A Nonparametric Model for Stationary Time Series

Isadora Antoniano-Villalobos[*]
Bocconi University, Milan, Italy.
isadora.antoniano@unibocconi.it

Stephen G. Walker
University of Texas at Austin, USA.
s.g.walker@math.utexas.edu

[*] PhD research funded by CONACyT.

Abstract

Stationary processes have been used as statistical models for dependent quantities evolving in time. Stationarity is a desirable model property, however, the need to define a stationary density limits the capacity of such models to incorporate the diversity of the data arising in many real life phenomena. Alternative models have been proposed, usually resulting in a compromise, sacrificing the ability to establish properties of estimators, in favor of greater modeling flexibility.

In this paper we present a family of time-honogeneous processes with nonparametric stationary densities, which retain the desirable statistical properties for inference, while achieving substantial modeling flexibility, matching those achievable with certain non–stationary models. For the sake of clarity we restrict attention to first order Markov processes.

Posterior simulation involves an intractable normalizing constant. We therefore present a latent extension of the model which enables exact inference through a trans-dimensional MCMC method. Numerical illustrations are presented.

Keywords: Markov model; Mixture of Dirichlet process model; Latent model; Dependent Dirichlet process; Time-homogeneous process.

1 Introduction

Since the advent of Bayesian posterior inference via simulation techniques (Escobar, 1988; Gelfand and Smith, 1990), it has been possible to estimate Bayesian nonparametric models. While the mixture of Dirichlet process (MDP) model, introduced by Antoniak (1974) and Lo (1984), remains one of the most popular models, the advances in simulation techniques have now allowed models to move away from standard set–ups involving independent and identically distributed observations, to cover more complex data structures, such as regression models and time series models. We cite the book of Hjort et al. (2010) which contains examples, references, and discussions of these various models; and specifically Chapter 6, for a review on MCMC based methods.

Before proceeding it is useful here to establish some notation. For the independent and identically distributed case, assume we have data $(y_1, \ldots, y_n)$. The basic mixture model takes the form

$$f(y) = \int k(y|\theta)\,\mathrm{d}P(\theta),$$

where $k(\cdot|\theta)$ is a density for all $\theta \in \Theta$ and P is a distribution function on Θ. If the prior for P is assigned as a Dirichlet process (Ferguson, 1973) then, according to Sethuraman (1994), there is a stick–breaking representation for P given by

$$P = \sum_{j=1}^{\infty} w_j\, \delta_{\theta_j},$$

where the weights (w_j) are defined in terms of independent and identically distributed (v_j), from the beta$(1, c)$ density, for some $c > 0$, as $w_1 = v_1$ and subsequently, for $j > 1$,

$$w_j = v_j \prod_{l<j}(1 - v_l).$$

The (θ_j) are typically taken as independent and identically distributed from some density function $g(\theta)$, and independently of the weights. Other stick–breaking constructions are possible, based on alternative beta distributions, which, obviously, correspond to different priors on P. See Ishwaran and James (2001) for more details.

Perhaps one area that has not been fully exploited from a Bayesian nonparametric point of view, and involving the Dirichlet mixture model, is time series data. There is a need, in the context of time series, for flexible models which can accommodate complex dynamics observed in real life data. While stationarity is a desirable property which facilitates estimation of relevant quantities, it is difficult to construct stationary models for which both the transition mechanism and the invariant density are sufficiently flexible. Many attempts have been made, often resulting in a compromise between flexibility and statistical properties. On the one hand, flexible transition mechanisms have been designed for which the resulting processes are not stationary; which may be purposefully achieved by the introduction of time as a covariate in a regression model (see e.g. Griffin and Steel, 2006, 2011; Zhu et al., 2005; Williamson et al., 2010), or may simply be a consequence of the model construction, for which stationarity conditions have not been established (see e.g. Müller et al., 1997; Tang and Ghosal, 2007). On the other hand, stationary models with a flexible transition mechanism have been proposed, for which the stationary density is restricted (Mena and Walker, 2005) or for which the construction is too complicated for efficient inference (Martínez-Ovando and Walker, 2011).

In this paper, we propose a model with nonparametric transition and stationary densities, which enjoys the advantages associated with stationarity, while retaining the necessary flexibility for both the transition and stationary densities. We demonstrate how posterior inference via MCMC can be carried out, focusing on the estimation of the transition density, both for stationary and non–stationary data–generating processes. For ease of exposition we only consider first order time series data and models, but the construction we propose can be adapted for higher order Markov dependence structures using multivariate normal kernels rather than one dimensional ones. Since the term non-stationarity is generally used in the literature to refer to general time varying processes, we clarify that our analysis is limited to time homogeneous processes for which the transition density does not vary with time.

At this point it is worth elaborating on the merits of having both the stationary and the transition densities represented as infinite mixture models. The idea is that no matter what new transition mechanism arises in reality, there are sufficient components within the model to absorb the changes. This clearly would not be true for a finite dimensional model. Hence, and illustration will bear this out, we believe that the combination of stationarity and infinite mixtures provides a powerful tool for modelling what might resemble highly irregular processes.

In Section 2 we describe the model and a latent variable extension is presented in Section 3. Section 4 provides a description of the MCMC method for posterior inference, for a particular choice of parametric mixture kernel. Finally, in Section 5 we present some examples for which inference is carried out, given a simulated sample from a known model. The first two examples involve the estimation of transition and stationary densities; the final two, the estimation of transition densities, when the stationary density does not exist.

2 Time series model

In this section we construct a flexible stationary model with nonparametric invariant and transition densities, i.e. both densities have an infinite mixture form. We begin with a simple normal first order stationary autoregressive model, AR(1).

Denote by

$$K_\theta(y,x) = \mathrm{N}_2\big((y,x)|(\mu,\mu),\Sigma\big),$$

the bivariate normal density with mean $\mu \in \mathbb{R}$ and covariance matrix

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

for some $-1 < \rho < 1$ and $\sigma^2 > 0$, making $\theta = (\mu, \rho, \sigma)$. Clearly,

$$K_\theta(y) = \int K_\theta(y|x)\, K_\theta(x)\, \mathrm{d}\, x = \mathrm{N}(y|\mu, \sigma^2),$$

so we can define a time homogeneous Markov process with stationary density $K_\theta(y)$, through the transition probability

$$\mathbb{P}\big(x_{n+1} \in A | x_n\big) = \int_A K_\theta(y|x_n) \mathrm{d}\, y; \quad A \in \mathcal{A},$$

with corresponding transition density given by the conditional

$$K_\theta(y|x) = \mathrm{N}\big(y|\mu + \rho(x-\mu), (1-\rho^2)\sigma^2\big).$$

Note that this corresponds to the normal AR(1) model, but the parametrization has been chosen to guarantee stationarity, without additional conditions on the parameters.

As with all simple parametric models, the dynamics of this process will be easily overwhelmed by real data. The aim of this paper is to construct a nonparametric version of the autoregressive model, by reproducing the construction of the transition mechanism as the conditional density for a given joint. We propose defining a nonparametric mixture directly over the bivariate density $K_\theta(y, x)$, therefore preserving the stationarity.

In the general case we take

$$f_P(y, x) = \int K_\theta(y, x)\, \mathrm{d}P(\theta).$$

In particular, this can be represented as

$$f_P(y, x) = \sum_{j=1}^{\infty} w_j\, K_{\theta_j}(y, x),$$

when P is a discrete probability measure given by

$$P = \sum_{j=1}^{\infty} w_j\, \delta_{\theta_j}.$$

Following the same principle used in the parametric case, we define the transition density as the conditional density

$$f_P(y|x) = \frac{\sum_{j=1}^{\infty} w_j\, K_{\theta_j}(y, x)}{\sum_{j=1}^{\infty} w_j\, K_{\theta_j}(x)}.$$

Then the transition probability measure is given by

$$\mathbb{P}\big(x_{n+1} \in A | x_n\big) = \int_A f_P(y|x_n) \mathrm{d}\, y; \quad A \in \mathcal{A},$$

which defines a first order time-homogeneous stationary process with invariant density

$$f_P(y) = \sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y).$$

The transition mechanism can be expressed as a nonparametric mixture of transition densities with dependent weights,

$$f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) \, K_{\theta_j}(y|x),$$

where

$$w_j(x) = \frac{w_j \, K_{\theta_j}(x)}{\sum_{j'=1}^{\infty} w_{j'} \, K_{\theta_{j'}}(x)}. \tag{1}$$

Therefore, we have constructed a model for which both the transition and the stationary densities are defined as nonparametric, i.e. infinite dimensional, mixtures.

