Stationary Mixture Transition Distribution (MTD) models via predictive distributions

Ramsés H. Mena and Stephen G. Walker
IIMAS–UNAM, México and University of Kent, UK
Received 29 Sep 2005 : final 31 March 2006

ABSTRACT. This paper combines two ideas to construct autoregressive processes of arbitrary order. The first idea is the construction of first order stationary processes described in Pitt et al. (2002) and the second idea is the construction of higher order processes described in Raftery (1985). The resulting models provide appealing alternatives to model non-linear and non-Gaussian times series.

Key words: AR model, Bayesian nonparametrics, MTD models, random probability measure, stationary process.

Running title: Stationary MTD models

MSC: 62M10, 62A15

1. Introduction

In Pitt *et al.* (2002), a flexible way of constructing *strictly stationary* autoregressive type AR(1) models with arbitrary marginal distributions was introduced. This article is concerned with the generalisation to higher order models and also to models allowing more flexible dependence structures.

The Pitt *et al.* (2002) constructions were based on a Gibbs sampler representation. Here we briefly review the idea: Suppose the required marginal density for the model is set to be $f_X(x)$. A conditional density $f_{Y|X}(y \mid x)$ is introduced and a well-defined transition density driving the AR(1)-type model $\{X_t\}$ can be defined as

$$p(x_{t-1}, x) = \int f_{X|Y}(x \mid y)\, f_{Y|X}(y \mid x_{t-1})\, \eta_1(dy) \tag{1}$$

where

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)\, f_X(x)}{\int f_{Y|X}(y \mid x)\, f_X(x)\, \eta_2(dx)},$$

and η_1 and η_2 are certain reference measures, in practice the Lebesgue or counting measure. It is easy to show that $f_X(\cdot)$ constitutes an invariant density for the transition (1); that is

$$f_X(x) = \int p(x_{t-1}, x) f_X(x_{t-1})\, \eta_2(dx_{t-1}). \tag{2}$$

This implies that the constructed first order AR(1)-type model with transition (1) defines a strictly stationary process with marginal density $f_X(\cdot)$. The associated random variable Y can be seen as a hidden or latent component in the model. Note that in order to avoid

complications, the domain of the "relevant parameter" in $f_{Y|X}(y \mid x)$ must coincide with the support of the required stationary distribution.

Given the required stationary distribution, the dependence in the model is imposed by the choice of $f_{Y|X}(y \mid x)$. Although, in general, such imposition might take many forms, Pitt *et al.* (2002) restrict this choice to those densities satisfying the linear dependence property given by

$$E[X_t \mid X_{t-1} = x] = \rho\, x + (1-\rho)\, \mu, \tag{3}$$

where $0 < \rho < 1$ and $\mu = E(X_t)$. This gives an autocorrelation function (ACF) with geometric decay, $\mathrm{Corr}(X_t, X_{t-h}) = \rho^h$.

It is worth pointing out that among the AR-type models available in the literature, many share the linear property mentioned above. A relatively recent review of these models can be found in Grundwald *et al.* (2000).

When studying time series models, an important issue is the dependence structure, namely the dynamics driving the process $\{X_t\}$. In the case of AR(1) models, such dependence is determined by the law of $\{X_t \mid X_{t-1}\}$, or $\{X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p}\}$ in the case of p-lagged models. In general, for non-linear and non-Gaussian models, complete specification of the densities corresponding to such conditional distributions is difficult to find and, when available, difficult to handle. From an application point of view, this difficulty has encouraged the use of restricted approaches to studying dependence structures in AR-type models. In particular, the modelling via Gaussian families has been popularised, and due to the second moment characterization of this distribution, the autocorrelation function has become the key measure of dependence.

One of the advantages of the approach undertaken by Pitt *et al.* (2002) is that, for the AR(1) case, equation (1) can be manipulated to lead to tractable expressions. When such a quantity has a closed form expression solving the underlying integral, then we can operate directly with the transition density. Alternatively, if the integral in (1) can not be simplified, the integral representation through the latent variable, Y, is useful and can be used, for example, to construct a "missing data" likelihood function.

For the sake of illustration, let us borrow one of the examples from Pitt *et al.* (2002). Suppose we want to construct an AR(1)-type model $\{X_t\}$ with a gamma stationary distribution, having density given by $\mathrm{Ga}(x; a, b) \propto x^{a-1}\, e^{-bx}\, I(x > 0)$, $a, b > 0$. Following the idea, we need to find a conditional density $f_{Y|X}(y \mid x)$. Set this choice to be the Poisson distribution with density given by $\mathrm{Po}(y; x\,\phi) \propto (x\,\phi)^y / y!\, I\{y = 0, 1, 2, \ldots\}$. Here $\phi := b\,\rho/(1-\rho)$ and $0 < \rho < 1$, so the ACF is given by ρ^h. The well known conjugacy property between these distributions leads to $f_{X|Y}(x \mid y) = \mathrm{Ga}(x; y + a; b + \phi)$.

