the posterior sample at step t after the burn-in period, and

$$E\left[\int \phi(x; g(u, \gamma, \beta, z), \sigma)\, \mathrm{d}P(u, \gamma, \sigma)|\beta^{(t)}, \theta_1^{(t)}, \ldots, \theta_n^{(t)}\right]$$
$$= \frac{\alpha \int \phi\left(X; g\left(u, \gamma, \beta^{(t)}, z\right), \sigma\right) \mathrm{d}G_0(u, \gamma, \sigma)}{\alpha + n} + \frac{\sum_{j=1}^{n} \phi\left(x; g\left(u_j^{(t)}, \gamma_j^{(t)}, \beta^{(t)}, z\right), \sigma_j^{(t)}\right)}{\alpha + n}.$$

The first term on the RHS of the above equation can be assessed using Monte Carlo integration, and the numerator of the second term can be simplified as $\sum_{j=1}^{p} n_j^{(t)} \phi\left(x; g\left(\tilde{u}_j^{(t)}, \tilde{\gamma}_j^{(t)}, \beta^{(t)}, z\right), \tilde{\sigma}_j^{(t)}\right)$.

4. Posterior consistency

Consistency may be thought of as a validation of the Bayesian procedure. As the sample size goes to infinity, the Bayesian assessment of the unknown parameter should be close to the true value if it exists. Consistency implies that the inference will eventually be free of the prior distribution. This section establishes posterior consistency for model (1) in two topologies.

Let μ be the prior for β and $DP_{\alpha G_0}$ be the DP prior on P. The distributions μ and G_0 discussed in this section include, but are not limited to, that defined in Section 2. Let the true conditional density be of the form $f_0(x|z) = \int \phi\left(z; g\left(u, \gamma, \beta_0, z\right), \sigma\right) \mathrm{d}P_0(u, \gamma, \sigma)$ for some β_0 and P_0 lying in the support of the prior distribution. Let $\mathscr{F} = \{f(x|z) = \int \phi(x; g(u, \gamma, \beta, z), \sigma)\, \mathrm{d}P(u, \gamma, \sigma) : \beta \in \operatorname{supp}(\mu),\ P \in \operatorname{supp}\left(DP_{\alpha G_0}\right)\}$ denote the space of the conditional densities. The posterior distribution is said to be consistent at f_0 if for every neighborhood U of f_0, we have that $\Pi\left(U|X^{(n)}\right) \to 1$ a.s., where Π denotes the posterior distribution. Here, all probability statements are given relative to the true distribution.

The notion of consistency depends on the notion of a topology. We consider the integrated L_1 and sup-L_1 metrics on the conditional densities. The integrated L_1 distance between two conditional densities f_1 and f_2 is defined as

$$d_v\left(f_1, f_2\right) = \int \|f_1(\cdot|z) - f_2(\cdot|z)\| v(z)\, \mathrm{d}z = \iint |f_1(y|z) - f_2(y|z)| v(z)\, \mathrm{d}y\, \mathrm{d}z, \tag{6}$$

where $v(z)$ is a probability density function assumed to be strictly positive for any $z \in \Re^k$. Different v leads to different metrics, but all these topologies are equivalent. The sup-L_1 distance is defined as

$$d_s(f_1, f_2) = \sup_{z \in \Re^k} \|f_1(\cdot|z) - f_2(\cdot|z)\|. \tag{7}$$

The sup-L_1 metric is much stronger than the integrated L_1 metric.

Posterior consistency has been explored for a wide class of order-1 Markov processes in Tang and Ghosal (2006) and Ghosal and Tang (2005). A classical tool for posterior consistency in the i.i.d. case is the Schwartz's theorem (Schwartz, 1965; Ghosal et al., 1999), which was extended to ergodic Markov processes by Tang and Ghosal (2006). The theory essentially requires two sufficient conditions:

(1) *Kullback–Leibler positivity*: $\Pi\left(f : K\left(f_0, f\right) < \varepsilon\right) > 0$ for any $\varepsilon > 0$, where $K\left(f_0, f\right) = \iint \pi_0(z) f_0(y|z) \log\left(f_0(y|z)/f(y|z)\right) \mathrm{d}y\, \mathrm{d}z$ and $\pi_0(z)$ is the invariant distribution of Z_i's.

(2) *Uniformly exponentially consistent tests*: There exist $\beta > 0$ and a sequence of tests $\phi_n\left(X^n\right)$ for testing $f = f_0$ *versus* $f \in U^c$ such that

$$E_{f_0}\left(\phi_n\right) < \exp(-n\beta) \quad \text{and} \quad \sup_{f \in U^c} \mathrm{E}_f\left(1 - \phi_n\right) < \exp(-n\beta),$$

for all $n > n_0$, where U is any neighborhood of f_0 and n_0 is a positive integer.