So far, we have only defined what Martínez-Ovando and Walker (2011) refer to as a benchmark model. In the past, however, this model was considered to be intractable due to the infinite mixture appearing in the denominator of the dependent weight expression. A contribution in the present paper is a method, delineated in the following section, to overcome such intractability, therefore enabling posterior inference for the model.

3 Likelihood function and latent model

Let us consider a sample $\boldsymbol{x_n} = (x_0, \ldots, x_n)$. The likelihood function for the model is given by

$$f_P(\boldsymbol{x_n}) = f_P(x_0) \prod_{i=1}^{n} f_P(x_i|x_{i-1}) = f_P(x_0) \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j(x_{i-1}) \, K_{\theta_j}(x_i|x_{i-1}) \right), \tag{2}$$

where the dependent weights are given by expression (1) and the first observation is assumed to arise from the stationary density,

$$f_P(x_0) = \sum_{j=1}^{\infty} w_j \, K_{\theta_j}(x_0).$$

However, in order to simplify the notation, in the following we will consider, without loss of generality, the conditional likelihood

$$f_P(\boldsymbol{x_n}|x_0) = \prod_{i=1}^{n} f_P(x_i|x_{i-1}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j(x_{i-1}) \, K_{\theta_j}(x_i|x_{i-1}) \right), \tag{3}$$

thus assuming a fixed initial point $X_0 = x_0$.

Expression (3) is familiar in the context of nonparametric mixture models, and different methods for posterior inference for this type of likelihood model have been proposed. Such methods are usually divided into two families: the so called marginal methods rely on integrating out the random distribution function from the model, thus removing the infinite dimensional parameters (see e.g. Escobar, 1988; MacEachern and Müller, 1998; Neal, 2000); other methods work by sampling a finite but sufficient number of variables at each iteration of a Markov chain simulation scheme with the desired stationary distribution (see e.g. Muliere and Tardella, 1998; Ishwaran and Zarepour, 2000; Papaspiliopoulos and Roberts, 2008; Kalli et al., 2011). We will use the latter idea, as it is more convenient in the present case, in which we also have to deal with the intractable component in the denominator of (1). Accordingly, we introduce for each i an allocation variable $d_i \in \{1, 2, \ldots\}$, and use the latent model

$$\begin{aligned} f_P(\boldsymbol{x_n}, \boldsymbol{d_n}) =& \prod_{i=1}^{n} w_{d_i}(x_{i-1}) \, K_{\theta_{d_i}}(x_i|x_{i-1}) \\ =& \frac{\prod_{i=1}^{n} w_{d_i} \, K_{\theta_{d_i}}(x_{i-1}) \, K_{\theta_{d_i}}(x_i|x_{i-1})}{\prod_{i=1}^{n} \sum_{j=1}^{\infty} w_j \, K_{\theta_j}(x_{i-1})}. \end{aligned} \tag{4}$$

Once again, in order to illustrate the ideas, while keeping the notation simple, we consider a bivariate Gaussian kernel and mix over the mean and correlation coefficient, keeping the variance fixed across mixture components, so as to avoid overparametrization (see further comments in section 5). In other words, in what follows, we take

$$K_{\theta_j}(y|x) = \mathrm{N}\Big(y|\mu_j + \rho_j(x - \mu_j), (1 - {\rho_j}^2)\sigma^2\Big);$$

$$K_{\theta_j}(x) = \mathrm{N}(x|\mu_j, \sigma^2).$$

In this case, the denominator in (4) can be rewritten as

$$\sigma^{-n} \prod_{i=1}^{n} \left(\sum_{j=1}^{\infty} w_j \, \exp\Big\{-\tfrac{1}{2} \, (x_{i-1} - \mu_j)^2/\sigma^2\Big\} \right).$$

We now observe that the product terms are bounded by 1, and hence it is possible to use the identity

$$\sum_{k=0}^{\infty}(1-c)^k = c^{-1},$$

which holds for any $0 < c < 1$. Consequently, for each i, we can substitute

$$\left[\sum_{j=1}^{\infty} w_j \, \exp\Big\{-\tfrac{1}{2}\,(x_{i-1}-\mu_j)^2/\sigma^2\Big\}\right]^{-1}$$

with

$$\sum_{k_i=0}^{\infty}\left[1-\sum_{j=1}^{\infty} w_j \exp\Big\{-\tfrac{1}{2}\,(x_{i-1}-\mu_j)^2/\sigma^2\Big\}\right]^{k_i},$$

bringing the infinite sum from the denominator to the numerator of the likelihood expression. Furthermore, we may use the (k_i) as latent variables and introduce, for each i and $l = 1, \ldots, k_i$, new allocation variables, $z_{i,l} \in \{1, 2, \ldots\}$, in the same spirit of the d_i introduced before. Therefore, a latent expression for dealing with the denominator in (4) is given by

$$\sigma^n \prod_{i=1}^{n}\prod_{l=1}^{k_i} w_{z_{i,l}} \left[1-\exp\Big\{-\tfrac{1}{2}\,(x_{i-1}-\mu_{z_{l,i}})^2/\sigma^2\Big\}\right].$$

This combines with the numerator to give the full joint latent model,

$$f_P(\boldsymbol{x_n}, \boldsymbol{d_n}, \boldsymbol{k_n}, \boldsymbol{z_n}) = \sigma^n \prod_{i=1}^{n} w_{d_i}\, \mathrm{N}\Big((x_i, x_{i-1})|(\mu_{d_i}, \mu_{d_i}), \Sigma_{d_i}\Big)$$
$$\times \prod_{l=1}^{k_i} w_{z_{i,l}} \left[1-\exp\Big\{-\tfrac{1}{2}\,(x_{i-1}-\mu_{z_{i,l}})^2/\sigma^2\Big\}\right],$$

for which inference can be achieved via posterior simulation through the usual methods, as we will show in the following section. Notice that the subindex associated with the matrix Σ in the above expression refers to the component-wise correlation coefficient ρ_{d_i}.

It is easy to check that the original likelihood is recovered by summing over the latent variables $\boldsymbol{d_n} = \{d_i : i = 1, \ldots, n\}$, the $\boldsymbol{k_n} = \{k_i : i = 1 \ldots, n\}$ and the $\boldsymbol{z_n} = \{z_{i,l} : i = 1, \ldots, n;\ l = 1, \ldots, k_i\}$. Their introduction, therefore, does not alter the model, but makes posterior simulation for the $(\mu_j), (w_j), \sigma$ and ρ possible via MCMC. We refer the reader to the supplementary material for more details, including the particular case in which the mixing is done over the means only, while the complete corvariance structure, represented by ρ and σ^2, remains constant across components.

4 Posterior inference via MCMC.

The Bayesian model is completed by defining assigning a prior to P; effectively, over σ and the $(w_j, \mu_j, \rho_j)_{j=1}^{\infty}$.A typical choice is that of a stick–breaking process prior, i.e. for independently distributed Beta(a_j, b_j) variables, $(v_j)_{j=1}^{\infty}$, for some $a_j, b_j > 0$ (see Ishwaran and James, 2001), let

$$w_1 = v_1, \quad \text{and for } j > 1, \quad w_j = v_j \prod_{l<j}(1 - v_l).$$

In this paper, we focus on a Dirichlet Process prior, for which $a_j = 1$ and $b_j = b$.