In this case the linear expectation (3) is satisfied, since

$$\begin{aligned} E(X_t \mid X_{t-1} = x) &= E\{E(X_t \mid Y) \mid X_{t-1} = x\} \\ &= \frac{1-\rho}{b}\, \{a + E\,(Y \mid X_{t-1} = x)\} \\ &= \rho\, x + (1-\rho)\frac{a}{b} \end{aligned} \tag{4}$$

leading to $\mathrm{ACF}(h) = \rho^h$. Furthermore, following expression (1) with η_1 taken to be the counting measure, the one-step transition density of the constructed AR(1)-type model is

given by

$$
\begin{aligned}
p(x_{t-1}, x_t) &= \sum_{y=0}^{\infty} \mathrm{Ga}(x_t; y+a, b+\phi)\, \mathrm{Po}(y; x_{t-1}\,\phi) \\
&= \frac{\exp\left\{-\frac{b}{1-\rho}[x_t + \rho x_{t-1}]\right\}}{(1-\rho)\rho^{\frac{a-1}{2}}} \sqrt{\frac{x_t}{x_{t-1}}}^{a-1} \mathrm{I}_{a-1}\left(\frac{2\,b\,\rho^{1/2}\sqrt{x_t\ x_{t-1}}}{1-\rho}\right), \qquad (5)
\end{aligned}
$$

where $x, x_{t-1} > 0$ and $\mathrm{I}_\nu(\cdot)$ denotes the modified Bessel function of the first kind with index ν. See Abramowitz and Stegun (1992). The availability of the transition density (5) allows us to estimate the underlying parameters, a, b, ρ, using maximum likelihood estimation (MLE).

It is worth emphasizing that the specific family assumed for the conditional density $f_{Y|X}(y \mid x)$ might take different forms, leading to different forms of dependence while maintaining the same marginal distribution. Although, in general, the approach used by Pitt *et al.* (2002) can handle a wide variety of dependence forms, the structure of a given dataset might easily depend on higher lagged observations. This clearly implies the need for models with the ability to capture this feature.

2. Strictly stationary MTD models

In this section we use a technique introduced by Raftery (1985) to generalise the construction of Pitt *et al.* (2002). Raftery considered one-step p-lagged transition densities constructed by discrete mixtures of the type

$$
f(x \mid x_{t-1}, \ldots, x_{t-p}) = \sum_{k=1}^{p} w_k\, p_k(x_{t-k}, x), \qquad (6)
$$

for each t, where $w_k \geq 0$ and $\sum_{k=1}^{p} w_k = 1$. In general, p_k is allowed to change with the lagged value. In this section we will concentrate on the case where the only dependence is given through the lagged value, namely x_{t-k}. Note that this approach presupposes the knowledge of the first p values of the series, or one simply dismisses them for estimation purposes. Raftery's technique leads to *Mixture of Transition Distribution* (MTD) models. MTD models have been used in many applied areas, we refer to Berchtold and Raftery (2002) for a review on these models.

Although, in general, MTD models can be used with practically any transition densities, their estimation typically relies in some stationarity assumptions based on the particular choice of the parametric form assumed for the transitions $p(\cdot, \cdot)$. In Le *et al.* (1996) the Gaussian case was studied and conditions for weak stationarity were presented. In fact, these authors noted the difficulty in establishing conditions for strict stationarity. The typical relation between stationarity conditions and the availability of feasible estimation methods represent a limitation of the potential use of MTD models for applications in non-Gaussian settings. Furthermore, when dealing with series not supported on the real line, it is not always clear which form of transition density should be used for modelling purposes. Constructing the underlying transitions via the approach proposed by Pitt *et al.* (2002) tackles both issues of the strict stationarity and also provides an easy way of constructing parametric forms for the transition distributions.

Before defining the models resulting from the merging of these two approaches, first let us denote by $X^{[t,i]} := (X_{t-1}, \ldots, X_{t-i})$ the random lagged values down to lag i starting at time t. Similarly, also define observed points $x^{[t,i]}$.

Proposition 1. *A MTD model $\{X_t\}_{t\in\mathbb{N}}$ is strictly stationary with marginal density f_X when $X_1 \sim f_X$ and for all $t \geq 2$*

$$f\big(x_t \mid x^{[t,t_p]}\big) = \sum_{k=1}^{t_p-1} w_k\, p(x_{t-k}, x_t) + \left(1 - \sum_{k=1}^{t_p-1} w_k\right) p(x_{t-t_p}, x_t), \tag{7}$$

where $t_p := (t-1) \wedge p$, $i \wedge p := \min\{i, p\}$, $\sum_{k=1}^{p-1} w_k \leq 1$ and the transition densities $p(x_k, \cdot)$ are given by (1).