The condition of Kullback–Leibler positivity implies the ergodicity of the AR process, and needs to be shown irrespective of the topology under consideration. For order-1 Markov process ($k = 1$), Tang and Ghosal (2006) showed that the space $\mathscr{F}$ is compact in the sup-L_1 metric under the assumptions that the AR link function is bounded and uniformly equicontinuous for $Z \in \Re$ and certain other regularity conditions, and then constructed uniformly exponentially consistent tests for each of finitely many balls that partition $\mathscr{F}$; see also Theorem 2 below. For bounded link functions, uniformly exponentially consistent tests can also be constructed in certain weaker topologies using the Hoeffding's inequalities (Tang and Ghosal, 2006; Tang, 2006). For more general link functions, the construction of tests is typically difficult. The construction of tests can be avoided using a martingale-based approach, which is introduced for i.i.d. cases by Walker (2003, 2004), and extended to Markov processes in Tang and Ghosal (2006) and Ghosal and Tang (2005). Although the martingale-based approach does not work for the sup-L_1 metric, it does not require the boundness assumption on the AR link function. Tang and Ghosal (2006) showed the posterior consistency of the DPMMs for Markov processes in the integrated L_1 metric under the following sufficient conditions: (1) Kullback–Leibler positivity, (2) the link function is uniformly equicontinuous for $Z \in C$ for any compact $C \subset \Re$, and (3) certain other regularity conditions; see also Theorem 1 below.

The following two theorems show the consistency of Model (1) in the integrated L_1 and sup-L_1 metrics. We omit the proof as the arguments are essentially similar to that in Tang and Ghosal (2006). Although we focus on model (1) in this paper, most results in Tang and Ghosal (2006) and Ghosal and Tang (2005) can be extended easily to DPMMs for ergodic time series with other AR link functions.

Theorem 1. *If the supports of* μ *and* G_0 *are both compact, the posterior consistency holds at* f_0 *in the integrated* L_1 *metric. That is, for any* $\varepsilon > 0$, $\Pi\left(U|\boldsymbol{X}^{(n)}\right) \to 1$ *a.s., where* $U = \{f : \int \|f(\cdot|z) - f_0(\cdot|z)\| v(z)\,\mathrm{d}z \leqslant \varepsilon\}$.

Remark 1. Proof of Theorem 1 is similar to that for Proposition 5.1 and Theorem 7.1 of Tang and Ghosal (2006) using the following facts:

- For any compact $C \in \Re^k$, the family of functions

  $$\{x \mapsto g(u, \gamma, \beta, z) : z \in C, \beta \in \mathrm{supp}(\mu), (u, \gamma) \in \mathrm{supp}(G_0)\}$$

  is uniformly equicontinuous; that is, for any $\varepsilon > 0$, there exists $\delta > 0$ such that $\sup_{u,\gamma,\beta}|g(u, \gamma, \beta, z_1) - g(u, \gamma, \beta, z_2)| < \delta$ whenever $\|z_1 - z_2\| < \varepsilon$.
- The space of conditional densities $\mathscr{F}$ is compact with respect to the semi-metric d_C defined as $d_C(f_1, f_2) = \sup_{z \in C} \|f_1(\cdot|z) - f_2(\cdot|z)\|$.

Theorem 2. *Under the assumption of Theorem* 1 *and that no point in the support of* G_0 *satisfies* $\prod_{j=1}^k \gamma_j = 0$, *the posterior consistency holds at* f_0 *in the* sup-L_1 *metric. That is, for any* $\varepsilon > 0$, $\Pi\left(U|\boldsymbol{X}^{(n)}\right) \to 1$ *a.s., where* $U = \{f : \sup_{z \in \Re^k} \|f(\cdot|z) - f_0(\cdot|z)\| \leqslant \varepsilon\}$.

Remark 2. The assumption of Theorem 2 ensures that the family of functions $\{x \mapsto g(u, \gamma, \beta_0, Z) : z \in \Re^k, \beta \in \mathrm{supp}(\mu), (u, \gamma) \in \mathrm{supp}(G_0)\}$ is uniformly equicontinuous, and hence that the space of conditional densities $\mathscr{F}$ is compact with respect to the sup-L_1 metric. This theorem can be proved in a way similar to that for Theorem 8.1 of Tang and Ghosal (2006).