For $\tau = \sigma^{-2}$ we use a gamma prior, and for each ρ_j, a discrete uniform prior on $R \subset (-1, +1)$, independently across j. The (μ_j) are taken independent and identically distributed from a base measure, which we choose to be a Normal distribution. With some care and reasonable restrictions on the priors, the results can be extended for component dependent variance (σ_j^2); however, this does not seem a good idea due to unidentificability issues introduced by the additional parameters (see the supplementary material for details).

Together with the joint latent model, the prior specification provides a joint density for all the variables which need to be sampled for a Monte Carlo based posterior estimation, i.e. the model parameters $\sigma, (w_j, \mu_j, \rho_j)_{j=1}^{\infty}$,and the latent variables $\left((z_{l,i})_{l=1}^{k_i},\, d_i,\, k_i\right)_{i=1}^{n}$.

There is still an issue due to the infinite state space for the $(z_{l,i}, d_i)$, corresponding to the infinite number of mixture components. One way to overcome this is to use the slice sampling technique of Kalli et al. (2011) (we refer the reader to the supplementary material for more details). However, as we have mentioned before, our model has to deal with the infinite mixtures of both the numerator and the denominator of the likelihood expression, which are represented by the two sets of indexing variables, (d_i) for the numerator $(z_{l,i})$ for the denominator. This increases the sensitivity of the simulation algorithm to the slice sampling parameters. Therefore, in the following, we will use a more stable algorithm based on an adequate random truncation (see Muliere and Tardella, 1998; Ishwaran and James, 2000, 2001).

As is well known, at each iteration of the MCMC, the $(w_j)_{j=1}^{\infty}$ can be calculated as $w_1 = v_1$ and $w_j = v_j \prod_{l<j}(1 - v_l)$ for $j > 1$. The (v_j) must be independently sampled from the full conditional distribution, which can easily be identified as

$$f(v_j|\cdots) = \text{Beta}(1 + n_j + N_j, b + n_j^+ + N_j^+),$$

where

$$n_j = \sum_i \mathbf{1}(d_i = j); \quad N_j = \sum_{i,l} \mathbf{1}(z_{i,l} = j); \quad n_j^+ = \sum_i \mathbf{1}(d_i > j); \quad N_j^+ = \sum_{i,l} \mathbf{1}(z_{i,l} > j).$$

Clearly, we cannot sample an infinite number of weights, therefore, we sample only $(w_j)_{j=1}^J$, for $J = \max(J_1, J_2)$, where $J_1 = \max\{z_{l,i}, d_i : i = 1, \dots, n; j = 1, \dots, k_i\}$ ensures the inclusion of all components active at the current iteration; and $J_2 = \min\{j : \prod_{l<j}(1 - v_l) < \epsilon\}$ for a sufficiently small $\epsilon > 0$, ensures that the sum of the J sampled weights is close enough to one. In other words, for $j > J$, w_j is almost zero and, therefore, the probability of sampling an index equal to such j at the current iteration is negligible.

A discrete prior for the component-wise correlation coefficient, ρ_j results in a discrete full conditional distribution, with

$$\mathbb{P}(\rho_j = r | \cdots) \propto \pi(r)(1 - r^2)^{-n_j/2} \exp\left\{ -\frac{\tau}{2} \sum_{d_i = j} \widehat{\mu}_i' \Sigma_r^{-1} \widehat{\mu}_i \right\},$$

where

$$\widehat{\mu}_i = \begin{pmatrix} x_i - \mu_{d_i} \\ x_{i-1} - \mu_{d_i} \end{pmatrix}, \quad \Sigma_r = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

for every $r \in R$. This can be sampled directly given a finite set R and a prior π over it.

The sampling of the (μ_j) is also not problematic. For each j, the prior for μ_j is a normal distribution, $\mathrm{N}(\mu_j | m, t^{-1})$, therefore the full conditional distribution can be written as

$$f(\mu_j | \cdots) \propto \mathrm{N}(\mu_j | m_j, t_j^{-1}) \prod_{i=1}^n \prod_{l=1}^{k_i} \left[1 - \exp\left\{ -\frac{(x_{i-1} - \mu_{z_{i,l}})^2}{2\sigma^2} \right\} \right]$$

where

$$m_j = \frac{1}{t_j} \left[mt + \tau \sum_{d_i = j} \frac{x_i + x_{i-1}}{1 + \rho_j} \right];$$

$$t_j = t + \frac{2n_j}{1 + \rho_j};$$

Since the product term in this full conditional is clearly bounded, we use a Metropolis-Hastings step, with proposal distribution given by the Gaussian factor in the distribution.

Before updating the σ it is convenient to introduce additional latent variables, $(u_{i,l})$, which allow us to substitute the term

$$\prod_{i=1}^n \prod_{l=1}^{k_i} \left[1 - \exp\left\{ -\frac{(x_{i-1} - \mu_{z_{i,l}})^2}{2\sigma^2} \right\} \right] \tag{5}$$

with a truncation term,

$$\prod_{i=1}^{n}\prod_{l=1}^{k_i}\mathbf{1}\left(u_{i,l} < 1 - \exp\left\{-\frac{(x_{i-1}-\mu_{z_{i,l}})^2}{2\sigma^2}\right\}\right).$$

Recall that $\tau = \sigma^{-2}$ is assigned a Gamma$(\tau|a,c)$ prior, which is conjugate for the precision of the Normal density kernel. Therefore, the full conditional for τ is a truncated Gamma,

$$f(\tau|\cdots) \propto \text{Gamma}(\tau|\widehat{a},\widehat{c})\mathbf{1}(\tau > T);$$
$$\widehat{a} = a + n/2;$$
$$\widehat{c} = c + \tfrac{1}{2}\sum_{i=1}^{n}\widehat{\mu}_i'\Sigma_{d_i}{}^{-1}\widehat{\mu}_i;$$
$$T = \max_{l,i}\left\{\frac{-2\log(1-u_{i,l})}{(x_i-\mu_{z_{i,l}})^2}\right\}.$$

Note that, since the variance is common for all mixture components, a small change in σ^2 may result in a small change in the product term (5), thus making the choice of the proposal distribution in a Metropolis-Hastings scheme inconvenient. Consequently, the use of the auxiliary variables $(u_{i,l})$ seems more effective in this case.

Finally, we need to describe how to update each k_i. Since the dimension of the sampling space changes with k_i, we use ideas for trans-dimensional MCMC developed in the context of model selection (Green, 1995; Godsill, 2001).

With probability $0 < p < 1$, we propose a move from k_i to $k_i + 1$ and accept it with probability

$$\min\left\{1, \frac{1-p}{p}\left[1-\exp\{-\frac{\tau}{2}(x_i-\mu_{z_{i,k_i+1}})^2\}\right]\right\}.$$

Clearly, the evaluation of this expression requires the sampling of the additional z_{i,k_i+1}. We take $z_{i,k_i+1} = j$ with probability w_j.

Whenever a move of this type is not propsed, and if $k_i > 0$, we accept a move to $k_i - 1$ with probability

$$\min\left\{1, \frac{p}{1-p}\left[1-\exp\{-\frac{\tau}{2}(x_i-\mu_{z_{i,k_i}})^2\}\right]^{-1}\right\}.$$

Thus, we have shown it is possible to perform posterior inference for the time series model. In the next section, we illustrate this in practice.

5 Illustrative examples.

In this section we present four examples, all of them involving simulated data, to illustrate the model presented in Section 2, and focusing on prediction. In the first example, data is simulated from the stationary model with a fixed known number of fully specified mixture components. In the second example, the data is generated by a stationary process which is not stated in terms of a nonparametric mixture, but in the form of a diffusion process. In these two examples, we use posterior simulation to recover the transition and stationary densities, the latter corresponding with the data histogram for a large enough sample. The excellent results for these two examples are not surprising.

However, that these excellent results also arise with data from a non–stationary model is of great interest. In the final two examples the data is generated from processes for which a stationary density does not exist. Nevertheless, both processes have fixed time homogeneous transition densities, and we are able to estimate them using the nonparametric stationary mixture model presented in this paper.

Therefore, the examples are chosen to illustrate how our model can be used for transition and invariant density estimation simultaneously, when the stationary density exists; yet remains suitable for transition density estimation, even when the data is not generated by a stationary process.