Proof. Without loss of generality, let us assume that $X_{t-1}, \ldots, X_{t-p}$ have marginal distribution with density given by $f_X(\cdot)$. Hence, we can treat the conditional densities with at least p lagged values. Defining $w_p := 1 - \sum_{k=1}^{p-1} w_k$ we have

$$f\big(x_t \mid x^{[t,p]}\big) = \sum_{k=1}^{p} w_k\, p(x_{t-k}, x_t). \tag{8}$$

Let $\mathcal{E}$ denote the support of $f_X(\cdot)$ and $\mathcal{Y}$ the support of the conditional density required for the transition density (1). In order to prove strict stationarity it is enough to verify

$$\begin{aligned}
&\int_{\mathcal{E}^p} f(x_t \mid x^{[t,p]})\, f_{X^{[t,p]}}(x^{[t,p]})\, \eta_2(dx_{t-1}) \cdots \eta_2(dx_{t-p}) \\
&= \sum_{k=1}^{p} w_k \int_{\mathcal{Y}} \int_{\mathcal{E}} f_{X|Y}(x_t \mid y_t)\, f_{Y|X}(y_t \mid x_{t-k})\, f_X(x_{t-k})\, \eta_2(dx_{t-k})\, \eta_1(dy_t) \\
&= \sum_{k=1}^{p} w_k \int_{\mathcal{Y}} f_{Y|X}(y_t \mid x_t)\, f_X(x_t)\, \eta_2(dy_t) \\
&= f_X(x_t).
\end{aligned} \tag{9}$$

The first equality comes from an application of Fubini's Theorem and from the hypothesis that $X_{t-1}, \ldots, X_{t-p}$ have marginal $f_X(\cdot)$. The second equality comes from the reversibility property underlying the construction.

The marginal densities corresponding to the blocks of dimension less than p can be verified to be also equal to f_X by using the same argument and applying equation (7). With this, any finite dimensional distribution can be fully specified retaining the same structure under time-shifts. Hence the result follows. □

The restriction on the weight of the second summand in (7) is done to ensure that the mixing proportions add to one even for those conditional densities with lagged values less than p.

In the context of the generalisation to high order models, the main advantage of the approach we are proposing, on combining the approaches of Pitt *et al.* (2002) and Raftery (1985), are the strict stationarity of the model and the integral representation of the transition mechanism driving the model. Proposition 1 uses the strict stationarity as a constructive feature rather than a property depending of some parameter values. Defining the transition distributions via the approach by Pitt *et al.* (2002) provides an appealing way to model non-linear and non-Gaussian dependence while retaining stationarity. Note that Proposition 1 also covers those conditional densities depending on lagged values less than p, providing a full characterisation of all finite dimensional distributions of $\{X_t\}_{t\in\mathbb{N}}$.

It is worth mentioning that the general approach of MTD models introduced by Raftery (1985) is not limited to stationary models, since it does not impose any condition on the marginal behavior, but rather only on the transition densities. In our approach, the transition

density required for the construction of the MTD model is built in a way that ensures the marginal distribution $f_X(\cdot)$ remains invariant.

In general, the approach introduced by Raftery (1985) is not always easy to handle, as shown in Le *et al.* (1996), where they confine themselves to Gaussian distributions. The main issue for non-Gaussian distributions in the general MTD framework is stationarity, which is not easily attained, even in a weak sense. This typically leads to estimation procedures which are complicated.

2.1. Correalation structure and higher moment properties

Constructing the transition densities through the approach described in Proposition 1, leads us to the following latent representation;

$$\begin{aligned} f(x \mid x^{[t,p]}) &= \sum_{k=1}^{p} w_k \int f_{X|Y}(x \mid y_t)\, f_{Y|X}(y_t \mid x_{t-k})\, \eta_1(dy_t) \\ &= \int f_{X|Y}(x \mid y_t)\, f(y_t \mid x^{[t,p]})\, \eta_1(dy_t), \end{aligned} \tag{10}$$

where

$$f(y_t \mid x^{[t,p]}) = \sum_{k=1}^{p} w_k\, f_{Y|X}(y_t \mid x_{t-k}). \tag{11}$$

Therefore, the latent structure enters as a typical finite mixture model. See McLachlan and Peel (2000).

Such a representation allows us to study all dependency properties of the stationary MTD model, even in cases where the transition densities are not known explicitly. This allows us to consider models with complex dependence structures.