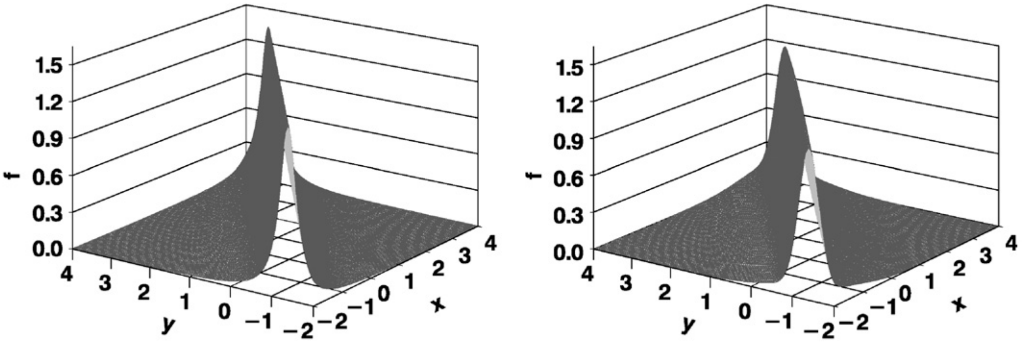
Remark 3. The assumption that no point in the support of G_0 satisfies $\prod_{j=1}^k \gamma_j = 0$ may not be dropped. See Remark 8.1 of Tang and Ghosal (2006). When $k = 1$, the assumption $\gamma_1 \neq 0$ is not restrictive since the order-1 Markov process reduces to the i.i.d. model under either condition $\beta = 0$ or $\gamma_1 = 0$.

We have put the compactness assumption on the supports of μ and G_0 to show consistency for technical reasons. We can use truncated prior distribution. If the compact support is chosen to be large enough, the truncation should have no or little effect on the numerical results. In application, we use the prior specified in Section 2.

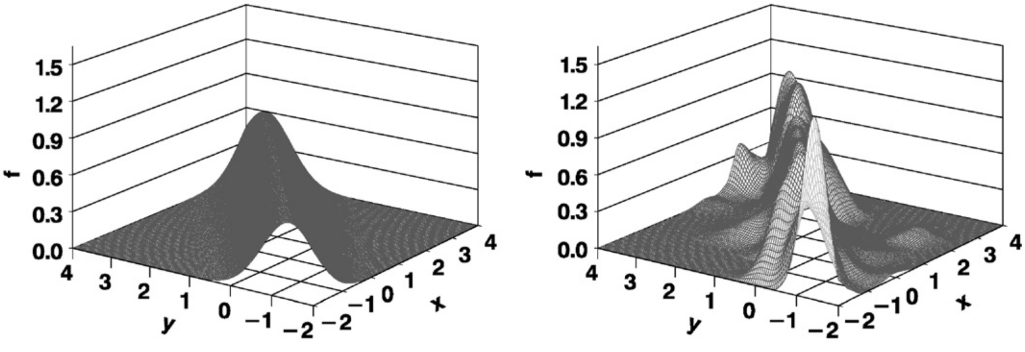
5. Simulations

This section illustrates the performance of the DPMM with simulated examples. Throughout, we set the prior parameters as follows: $\alpha = 1$, $u_\beta = 0$, $\sigma_\beta^2 = 10^{12}$, $u_0 = \bar{x}$, $\sigma_u = 3\hat{s}$, $u_\gamma = 0$, $\sigma_\gamma = 1/(3\hat{s}k)$, $v_1 = 1/(1000s^2)$ and $v_2 = 1/1000$, where $\bar{x}$ is the mean and $\hat{s}$ is the standard error of the observed time series. The σ_γ is chosen to be small to reflect the prior belief that the underlying process is approximately a linear model or mixture of linear models. The prior densities for β, u_i's and σ_i's are quite flat in the parameter space to reflect the vague prior belief. Our experience suggests that the posterior inference is not very sensitive to the particular prior parameter choice within a wide range.

In simulation one and two, the DPMM is compared with the MAR of Wong and Li (2000) and double kernel local linear approach of Fan et al. (1996) for estimating the conditional density of order-1 AR process. The MAR models are fitted with the MIXREG package in R software. The kernel methods are fitted with the HDRCDE package in R software, where the bandwidth is chosen according to the normal reference rule (Hyndman and Yao, 2002). We also try a series of different bandwidths in the kernel approach. For all examples studied in this paper, the shapes of the kernel



(a) Truth (b) DPMM



(c) AR(1) (d) kernel method

Fig. 2. Simulation one: plots of true and estimated conditional densities: (a) Truth, (b) DPMM, (c) AR(1), (d) Kernel method.

conditional density estimates seem quite insensitive to the choice of bandwidth in the z-direction. Similar to Fan and Yim (2004), for each simulated sample, the performance of each method is evaluated by the root mean-squared error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_j \left[\hat{f}\left(x_j|z_j\right) - f\left(x_j|z_j\right)\right]^2 I\left(z_j \in [a,b], x_j \in [a,b]\right)}{\sum_j I\left(z_j \in [a,b], x_j \in [a,b]\right)}},$$

where $\left(z_j, x_j\right)$'s are grid points that are even distributed across certain regions of interest and $[a, b]$ is an interval where most observed data lie.