5.1 Example 1: Stationary mixture model

We generate a sample of size $n = 1000$ from the stationary mixture model described in Section 2, with three mixture components and true parameters $\mu_0 = (-1, 0, 3)'$, $w_0 = (0.1, 0.4, 0.5)'$, $\sigma_0^2 = 1$ and common $\rho_0 = 0.8$. Figure 1 (upper plot) shows the data, in blue, along with a heat plot for the true predictive density $f_0(x_i|x_{i-1})$, for $i = 1, \ldots, n$. We then perform posterior inference for both the stationary and the transition densities, via MCMC sampling, considering the full likelihood of expression (2). We use a Dirichlet process prior with mass parameter $b = 0.1$ for the mixing probability P. The base measure, defined in the previous section, requires the specification of some hyperparameters, and we take, for the μ_j, $m = \overline{x}_n$ and $t = 1/s_n$, the sample mean and precision respectively; $a = 1$, $c = 0.1$ for the $\tau = \sigma^{-2}$; and $R = \{0.001, \ldots, 0.999\}$ for the ρ_j.

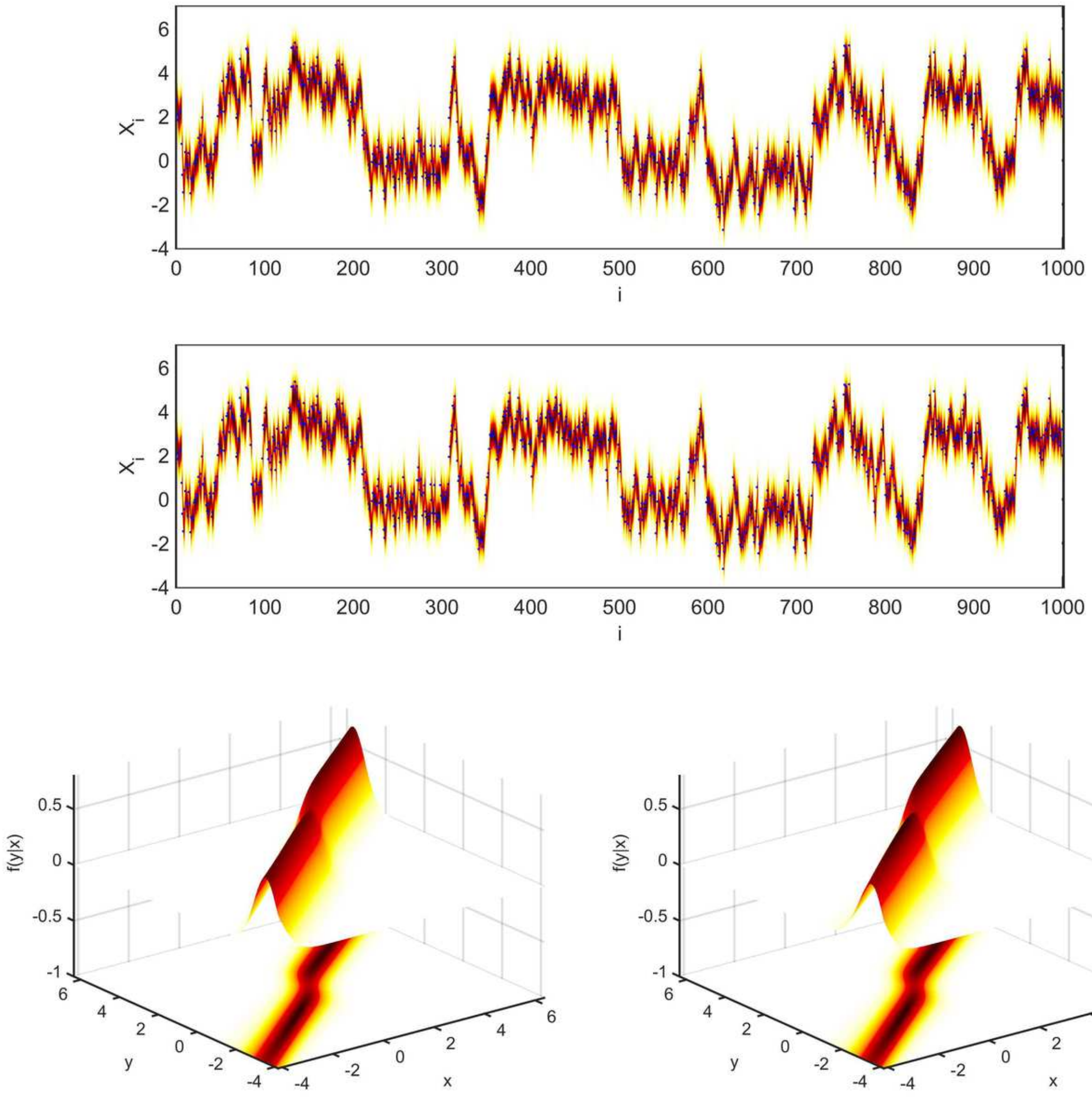

Figure 1: True (above) and estimated (middle) predictive densities for $n = 1000$ data points produced from the stationary mixture model with 3 mixture components; in both images, the blue dots represent the data. Below, the true (left) and estimated (right) transition density surfaces. The colors indicate the predictive density, with darker colors indicating higher density values.

Figure 1 shows, in the middle, a heat plot of the estimated predictive densities, which correspond to the Monte Carlo average of the posterior sample produced by the Markov Chain scheme for the latent model, evaluated at each data point. We use a Monte Carlo sample size of 2000 after a burn in period of 48000 iterations. By comparing the upper and middle plots in this figure, we conclude that the transition structure generating the data is indeed recovered by the estimation procedure.

Note that the estimate for the true transition density $f_0(y|x)$ is given by

$$f_n(y|x) = \int f(y|x) \mathrm{d}\Pi^n(f).$$

We therefore may obtain point-wise Monte Carlo estimates, for different values of x and y, thus producing a transition density surface, illustrated at the bottom of Figure 1. Once again, the similarity of the true surface (left plot) with the estimated one (right plot), suggests a success of the estimation mechanism.

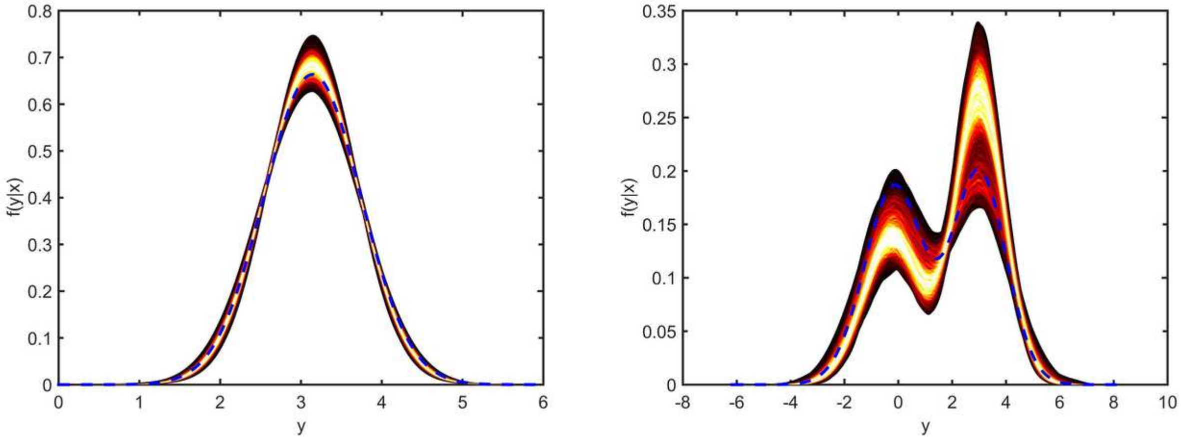

Figure 2: For the stationary mixture model with 3 mixture components, on the left, the transition density from the last data point $x = X_n$; on the right, the stationary density . The true curves are shown in blue and the color plot indicates point-wise posterior estimation, with darker color corresponding to a higher posterior probability.