As previously mentioned, a feature of the models in Pitt *et al.* (2002) is the linear expectation property (3). In the construction at issue, this property also extends to the p-order case. We have

$$E(X_t \mid X^{[t,p]}) = \sum_{k=1}^{p} w_k \left\{\rho\, X_{t-k} + (1-\rho)\,\mu\right\} = \rho \left(\sum_{k=1}^{p} w_k\, X_{t-k} \right) + (1-\rho)\,\mu. \tag{12}$$

Some other quantities of interest can be obtained while keeping this linear property. For example, if $p = 2$, we have

$$f(x_t \mid x_{t-1}, x_{t-2}) = w\, p(x_{t-1}, x_t) + (1-w)\, p(x_{t-2}, x_t)$$

and

$$E(X_t \mid X_{t-1}, X_{t-2}) = \rho\left\{w\, x_{t-1} + (1-w)\, x_{t-2}\right\} + (1-\rho)\,\mu.$$

Let us use the notation $E_x(\cdot) = E(\cdot \mid X_1 = x)$. Then the lagged predictive moments are given by

$$\begin{aligned} E_x(X_2) &= \rho\, x + (1-\rho)\,\mu \\ E_x(X_3) &= E_x[E(X_3 \mid X_2, X_1)] = E_x\left[\rho\,(w\, X_2 + (1-w)\, X_1) + (1-\rho)\,\mu\right] \\ &= \rho\, w\, E_x(X_2) + \rho\,(1-w)\, x + (1-\rho)\,\mu \\ E_x(X_4) &= E_x[E(X_4 \mid X_3, X_2)] = \rho\, w\, E_x(X_3) + \rho\,(1-w)\, E_x(X_2) + (1-\rho)\,\mu \\ &\vdots \\ E_x(X_h) &= \rho\, w\, E_x(X_{h-1}) + \rho\,(1-w)\, E_x(X_{h-2}) + (1-\rho)\,\mu, \end{aligned} \tag{13}$$

where $\mu = E(X)$, assuming it exists, denotes the mean of the given stationary distribution. If we further assume that $\mu_{(2)} := E(X^2)$ also exists, then we can compute the following difference equation for the autocovariance,

$$\begin{aligned}\gamma(h) &= E(X_{h+1}X_1) - \mu^2 = E[X_1\, E(X_{h+1} \mid X_1)] - \mu^2. \\ &= \rho\, w\, E(X_1 X_h) + \rho\,(1-w)\, E(X_1 X_{h-1}) + (1-\rho)\,\mu^2 - \mu^2 \\ &= \rho\, w\, \gamma(h-1) + \rho\,(1-w)\,\gamma(h-2). \end{aligned} \tag{14}$$

In terms of the autocorrelation function we have,

$$r(h) = \frac{\gamma(h)}{\mu_{(2)} - \mu^2} = \rho\,[w\, r(h-1) + (1-w)\, r(h-2)]\,, \quad h \geq 2, \tag{15}$$

with initial conditions

$$r(0) = 1 \quad \text{and} \quad r(1) = \rho. \tag{16}$$

The general solution to this difference equation is given by

$$r(h) = B_1\, z_1^h + B_2\, z_2^h, \tag{17}$$

where z_1, z_2 are the solutions to the quadratic equation given by $z^2 - \rho\, w\, z - \rho\,(1-w) = 0$, and B_1, B_2 are determined with the initial conditions. Therefore, the ACF can be determined through

$$r(h) = \frac{1}{2^{h+1}} \left\{ \left[1 + \frac{2\,\rho - \rho\, w}{A} \right] [\rho\, w + A]^h + \left[1 - \frac{2\,\rho - \rho\, w}{A} \right] [\rho\, w - A]^h \right\}, \tag{18}$$

where $A = \sqrt{\rho^2\, w^2 + 4\,\rho - 4\,\rho\, w}$.

In general, p-lagged stationary MTD models with linear property (12) satisfy the p-order difference equation given by

$$r(h) = \rho \sum_{k=1}^{p} w_k\, r(h-k), \quad h \geq p.$$

Notice that, for the above example, the ACF does not depend on the choice of stationary distribution. Typically, tools for analysing dependency in time series models are based on second order moments. However, higher moments can be crucial in the analysis of dependence, this is the case of the well known ARCH model (Engle (1982)).

As we mentioned before, Grundwald *et al.* (2000) noted that most non-Gaussian stationary models available in the literature are limited to the simple linear dependence (3). The representation (1) for the transition density in the construction by Pitt *et al.* (2002) not only provides us with means of constructing models with more complex dependence structures but also to study higher moments. For example, in the gamma-Poisson AR(1) model (5), of Section 1, we have that for $s, l \geq 1$

$$\begin{aligned} E[X_t^s\, X_{t-1}^l] &= (b+\phi)^{-s} \sum_{y=0}^{\infty} \frac{\Gamma(y+s+a)}{\Gamma(y+a)}\, \frac{\phi^y}{y!}\, \frac{b^a}{(\phi+b)^{y+l+a}}\, \frac{\Gamma(y+l+a)}{\Gamma(a)} \\ &= \frac{(1-\rho)^{a+s+l}\, \Gamma(s+a)\, \Gamma(l+a)}{b^{\,s+l}\, \Gamma(a)^2}\; {}_2F_1(s+a, l+a; a; \rho), \end{aligned} \tag{19}$$

where ${}_2F_1([\cdot,\cdot];\,\cdot\,;\,\cdot\,)$ denotes Gauss's hypergeometric function. See Abramowitz and Stegun (1992). Using the result (19), it can be seen that, in particular, the 1-lag ACF corresponding to the squared process $\{X_t^2\}$ is given by

$$\mathrm{Corr}(X_t^2, X_{t-1}^2) = \frac{(\rho + 2 + 2a)\rho}{2a + 3}. \tag{20}$$

For this cross moment, the dependence also includes the parameters contained in the stationary distribution; a difference from the row ACF.