Simulation 1: This simulation illustrates the DPMM for handling non-normal data. A sample of size 150 is simulated from a simple AR(1) model

$$X_i = \tau + \rho X_{i-1} + \varepsilon_i \quad (i \geqslant 1), \tag{8}$$

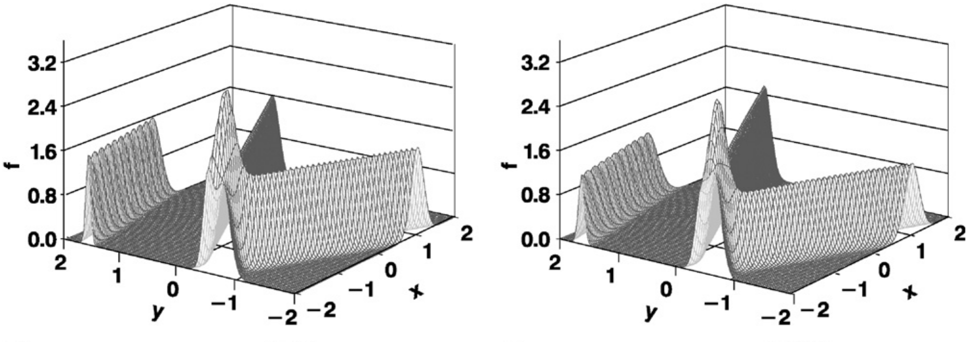
where $X_0 = 0$, $\rho = 0.8$, $\tau = 0.2$ and $\{\varepsilon_i\}$ are independent random variables from the Cauchy distribution $f(\varepsilon_i) = 4/[\pi(1 + 16\varepsilon_i^2)]$ truncated to the interval $[-3, 3]$. Fig. 2 plots the true and estimated conditional densities. AR(1) fails to detect the sharp central peak in the true conditional density because of the normal assumption on the error terms. The kernel estimate in Fig. 2(d) is not smooth. There is a trade-off between goodness of fit and smoothness in the kernel approach while choice of the right amount of smoothing is not a serious problem in DPMM. The conditional density estimated by DPMM (Fig. 2(b)) is the closest to the truth.

Table 1 presents the mean and standard error of the RMSE evaluated over 100 random samples of size 150. For each replication, we estimate $f\left(x_j|z_j\right)$ on a 51×51 grid on the space $[a = -2, b = 4] \times [-2, 4]$. The DPMM shows the best performance in term of the RMSE criterion.

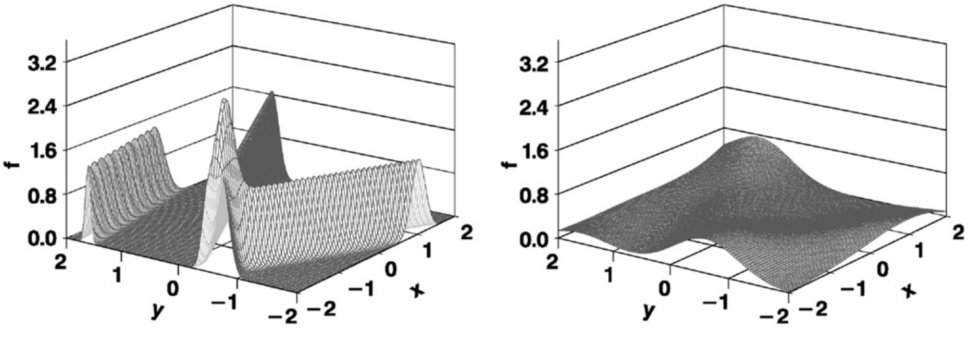
Table 1
Summary of the RMSE for simulation one and two

	Simulation one		Simulation two	
	Mean	Std	Mean	Std
DPMM	0.076	0.035	0.090	0.022
MAR	0.170	0.015	0.093	0.035
Kernel	0.206	0.071	0.419	0.006

The mean and standard error (std) of the RMSE are evaluated over 100 simulated samples.



(a) Truth (b) DPMM



(c) MAR (d) kernel method

Fig. 3. Simulation two: plots of true and estimated conditional densities: (a) Truth, (b) DPMM, (c) MAR, (d) kernel method.