From the predictive point of view, one may be interested in predicting, for instance, the next value, x_{n+1} in the data sequence. This can be done, through the predictive density $f_n(y|x_n)$. Clearly, from a Bayesian point of view, the estimation of such density would be incomplete without taking into account the complete estimated posterior. The left hand side of Figure 2 shows the true transition density given the last data point, i.e. $f_0(y|x_n)$, and a heat plot of the point-wise posterior estimates of the transition density. In other words, the colors illustrate the estimated posterior probability for each point of the predictive density; we can see that posterior estimation is highly concentrated around the true transition density, represented by the blue dashed line.

The right hand side of Figure 2 shows a similar plot, this time for the stationary density of the process. One more time, the true curve, $f_0(x)$ is shown in blue and the colors illustrate the point-wise posterior estimation. As might be expected, the transition density is recovered by the model better than the stationary density. This can be attributed to the fact that each new data point provides more information about the transition mechanism, while the information about the invariant measure is disturbed by the dependence between data points. However, given that the sample size is relatively small for this type of analysis, we believe the estimates to be satisfactory.

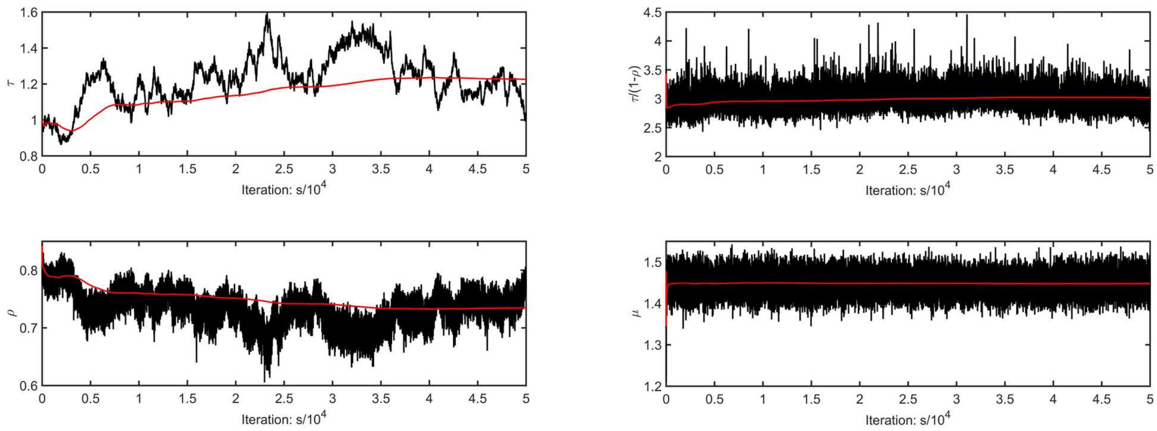

Figure 3: For the stationary mixture model with 3 mixture components, trace plot (in black) and moving average (in red) for: the common kernel precision τ (top left); the mean of the active kernel correlations vector $\sum \rho_{d_i}/n$ (bottom left) the monitoring transformation $\sum[\tau/n(1-\rho_{d_i}^2)]$ (top right); and the mean of the active kernel means vector $\sum \mu_{d_i}/n$ (bottom right).

In order to assess the convergence of the model, one may look directly at the trace plots of the model parameters. However, as is often the case for nonparametric mixture models, unidentificability issues complicate the interpretation of such plots. The left hand side of Figure 3 illustrate this issue. On the top, we can see the trace plot (in black) and moving average plot (in red) for the precision parameter $\tau = \sigma^{-2}$, while on the bottom we show a summary of the component-wise correlation coefficients, given by $\sum \rho_{d_i}/n$, which corresponds to the mean of the individual correlation coefficients of the components to which the data points are assigned at a given iteration, thus resulting in a weighted average, a common practice when monitoring convergence in the context of mixture models. In these plots, the interaction between the precision and correlation parameters is evident; even when the moving averages show signs of convergence, these plots are not reliable. Instead, we monitor a transformation of both conflicting parameters, which corresponds to the precision of the component-wise conditional density $f_j(y|x)$, given by $\tau/(1-\rho_j^2)$. Once again, we use the weighted average of such transformation over occupied components to construct the trace plot and moving averages represented on the top right hand side of Figure 3; on the bottom we see the analogous plot for the mean parameters μ_j. Both plots show lead us to believe that the burn-in period we are using is more than sufficient.

5.2 Example 2: Stationary diffusion

Here, we consider a diffusion process $X = \{X_t : t \geq 0\}$ defined as the solution to the stochastic differential equation (SDE)

$$\mathrm{d}X_t = \theta \frac{X_t}{\sqrt{1+X_t^2}} \mathrm{d}t + \mathrm{d}W_t,$$

which we refer to as the Hyperbolic diffusion. For $\theta < 0$, this is known to be a stationary process, with invariant density

$$f(x) \propto \exp\left\{2\theta\sqrt{1+x^2}\right\}.$$

The transition density, however, cannot be calculated explicitly (see e.g. Bibby and Sorensen, 1995, for more details).

A sample of size $n = 1000$ of observations, at times $t_i = i$, is generated using the exact simulation algorithm of Beskos et al. (2006), from the Hyperbolic diffusion with true parameter $\theta_0 = -2$. The top plot in Figure 4 shows the data, along with a heat plot of the true transition density $f_0(x_i|x_{i-1})$. Since an explicit form for the latter is not available, we replace it with a smoothed histogram of a sample generated, for each $i = 1, \ldots, n$, via exact simulation.