From a simulation perspective, we have the following mechanism for moving to X_t from $\{X_{t-1}, \ldots, X_{t-p}\}$. We take

$$X_{\delta_t} = X_{t-k} \text{ with probability } w_k, \;\; k = 1, \ldots, p$$

then take

$$Y_t \sim f_{Y|X}(y_t|x_{\delta_t})$$

and

$$X_t \sim f_{X|Y}(x_t|y_t).$$

A complete data likelihood based on the complete data set $\{X_t, Y_t, \delta_t\}$ is available for inference purposes.

3. Dependence structure via random measures

We have seen in Section 1 that the approach of Pitt *et al.* (2002) requires the specification of a parametric distribution with density $f_{Y|X}(y \mid x)$. Essentially, any parametric conditional distribution is allowed. For instance, in the gamma-Poisson example presented in the introduction we could have chosen $f_{Y|X}(y \mid x) = \mathrm{Ga}(y; x, 1)$ instead of $\mathrm{Po}(y; x\,\phi)$, which also leads to a stationary model with marginal distributions $\mathrm{Ga}(a, b)$. This distribution determines the dependence structure driving the model, and therefore plays an important role in the underlying construction. Specific choices can be of interest when some intuition about the latent variable is available. However, in many situations this is not the case.

In order to overcome this issue, Mena and Walker (2005) resorted to randomize the choice of such a distribution by using ideas from Bayesian nonparametric methods. The idea is simple, if we look closer at the construction of Pitt *et al.* (2002), the distribution $f_{Y|X}(y \mid x)$ can be seen as the posterior distribution corresponding to a prior distribution on the latent variable Y, and hence $p(x_{t-1}, x_t)$ becomes the posterior predictive distribution. For a nonparametric assumption about Y, it seems reasonable to assume that such a component is itself a random distribution function.

An example, which appears in Pitt *et al.* (2002), can be obtained using the predictive distribution of the Dirichlet process; see Ferguson (1973). Specifically, instead of considering $f_{Y|X}(y \mid x)$ given in a parametric form, let us consider the posterior distribution, based on one-observation, corresponding to the Dirichlet process,

$$F \mid X \sim \mathscr{D}(c\,G + \delta_X),$$

where $c > 0$, $\mathscr{D}(\mu)$ denotes the Dirichlet process with driving measure μ, $G := E_{\mathscr{D}}[F]$ denotes the distribution function of the required stationary distribution and δ_x is the Dirac measure. The predictive distribution function corresponding to such a process is given by

$$P[X_t \leq x_t \mid X_{t-1} = x_{t-1}] = \frac{c\,G(x_t) + \delta_{x_{t-1}}((-\infty, x_t])}{c + 1}. \tag{21}$$

In this nonparametric setting, F is now equivalent to the latent component Y. A high order analog of this model can be constructed using the MTD idea described in the previous section. The idea involves taking

$$X_t|F_t \sim F_t$$

with

$$F_t \mid X_{t-1}, \ldots, X_{t-p} \sim \sum_{k=1}^{p} w_k \, \mathscr{D}\left(cG + \delta_{X_{t-k}}\right). \tag{22}$$

Following Proposition 1, it is easy to show that $\{X_t\}$ is stationary with marginal distribution function G. That is,

$$\mathrm{E}\left\{ \sum_{k=t-p}^{t-1} w_k \frac{cG(x) + \mathbf{1}(X_k \leq x)}{c+1} \right\}$$

is easily seen to be $G(x)$. Based on this result we can actually see that

$$X_t|X_{t-1}, \ldots, X_{t-p} \begin{cases} \sim G & \text{with probability } c/(c+1) \\ \\ = X_{t-k} & \text{with probability } w_k/(c+1). \end{cases}$$

It is worth mentioning that the stationary MTD via Dirichlet predictive distributions, resembles the DAR models, first introduced by Jacobs and Lewis (1978) under a different construction.

If one-data based predictive distributions, resulting from a non-parametric Bayesian scheme, take a tractable form, the underlying construction is also valid using other random measures. For example, we can use all random distributions resulting from normalisation of increasing additive processes, see Regazzini *et al.* (2003), Lijoi *et al.* (2005a,b). In this spirit normalised log-Gaussian processes were used, in Mena and Walker (2005) to build AR(1)-type models.