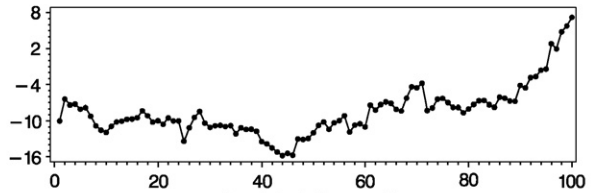
Simulation 2: We simulate 100 random samples of size 150 from the following 3-component MAR model:

$$f\left(X_i|X_{i-1}\right)=\tfrac{1}{6}\phi\left(X_i;\tfrac{1}{5}X_{i-1}+2,\tfrac{1}{24}\right)+\tfrac{1}{2}\phi\left(X_i;\tfrac{1}{2}X_{i-1}+\tfrac{1}{4},\tfrac{1}{8}\right)+\tfrac{1}{3}\phi\left(X_i;-\tfrac{1}{2}X_{i-1}-\tfrac{3}{2},\tfrac{1}{12}\right)\quad(i\geqslant 1),$$

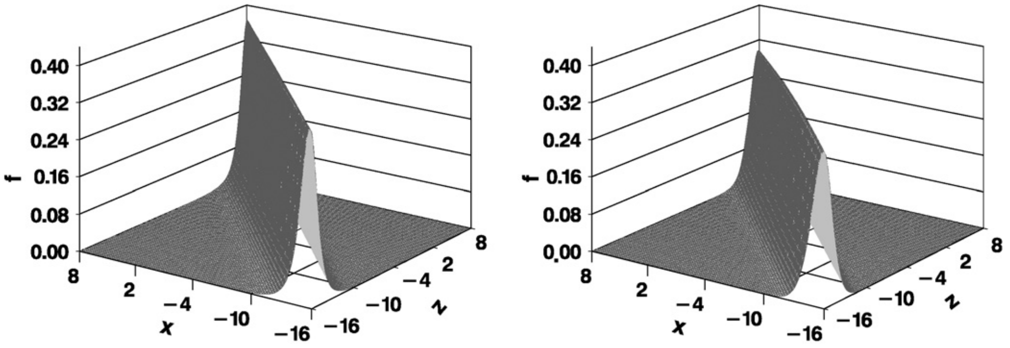
where $X_0=0$. Fig. 3 plots the true conditional density and that estimated using one random realization. Both DPMM and MAR recover the true shape of the conditional density reasonably well. The kernel method fails for this example. Table 1 summarizes the results based on 100 replications, where $f\left(x_j|z_j\right)$ is estimated over 51×51 grid points evenly distributed on $[-2,2]\times[-2,2]$. DPMM and MAR show comparable performance while the mean of RMSE produced by the kernel method is almost four times bigger than that by DPMM and MAR. The advantage of DPMM is that the number of mixture components is not specified in advance and is automatically determined by data. This example demonstrates the ability of the DPMM in detecting multimodality.

Simulation 3: A nearly non-stationary time series of size 100 is simulated from

$$X_i=0.99X_{i-1}+\varepsilon_i\quad(i\geqslant 1),\tag{9}$$



(a) simulated time series



(b) true conditiional density (c) density estimated by DPMM

Fig. 4. Simulation 3: simulated nearly non-stationary data: (a) simulated time series, (b) true conditional density, (c) density estimated by DPMM.

where $X_0 = -10$ and $\{\varepsilon_i\}$ are independent and follow t_3-distribution. Fig. 4 plots the simulated data. This particular realization is interesting. The AR process defined by model (9) is ergodic while the observed series seem to be non-stationary. The best MAR model (model order k is fixed at 1) identified by the BIC criterion is a non-stationary AR(1) model with conditional expectation $E(X_i|X_{i-1}) = 0.388 + 1.025X_{i-1}$. In DPMM, the ergodicity of the process is enforced. The conditional density estimated by DPMM (Fig. 4(c)) seems close to the truth (Fig. 4(b)).

6. Analysis of Canadian lynx data

This section analyzes the famous Canadian lynx data. The data set consists of the annual record of the Canadian lynx trapped in the Mackenzie river district in northwest Canada for the period 1821–1934 inclusive, with a total of 114 observations. The data have been analyzed by many methods; see Wong and Li (2000) for a survey. The log-transformed time series is displayed in Fig. 5. Clearly, the data are cyclical with a period of approximate 10 years and the ascent periods (around 6 years) tend to exceed the descent periods (around 4 years) by approximately 50%. So this series is time-irreversible and nonlinear. The histogram (Fig. 6(a)) indicates that the marginal distribution is bimodal. It would be interesting to explore whether the conditional distribution is multimodal. Some parametric models such as ARIMA and threshold autoregressive model (Tong, 1990) are not suitable for this question since they implicitly assume that the underlying conditional density is unimodal. Wong and Li (2000) calculated the rate of ascent ((change in X/number of years) during an ascending period) and rate of descent ((change in X/number of years) during a descending period) for each cycle and found that the rates of descent were more variable than the rates of ascent. Since in theory the conditional variances for multimodal conditional distributions tend to be larger than that for unimodal conditional distributions, the phenomenon observed by Wong and Li (2000) suggests that there might exist multimodality in the conditional densities.