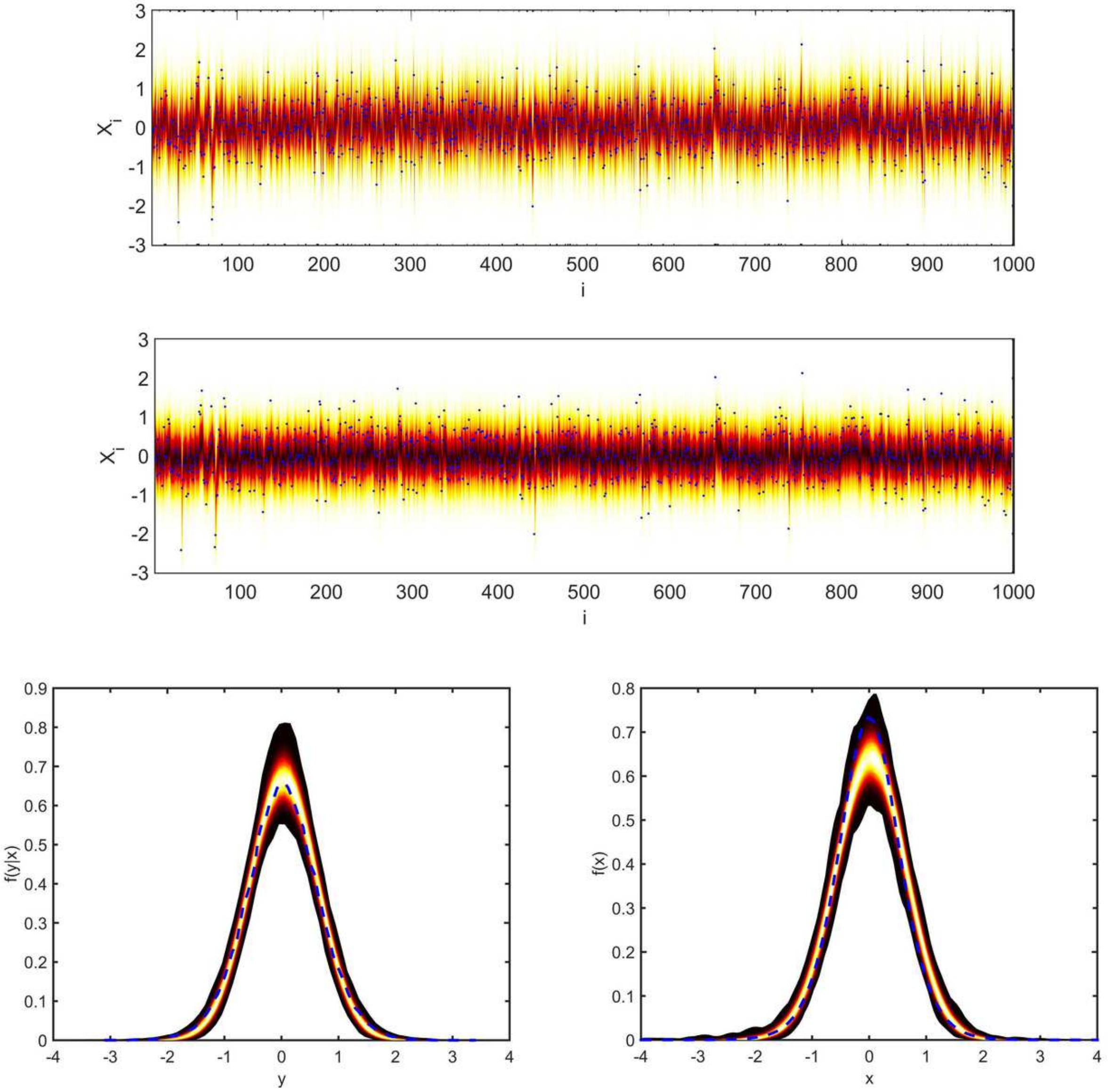

Figure 4: Posterior inference for $n = 1000$ data points produced from the hyperbolic diffusion with parameter $\theta = -2$. The top and middle images represent the true and estimated predictive densities: darker colors indicate higher density values; the blue dots represent the data. On the bottom, the left plot shows the estimated transition (solid line) and histogram of a true sample; the right plot shows the true stationary density (dashed line) and a color plot indicating point-wise posterior estimation, this time, lighter colors indicate higher posterior density.

The SDE provides a parametric description of the process. However, a nonparametric model should be flexible enough to recover the dynamics of the data generating mechanism. To illustrate this, we do posterior inference using the stationary time series mixture model described in this paper. We use the same hyperparameters as for the previous example. Posterior inference is, once again, carried out for the stationary and the transition densities, through posterior simulation for the latent model via MCMC, with a Monte Carlo sample size of 2000 after a burn in period of 48000 iterations. The middle plot in Figure 4 shows the estimated transition density at each data point. The right and left plots at the bottom of Figure 4 show the results. The estimated transition density (black line) is compared to a histogram (in blue) of a sample of size $2,000$ points generated from the true diffusion transition, via exact simulation. The normalizing constant for the true stationary density is calculated by numerical integration. Both the stationary and the transition densities can be seen to be recovered by the model.

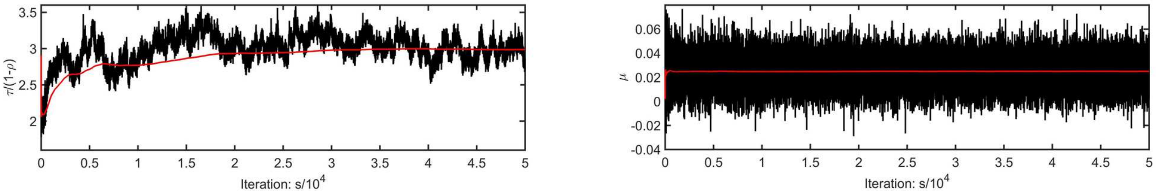

Figure 5: For the hyperbolic diffusion with parameter $\theta = -2$, trace plot (in black) and moving average (in red) for the monitoring transformation $\sum[\tau/n(1-\rho^2_{d_i})]$ (top) and the mean of the active kernel means vector $\sum \mu_{d_i}/n$ (bottom).

The data in this example was not generated from the model used for inference and, therefore, there is no reason to expect that the posterior distribution should concentrate around a fixed theoretically "true" value. Thus, the shape of the trace plot for the monitoring transformation $\tau/(1-\rho_j^2)$ on the left hand side of Figure 5 is not surprising, and convergence of the Markov Chain can be assessed by observing the moving average (in red). This further is confirmed by the corresponding plot for the mean parameter summary, on the right hand side of the figure.

5.3 Example 3: Standard Brownian motion

Standard Brownian motion is a typical example of a non stationary process. For discrete observations at times $t_i = i$, the transition density is known and given by $f(x_i|x_{i-1}) = \mathrm{N}(x_i|x_{i-1}, 1)$, the standard normal distribution centred at x_{i-1}.

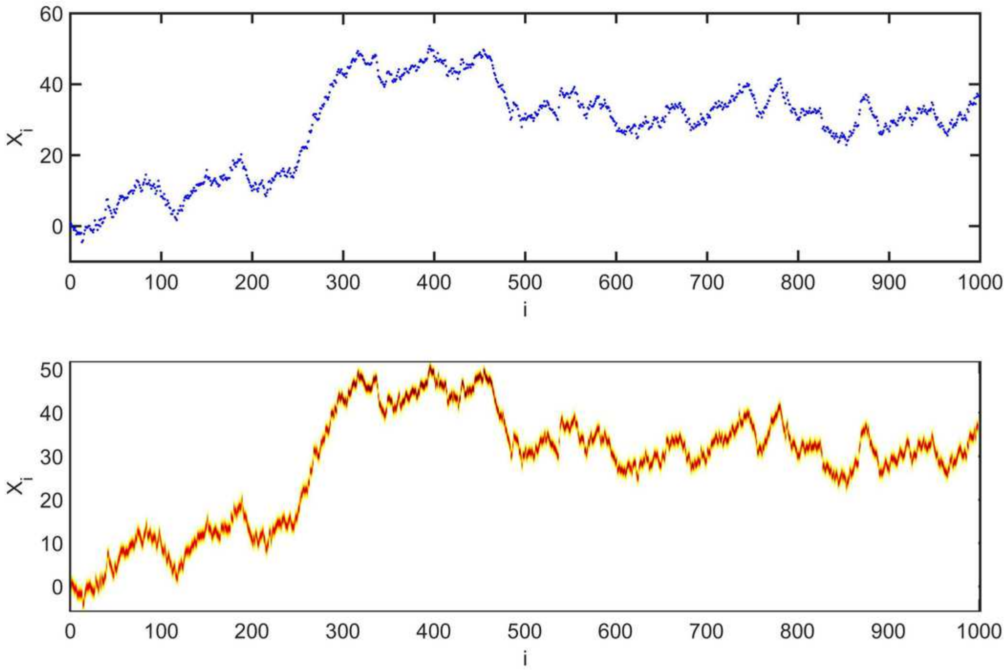

Figure 6: Data (above) and estimated transition densities(bottom) for $n = 1000$ data points produced from a standard Brownian motion and modelled via the stationary nonparametric mixture presented in this paper, with darker colors indicating higher density values.

Figure 6 (top) shows, in blue, a sample of $n = 1,000$ observations at times $t_i = i$ from a standard Brownian Motion path. Inference is carried out using a mixture over the mean, μ and the correlation coefficient ρ of the parametric components. We use an MCMC posterior sample size of 2000 after a burn-in period of 48000 iterations.

The mixture model we propose for time series is stationary. However, for any fixed sample size, it is flexible enough to capture the dynamics of the data, in the sense that we may use the model to estimate the transition density. The bottom plot of Figure 6 shows a heat plot of the MCMC posterior sample of the transition density $f(y|x)$ at each data point. Knowing that the true transition density is a normal density with unitary variance centered at the previous visited state, it is possible to verify the quality of the estimation. Furthermore, the plot on the left hand side of Figure 7, corresponds to the predictive density, i.e. the estimated conditional density given the last observation, $\mathbb{E}[f(x_{n+1}|x_n)|\boldsymbol{x_n}]$. The sample size is relatively small, for this type of problem, yet the model can recover the transition density. In this case, there is no stationary density to estimate. The model structure, for which only the conditional density appears in the likelihood, given $X_0 = 0$, guarantees that all of the information contained in the data is used to estimate the transition density.