4. Estimation

The approach we are proposing uses Bayes theorem to construct the predictive distribution; it is then used as the transition kernel in the construction of MTD models. Rather than being used as an inference procedure, Bayes theorem is used in the construction of the model. In order to emphasise this point, we have chosen to estimate the model using non-Bayesian procedures. However, adopting a Bayesian approach could be done through a simple modification of the analysis presented in Diebolt and Robert (1994).

In Le *et al.* (1996), estimation of parameters and weights is done by the *expectation maximisation* (EM) algorithm. Their method consists in conditioning on a latent variable Z_t, taking values $1, \ldots, p$, with probabilities $w_1, \ldots, w_p$. Typically, it is assumed that

$$\Pr(Z_t = z_t) = {w_1}^{z_{1t}} {w_2}^{z_{2t}} \cdots {w_p}^{z_{pt}}.$$

Hence, given $Z_t = (0, \ldots, 1, \ldots, 0)$, that is one in the k-th entry, density (8) reduces to $p(x_{t-k}, x_t)$, for $k = 1, \ldots, p$. For a review of this method we refer to McLachlan and Peel (2000). Here we adapt it to our model. In practice, it is convenient to work with a p-dimensional latent vector, Z_{kt} defined to be one or zero according to whether the lagged value is k or not.

Given a sample $x = (x_1, \ldots, x_n)$, $n > p$, the logarithm of the augmented data likelihood can easily seen to be equal to

$$\log L_{x,z}(\theta) = \sum_{k=1}^{p} \sum_{t=1}^{n} z_{kt} [\log w_k + \log p(x_{t-k}, x_t; \theta)], \tag{23}$$

where θ denotes all the parameters in the model. Note that a difference from Le *et al.* (1996) is that we are considering the stationary part, the distribution of the first p values, in the likelihood.

Hence, given an initial, or current, value of parameters $\theta^{(i)}$ and $w_k^{(i)}$ for $k = 1, \ldots, p$, the E-Step involves takng the expectation of (23) with respect to the conditional distribution of $\mathbf{Z} \mid \mathbf{X}$, which reduces to

$$\tau_k\big(x_t, \theta^{(i)}\big) := E[Z_{kt} \mid \mathbf{x}] = \frac{w_k^{(i)}\, p\big(x_{t-k}, x_t; \theta^{(i)}\big)}{f\big(x_t \mid x^{[t,p]}; \theta^{(i)}\big)}.$$

The M-Step involves maximising

$$\theta^{(i+1)} = \max_{\theta} Q(\theta \mid \theta^{(i)}), \tag{24}$$

where

$$Q(\theta \mid \theta^{(i)}) = \sum_{k=1}^{p} \sum_{t=1}^{n} \tau_k\big(x_t, \theta^{(i)}\big) \, [\log w_k + \log p(x_{t-k}, x_t; \theta)]. \tag{25}$$

The weights are updated via

$$w_k^{(i+1)} = \frac{\sum_{t=1}^{n} \tau_k\big(x_t, \theta^{(i)}\big)}{n - k}, \quad \text{for } k = 1, \ldots, p. \tag{26}$$

So far we have not used the fact that the one-step transition component can be also be decomposed through a latent variable Y. As we mentioned before, the parametric setting of Section 2 provides us with representation (1) for the transition $p(x_{t-k}, x_t)$. However, such an integral might not lead to a tractable expression or may even not be available explicitly. In such cases, it can be of interest to consider the complete data likelihood, also including the latent variables $Y = (Y_2, \ldots, Y_n)$. This, leads to consider the complete data log-likelihood given by

$$\log L_{x,z,y}(\theta) = \log f_X(x_1; \theta) + \sum_{k=1}^{p} \sum_{t=1}^{n} z_{kt} \{\log w_k + \log[f_{X|Y}(x_t \mid y_t; \theta)\, f_{Y|X}(y_t \mid x_{t-k}; \theta)]\}. \tag{27}$$

Hence, the EM algorithm will involve taking the expectation of (27) with respect to $Z, Y \mid X$. If we consider the decomposition of such a distribution as

$$\text{Law}\{Y, Z \mid X\} = \text{Law}\{Y \mid X, Z\} \text{Law}\{Z \mid X\}$$

then the only additional requirement is to consider the component-wise distribution

$$f(y_{kt} \mid x_t, x_{t-k}) \propto f_{X|Y}(x_t \mid y_{kt})\, f_{Y|X}(y_t \mid x_{t-k}). \tag{28}$$