We fit the DPMM. Partly for simplicity, we use the prediction mean square error (PMSE) criterion to determine the model order k

$$\mathrm{PMSE} = m^{-1} \sum_{i=n-m+1}^{n} (X_i - \hat{X}_i)^2, \tag{10}$$

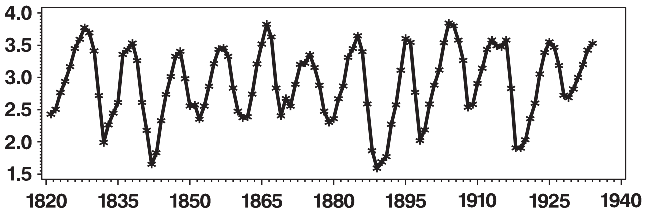
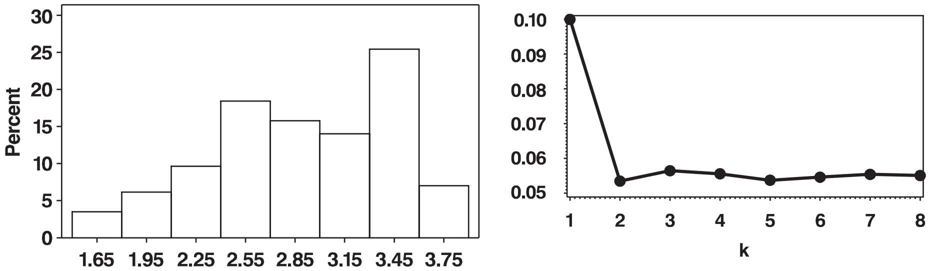


Fig. 5. Common logarithm of the Canadian lynx data.



(a) Histogram

(b) PMSE estimated by DPMM using last 30 observations

Fig. 6. Lynx data: (a) Histogram, (b) PMSE estimated by DPMM using last 30 observations.

where $\hat{X}_i$'s are the one-step ahead prediction of X_i and the last m observations are used to evaluate PMSE. Fig. 6(b) plots the PMSE estimated using the last $m = 30$ observations. The PMSE drops dramatically from $k = 1$ to 2 and remains approximately constant when $k \geqslant 2$, suggesting a DPMM of order $k = 2$ might be suitable for this data. The PMSE criterion seems insensitive to the choice of m in this example, as the estimated PMSE exhibits similar pattern for $m = 10, 15, 20, 25, 35, 40$. The estimated conditional densities for the last cyclical period (from $i = 106$ to 113) are depicted in Fig. 7 (solid line), together with the actual values of x_i at $i - 1$, i and $i + 1$. The conditional densities are unimodal when the series is ascending to a peak and show weak bimodalities during a descending period ($i = 106$–108).

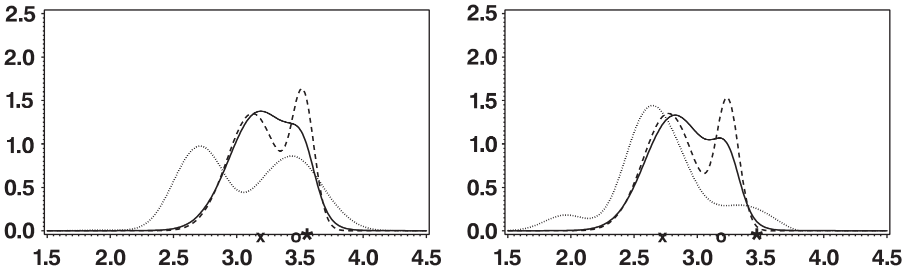
Fig. 7 also displays the conditional densities estimated by the MAR (dashed line) and the kernel method (dotted line). The MAR estimates are based on the model fitted by Wong and Li (2000):

$$\begin{aligned} f(X_i|X_{i-1}, X_{i-2}) = {} & 0.3163\phi(X_i; 0.7107 + 1.1022X_{i-1} - 0.2835X_{i-2}, 0.0877) \\ & + 0.6837\phi(X_i; 0.9784 + 1.5279X_{i-1} - 0.8871X_{i-2}, 0.2020). \end{aligned}$$

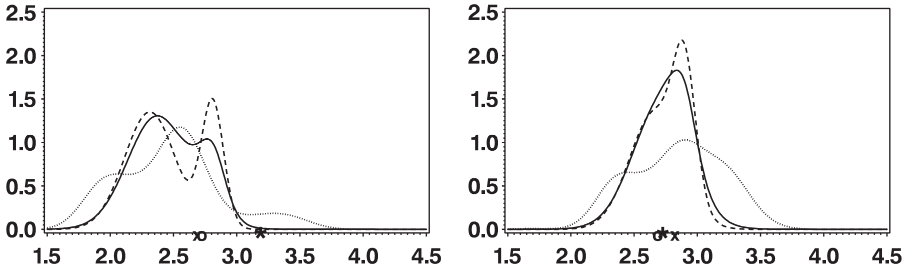
Similarly, the MAR estimates are bimodal only when the series is descending to a trough. However, the MAR estimates exhibit much sharper peaks than the DPMM estimates. On the contrary, the estimated kernel conditional densities seem quite flat. The kernel estimates show strong bimodality at step $i = 106$ and weak multimodality at many other steps. However, all multimodalities except the one at $i = 106$ will disappear if we increase the bandwidth in the z-direction gradually. Wong and Li (2001) also fitted the data with the logistic mixture autoregressive model. Although the shapes of the estimated conditional densities vary greatly with the methods, all the above models support that the conditional densities contain multimodalities.