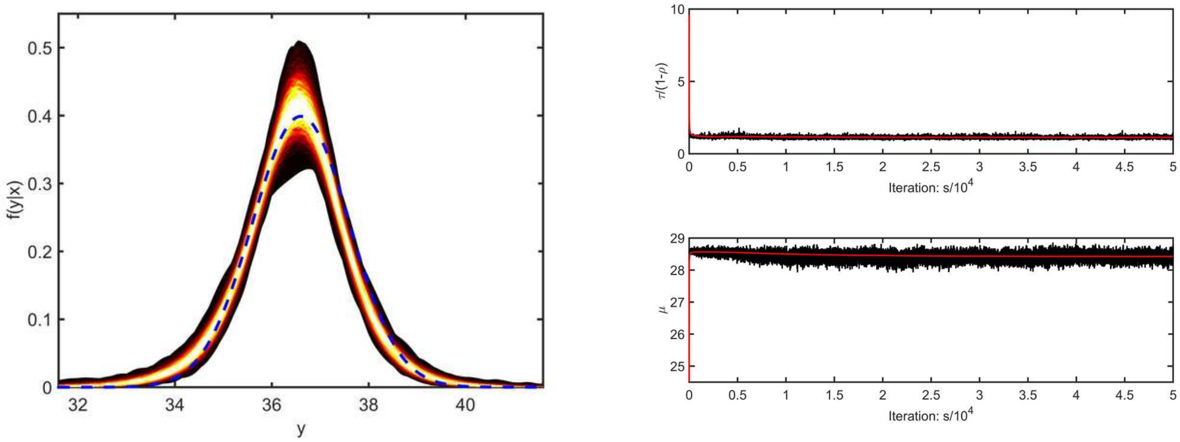

Figure 7: For $n = 1000$ data points produced from a standard Brownian motion and modeled via the stationary nonparametric mixture presented in this paper: on the left, the true transition density (blue) and a color plot indicating point-wise posterior estimation, with lighter color corresponding to a higher posterior probability; on the right trace plot (in black) and moving average (in red) for the monitoring transformation $\sum[\tau/n(1-\rho_{d_i}^2)]$ (top) and the mean of the active kernel means vector $\sum \mu_{d_i}/n$ (bottom).

The plots on the right of Figure 7 give clear indication of the convergence of the Markov Chain scheme for posterior simulation.

5.4 Example 4: Non-stationary diffusion

Finally, we consider a stochastic process $X = \{X_t : t \geq 0\}$ defined as a weak solution to the SDE

$$\mathrm{d}X_t = \sin(X_t - \theta)\mathrm{d}t + \mathrm{d}W_t.$$

A sample of size $n = 1000$ of observations, at times $t_i = i$, is generated using the exact simulation algorithm of Beskos et al. (2006), from this Sine diffusion, with true parameter $\theta_0 = 2$. The data can be seen at the top of Figure 8. Posterior inference is carried out for the transition density, through posterior simulation, via the MCMC algorithm for the latent model presented

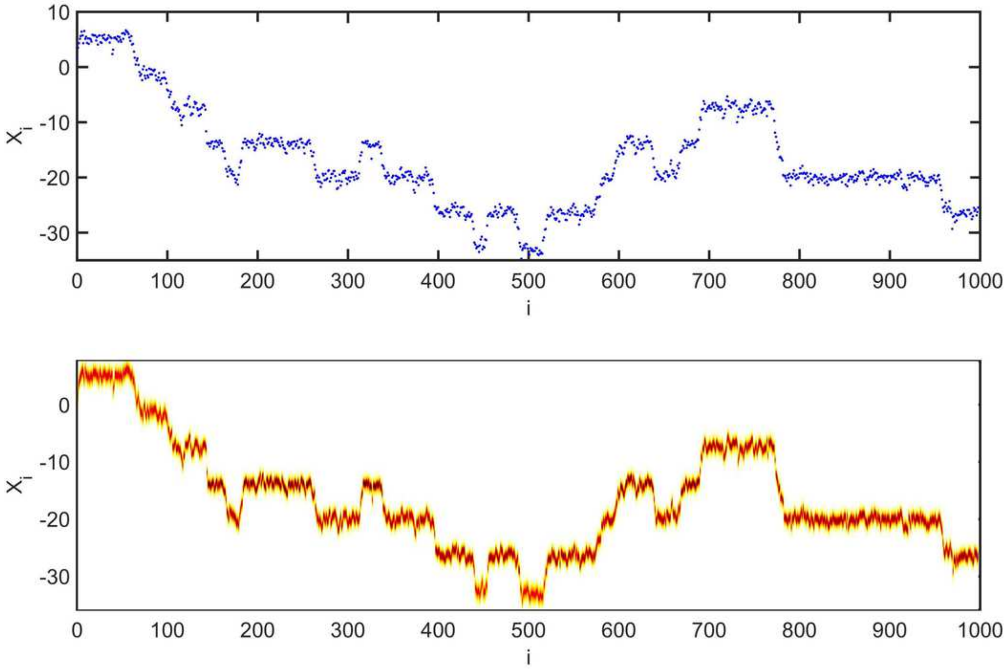

Figure 8: Data (above) and estimated transition densities(bottom) for $n = 1000$ data points produced from the sine diffusion with parameter $\theta_0 = 2$ and modelled via the stationary nonparametric mixture presented in this paper, with darker colors indicating higher density values.

in Section 2. Once again, the Monte Carlo sample size is 2000 after a burn in period of 48000 iterations with the same hyperparameters used for the previous examples. A color plot of the estimated transition density at each data point is shown at the bottom of Figure 8. There is no analytic expression available for the true transition density of this diffusion, however comparing with the data, we can appreciate that the dynamics of the process have been captured by the model.

Figure 9 shows the estimated transition density $f(y|x)$ $x = -20$ (left plot). The true transition density for this data is unavailable, but the estimate is compared against a smoothed histogram of a sample of size 10000, generated from the true model via exact simulation. Given the irregularity of the data, and the relatively small sample size, the heavier tails of the estimated density with respect to the exact simulated sample is justified. Overall, the transition density estimate can be considered accurate. In particular, the point $x = -20$ was chosen to illustrate how the model performance is improved for areas of the sample space frequently visited by the particular path represented by the data. On the right, we show the estimated m-step transition $f(x_{n+m}|x_n)$ which, as expected reflects the complex dynamics of the data-generating process.

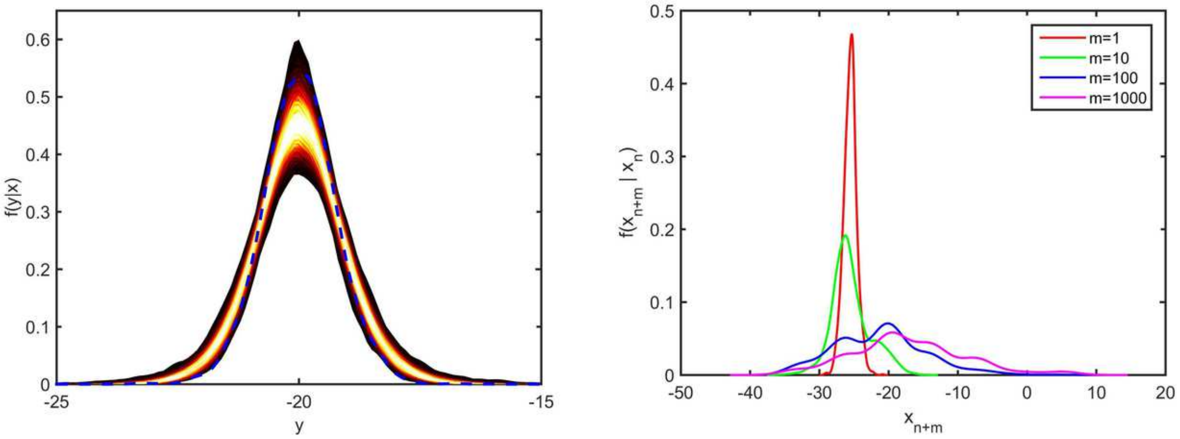

Figure 9: On the left, the predictive (transition) density, corresponding to an arbitrary point $x = -20$, for a sample of $n = 1000$ data points from the sine diffusion with parameter $\theta_0 = 2$; the color plot indicates point-wise posterior estimation, with lighter color corresponding to a higher posterior probability, while the blue like corresponds to a smoothed histogram of an exact sample from the diffusion. On the right, the estimated m-step transitions from the last data point $x_n = -25.43$.