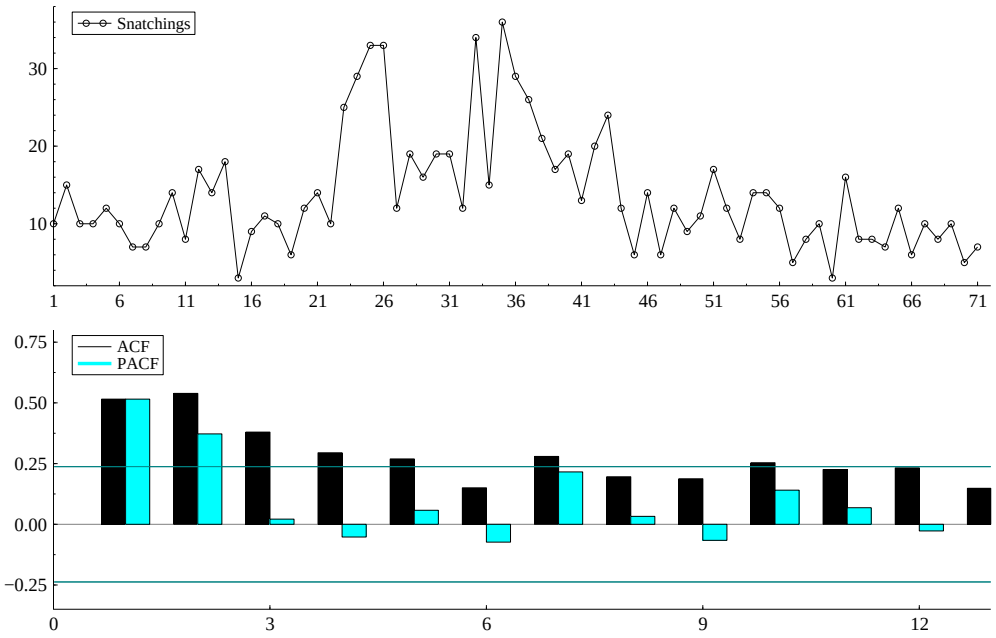

Figure 1: Number of Hyde Park purse snatchings in Chicago within every 28 day periods; Jan'69 - Sep '73. The upper plot shows the data series and the bottom one the autocorrelation and partial autocorrelation function.

5. Illustration

For our illustration we have selected a real data set consisting of the number of Hyde Park purse snatchings in Chicago within every 28 day periods; Jan'69 - Sep '73. The data source is McCleary and Hay (1980).

From the nature of the data, shown in Figure 1, we can infer that the support is the set of non-negative reals. Hence, a reasonable choice for the stationary distribution could be the gamma distribution with mean a/b. The data also exhibit high autocorrelation, at least back to lag five. It also shows certain recurrences, therefore justifying the use of a stationary model. With this observations it is reasonable to assume the Stationary MTD model with transition probabilities given by (5).

We apply the EM algorithm described in Section 4. The maximisation step was performed through the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm. See Press *et al.* (1992) for more on this algorithm. Table 5 presents the results. From these results we can infer that the lagged values X_{t-1}, X_{t-2}, X_{t-3} and X_{t-7} are significant, agreeing with the results observed in the ACF in Figure 1. The histogram corresponding the the snatchings data together with the fitted stationary distribution are shown in Figure 2.

Notice that, for this data set, it is also possible to fit a classical AR(p) model by preliminary transformations to (second order) stationarity, e.g. by removing mean or differentiating. However, doing this would not consider any of the higher moment properties underlying the MTD model.

Iter.	a	b	ρ	w_1	w_2	w_3	w_4	w_5	w_6	w_7
20	4.45	0.34	0.674	0.387	0.294	0.0733	0.0452	0.00717	0.00945	0.184
40	4.48	0.343	0.659	0.401	0.345	0.0444	0.0167	0.000765	0.00114	0.191
60	4.48	0.344	0.654	0.406	0.363	0.0301	0.00681	8.34E-05	0.000141	0.193
80	4.48	0.344	0.651	0.41	0.371	0.0218	0.00289	9.11E-06	1.78E-05	0.194
100	4.48	0.344	0.65	0.412	0.375	0.0165	0.00125	9.98E-07	2.26E-06	0.195
120	4.48	0.344	0.65	0.414	0.378	0.0128	0.000543	1.09E-07	2.88E-07	0.195
140	4.48	0.344	0.649	0.415	0.379	0.0102	0.000237	1.20E-08	3.67E-08	0.195
160	4.48	0.344	0.649	0.416	0.38	0.00819	0.000103	1.32E-09	4.71E-09	0.196
180	4.48	0.344	0.649	0.416	0.381	0.00667	4.52E-05	1.45E-10	6.04E-10	0.196
200	4.48	0.344	0.648	0.417	0.382	0.00548	1.98E-05	1.59E-11	7.76E-11	0.196
300	4.48	0.344	0.648	0.418	0.383	0.00222	3.16E-07	2.56E-16	2.76E-15	0.196
400	4.49	0.344	0.648	0.419	0.384	0.000961	5.04E-09	4.16E-21	9.89E-20	0.196
500	4.49	0.344	0.648	0.419	0.384	0.000428	8.05E-11	6.75E-26	3.56E-24	0.197

Table 1: Results for the EM estimation algorithm for a stationary MTD-AR(2) model applied to the snatchings data set.