7. Discussion

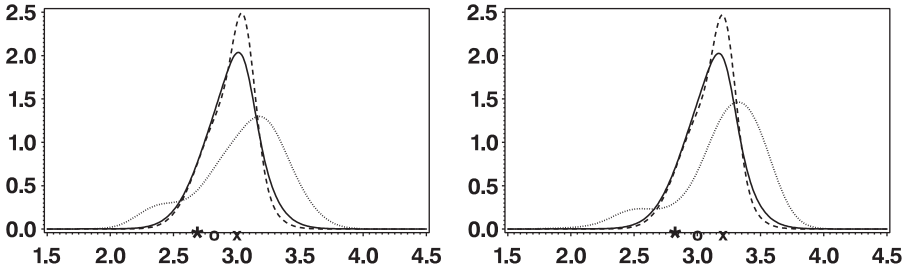
We have introduced a Bayesian nonparametric model for estimating the conditional density of ergodic time series and shown that it is consistent in two different topologies. The merit of mixture-type AR models lies in their ability to describe the conditional density and model changing conditional variance of the time series (Wong and Li, 2000). We further demonstrate that the Dirichlet mixture model is more flexible than the finite MAR model of Wong and Li (2000) in the analysis of data with multimodality and with non-normality. It is straightforward to incorporate exogeneous variables into the model. Although we focus on the analysis of time series, the proposed model could be viewed as a



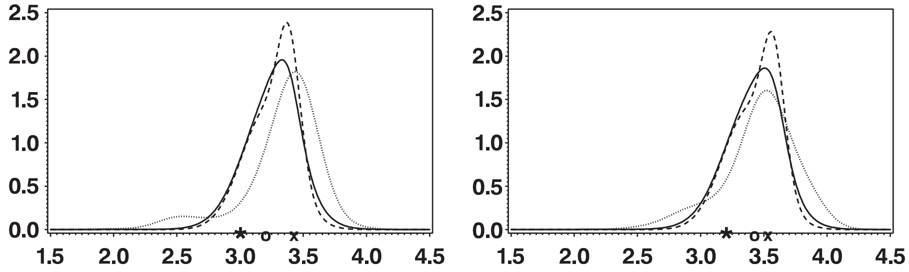
(a) $i = 106$ (b) $i = 107$



(c) $i = 108$ (d) $i = 109$



(e) $i = 110$ (f) $i = 111$



(g) $i = 112$ (h) $i = 113$

Fig. 7. Lynx data: estimated conditional densities $\hat{f}(\cdot|x_i, x_{i-1})$ from $i = 106$ to 113: DPMM (solid curve), MAR (dashed curve) and kernel method (dotted curve). The symbols '*','o' and 'x' on the horizontal axis represent the actual values of x_i at $i-1, i, i+1$, respectively.

semi-parametric model suitable for the regression problem, where X_i's are the dependent variables and Z_i's are the explanatory variables. Posterior consistency in semi-parametric regression models with independent errors has been established by Amewou-Atisso et al. (2003).

Because of the special properties of the hyperbolic tangent function, we have chosen it as the AR link function to model a strictly ergodic time series. However, other choices are possible. Tang (2006) shows that the Dirichlet mixture model (1) is uniformly ergodic for any bounded link function g. For more general link functions, the parameters of the model are usually subject to complex constraints in order to ensure the AR process to be stable, causing difficulty in specifying prior distributions. For linear AR models to be stationary, the roots of the AR characteristic polynomial need to lie in the unit circle. Sufficient conditions for finite mixture AR models to be stationary were established by Wong and Li (2000), and that for Dirichlet mixture models for Markov processes were explored in Tang and Ghosal (2006). In Huerta and West (1999), a prior structure is defined directly on the roots of the AR characteristic polynomial to model stationary linear time series. Similarly, a potential way to model ergodic nonlinear time series is to put hyperpriors on the roots of the AR characteristic polynomial through the DP.