As for previous examples, we present, in Figure 10, trace plots of key quantities supporting the convergence of the Markov Chain posterior simulation scheme within the burn-in period.

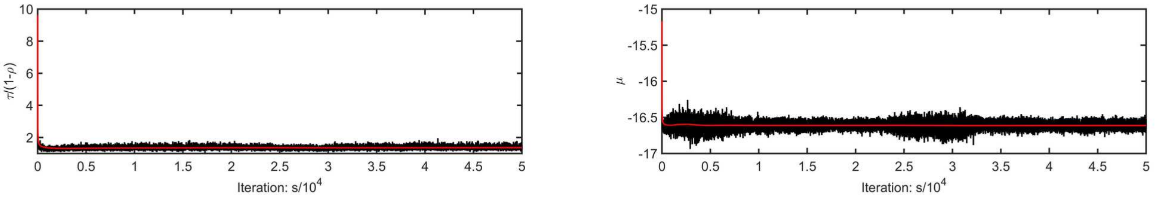

Figure 10: Trace plot (in black) and moving average (in red) for the monitoring transformation $\sum[\tau/n(1-\rho_{d_i}^2)]$ (left) and the mean of the active kernel means vector $\sum \mu_{d_i}/n$ (right).

6 Discussion

In summary, we have presented a stationary Markov model for which both the transition and stationary densities are nonparametric infinite mixture models. The construction is based on an infinite mixture of joint parametric kernels $k_\theta(y,x)$ for which both marginals are identical. The stationary density for the process is then given as the infinite mixture of such marginals, and the transition density is the corresponding conditional density, given by the ratio between the joint and the marginal mixtures. The infinite sum in the denominator can be seen as an intractable normalizing constant for the transition density. We have proposed a method for MCMC posterior inference based on the introduction of suitable latent variables.

We have illustrated the use of the stationary nonparametric model for posterior estimation of the transition and stationary densities when the data is generated by some true but arbitrary stationary process. In this case, a fixed true joint density for pairs of observations is available, and the model is able to recover it. At the same time, the stationary density is estimated and estimation of the transition density coincides with the ratio of the two.

When the data is generated by a non stationary but time homogeneous process, the model is still able to estimate the transition density, as we have empirically shown through some examples. In this case, there are no fixed marginal and joint densities to replicate, so the numerator and the denominator in the transition density expression do not have a direct interpretation. It is a known fact that a ratio can remain constant even when the numerator and denominator change. An analogous phenomenon explains the capacity of a stationary model to replicate a non stationary transition mechanism.

In order to have some intuition as to why, using a stationary model which is large enough, we may capture transition densities which arise from non-stationary models, let us write the transition mechanism from the stationary model as

$$f_p(y|x) = \frac{\sum_{j=1}^{\infty} w_j \, K_{\theta_j}(y,x)}{\sum_{j=1}^{\infty} w_j \, K_{\theta_j}(x)} = \frac{J}{M}.$$

The likelihood function is equal to the product of such transitions and thus, posterior estimation will recover the true transition density $f_0(y|x)$. If the data are coming from a stationary process then we know that such transition has associated bivariate and marginal densities $f_0(y,x)$ and $f_0(x)$. Thus can reasonably expect the posterior for J and the posterior for M to both converge to these true values, hence the capacity of the model to simultaneously estimate the stationary density of the process. On the other hand, if the data are not coming from a stationary model then clearly there are no associated bivariate and marginal densities to estimate and therefore, rather than forcing the convergence of the posterior of J and M separately, the shape of the likelihood function ensures that the posterior of J/M converges, thus correctly estimating the transition $f_0(y|x)$, provided the x falls within the range of the data. A formalization of this idea would likely require an account of the mass assigned to neighbourhoods of x by the empirical distribution.

We emphasize that the model we propose is intended for inference on the transition density $f(y|x)$ and, if existent, the stationary density $f(x)$. As is often the case with nonparametric mixture models, there are identifiability issues to consider and, therefore, the number of components and the values for the specific parameters are not interpretable. This was briefly discussed in Section 5, and more precisely, in Example 1, when choosing the quantities to be used for assessing the convergence of the MCMC procedure.

We have demonstrated the latent model construction and MCMC algorithm for a particular choice of parametric joint kernel, the bivariate Gaussian density. However, other kernel choices are available. Consider a measurable space $(\mathbb{X}, \mathcal{A})$ and denote by $K_\theta(y,x)$ any bivariate density on $\mathbb{X} \times \mathbb{X}$ with respect to some reference measure , for which the marginals are identical; i.e

$$K_\theta(y) = \int K_\theta(y,x)\,\mathrm{d}\,x \quad \text{and} \quad K_\theta(x) = \int K_\theta(y,x)\,\mathrm{d}\,y.$$

Clearly,

$$K_\theta(y) = \int K_\theta(y|x)\,K_\theta(x)\,\mathrm{d}\,x,$$

and therefore the model construction would still result in a flexible stationary time series.

Furthermore, the condition requiring both marginal densities to be equal is only needed to guaranty the stationarity of the mixture Markov model. Arbitrary joint kernels can be used to construct general autoregressive models if stationarity is not an issue. This includes the definition of multivariate time series models. An adequate choice of kernels in this case is not obvious and requires a careful study that goes beyond the scope of the present paper. Similarly, a nonparametric mixture is defined over multivariate kernels can be used to define a higher order Markov dependence structure. If the joint kernel includes $m+1$ random variables, an order m Markov transition can be defined as the ratio between the joint mixture over the m-variate marginal from which the $(m+1)$-th variable has been integrated out. Once again, a careful choice of the correlation structure in the multivariate kernels is paramount to the effectiveness of the model.

Acknowledgements

The authors are grateful for the careful reading and suggestions of two anonymous referees.

References

C. E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society*, 68(3):333–382, 2006.

B. M. Bibby and M. Sorensen. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1(1):17–39, 1995.

M. D. Escobar. *Estimating the means of several normal populations by nonparametric estimation of the distribution of the means.* PhD thesis, Department of Statistics, Yale University, 1988.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

S. J. Godsill. On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2001.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

J. E. Griffin and M. F. J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

J. E. Griffin and M. F. J. Steel. Stick-breaking autoregressive processes. *Journal of Econometrics*, 162(2):383–396, 2011.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian Nonparametrics.* Cambridge University Press, 2010.

H. Ishwaran and L. F. James. Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11:508–532, 2000.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

H. Ishwaran and M. Zarepour. Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21:93–105, 2011.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

S. N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

J. C. Martínez-Ovando and S. G. Walker. Time-series modelling, stationarity and Bayesian nonparametric methods. Technical report, Banco de México, 2011.

R. H. Mena and S. G. Walker. Stationary autoregressive models via a Bayesian nonparametric approach. *Journal of Time Series Analysis*, 26(6):789–805, 2005.

P. Muliere and L. Tardella. Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, 26:283–297, 1998.

P. Müller, M. West, and S. N. MacEachern. Bayesian models for non-linear auto-regressions. *Journal of Time Series Analysis*, 18:593–614, 1997.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Y. Tang and S. Ghosal. A consistent nonparametric Bayesian procedure for eestimating autoregressive conditional densities. *Computational Statistics and Data Analysis*, 51:4424–4437, 2007.

Sinead Williamson, Peter Orbanz, and Zoubin Ghahramani. Dependent Indian Buffet processes. *Journal of Machine Learning Research - Proceedings Track*, 9:924–931, 2010. URL `http://dblp.uni-trier.de/rec/bibtex/journals/jmlr/WilliamsonOG10`.

X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive Dirichlet process mixture models. Technical Report CMU-CALD-05-104, Carnegie Mellon University, Pittsburgh, 2005.