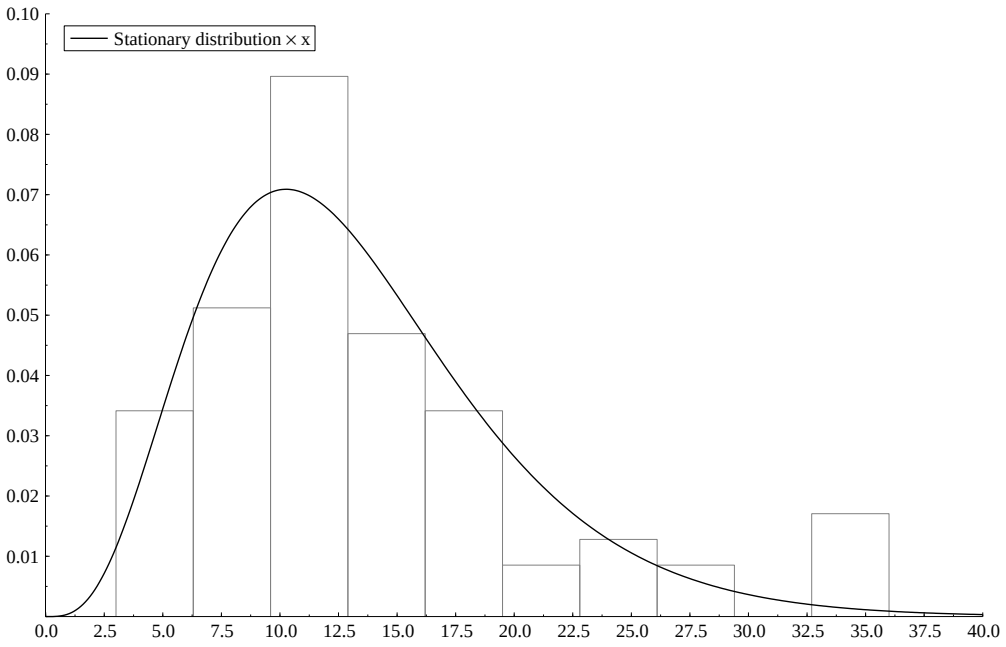

Figure 2: Fitted stationary distribution for the number of Hyde Park purse snatchings in Chicago.

Acknowledgments

The authors are grateful for the comments of two referees. The research of the second author was partially supported by an EPSRC Advanced Research Fellowship.

References

Abramowitz, M. and Stegun, I. A., editors (1992). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications Inc., New York. Reprint of the 1972 edition.

Berchtold, A. and Raftery, A. (2002). The Mixture Transition Distribution (MTD) model for high-order Markov chains and non-Gaussian time series. *Statist. Sci.*, **17**, 328–356.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc. B.*, **56**, 363–375.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica.*, **50**(4), 987–1008.

Grundwald, G. K., Hyndman, R. J., Tedesco, L., and Tweedie, R. L. (2000). Non-gaussian conditional linear AR(1) models. *Aust. N. Z. J. Stat.*, **42**, 479–495.

Jacobs, P. A. and Lewis, P. A. W. (1978). Discrete time series generated by mixtures. III: Autoregressive processes (DAR(p)). Technical Report NPS55-78-022, Naval Posgraduate School, Monterey, California.

Le, N. D., Martin, R. D., and Raftery, A. E. (1996). Modeling flat stretches, burst and outliers in time series using mixture transition distribution models. *J. Amer. Stat. Assoc.*, **91**, 1504–1515.

Lijoi, A., Mena, R. H., and Prünster, I. (2005a). Bayesian nonparametric analysis for a generalized dirichlet process prior. *Stat. Inference Stoch. Process.*, **8**, 283–309.

Lijoi, A., Mena, R. H., and Prünster, I. (2005b). Hierarchical mixture modelling with normalized inverse gaussian priors. *J. Amer. Stat. Assoc.*, **100**, 1278–1291.

McCleary, R. and Hay, R. (1980). *Applied Time Series Analysis for the Social Sciences*. SAGE Publications.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley.

Mena, R. H. and Walker, S. G. (2005). Stationary autoregresive models via a bayesian nonparametric approach. *To appear in the Journal of Time Series Analysis*.

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002). Constructing first order autoregressive models via latent processes. *Scand. J. Statist.*, **29**, 657–663.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge University Press.

Raftery, A. E. (1985). A model for high order Markov chains. *J. R. Statist. Soc. B.*, **47**, 528–539.

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.*

R. H. Mena. Departamento de probabilidad y estadística, IIMAS-UNAM, A.P. 20-726, Mexico, D.F. 01000. Mexico. Tel. +52 (55) 56223583 Ext. 3542. Fax +52 (55) 56223621.
Email: ramses@sigma.iimas.unam.mx