Partly for simplicity, we use the PMSE criterion to determine the order of the AR process in the analysis of Canadian lynx data. It works well for this example. However, the PMSE criterion works only for long time series, and may be sensitive to the choice of m especially when m is small (m is the number of observations for evaluating PMSE). As suggested by the referees, a more elegant way to deal with the model order is to treat it as an additional model parameter using the reversible jump MCMC methods (Richardson and Green, 1997). Such a full Bayesian approach for model selection has been applied to the problem of parametric Bayesian inference of AR processes (Vermaak et al., 2004). It is possible to extend this approach to Dirichlet mixture models.

Acknowledgments

We would like to thank the referees and the editor for their helpful comments and suggestions.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 2003. Posterior consistency for semiparametric regression problems. Bernoulli 9 (2), 291–312.

Blackwell, D., MacQueen, J., 1973. Ferguson distributions via polya urn schemes. Ann. Statist. 1, 353–355.

Choudhuri, N., Ghosal, S., Roy, A., 2005. Bayesian methods for function estimation. In: Dey, D. (Ed.), Handbook of Statistics, vol. 25. Marcel Dekker, New York, pp. 373–414.

Clemencon, S.J.M., 2000. Adaptive estimation of the transition density of a regular Markov chain. Math. Methods Statist. 9, 323–357.

Escobar, M., West, M., 1995. Bayesian density estimation and inference using mixtures. J. Am. Statist. Assoc. 90, 577–588.

Fan, J., Yao, Q., Tong, H., 1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika 83, 189–206.

Fan, J., Yim, T., 2004. A crossvalidation method for estimating conditional densities. Biometrika 91 (4), 819–834.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1, 209–230.

Ghosal, S., Tang, Y., 2005. Bayesian consistency for Markov processes. Sankhya, to appear.

Ghosal, S., van der Vaart, A.W., 2001. Entropies and rates of convergence of maximum likelihood and Bayes estimation for mixtures of normal densities. Ann. Statist. 29 (5), 1233–1263.

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 1999. Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist. 27 (1), 143–158.

Ghosal, S., Ghosh, J.K., van der Vaart, A.W., 2000. Convergence rates of posterior distributions. Ann. Statist. 28 (2), 500–531.

Gu, C., Wang, J., 2003. Penalized likelihood density estimation: direct cross-validation and scalable approximation. Statist. Sinica 13, 811–826.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, Berlin.

Huerta, G., West, M., 1999. Priors and component structures in autoregressive time series models. J. Roy. Statist. Soc. Ser. B 61 (4), 881–899.

Hyndman, R., Yao, Q., 2002. Nonparametric estimation and symmetry tests for conditional density functions. J. Nonparametric Statist. 14 (3), 259–278.

Jeon, Y., Lin, Y., 2006. An effective method for high dimensional log-density ANOVA estimation, with application to nonparametric graphical model building. Statist. Sinica 16, 353–374.

Lo, A.Y., 1984. On a class of Bayesian nonparametric estimates I: density estimates. Ann. Statist. 12, 351–357.

MacEachern, S.N., Müller, P., 1998. Estimating mixture of Dirichlet process models. J. Comput. Graphical Statist. 7, 223–228.

Müller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83, 67–79.

Müller, P., West, M., MacEachern, S., 1997. Bayesian models for non-linear autoregressions. J. Time Ser. Anal. 18 (6), 593–614.

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components. J. Roy. Statist. Soc. Ser. B 59 (4), 731–792.

Rosenblatt, M., 1969. Conditional probability density and regression estimates. In: Krishnaiah, P. (Ed.), Multivariate Analysis II. Academic Press, New York, pp. 25–31.

Schwartz, L., 1965. On Bayes procedures. Z. Wahr. Verw. Gebiete 4, 10–26.

Tang, Y., 2006. A Hoeffding-type inequality for ergodic time series. J. Theor. Probab., in press, preprint. Downloadable from ⟨http://www4.stat.ncsu.edu/∼sghosal/papers/Tang.pdf⟩.

Tang, Y., Ghosal, S., 2006. Posterior consistency of Dirichlet mixtures for estimating a transition density. J. Statist. Plann. Inference, in press, available online 9 June 2006.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22, 1701–1762.

Tong, H., 1990. Non-Linear Time Series. Oxford University Press, New York.

Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.J., 2004. Reversible jump Markov Chain Monte Carlo strategies for Bayesian model selection in autoregressive processes. J. Time Ser. Anal. 25 (6), 785–809.

Walker, S.G., 2003. On sufficient conditions for Bayesian consistency. Biometrika 90, 482–488.

Walker, S.G., 2004. New approaches to Bayesian consistency. Ann. Statist. 32, 2028–2043.

West, M., Harrison, J., 1997. Bayesian Forecasting and Dynamic Models. second ed.. Springer, New York.

Wong, C.S., Li, W.K., 2000. On a mixture autoregressive model. J. Roy. Statist. Soc. Ser. B 62 (1), 91–115.

Wong, C.S., Li, W.K., 2001. On a logistic mixture autoregressive model. Biometrika 88, 833–846.