Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

On Construction and Estimation of Stationary Mixture Transition Distribution Models

Xiaotian Zheng, Athanasios Kottas & Bruno Sansó

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 00, NO. 0, 1–11
https://doi.org/10.1080/10618600.2021.1981342

Taylor & Francis
Taylor & Francis Group

Check for updates

On Construction and Estimation of Stationary Mixture Transition Distribution Models

Xiaotian Zheng, Athanasios Kottas, and Bruno Sansó

Department of Statistics, University of California, Santa Cruz, CA

ABSTRACT
Mixture transition distribution (MTD) time series models build high-order dependence through a weighted combination of first-order transition densities for each one of a specified number of lags. We present a framework to construct stationary MTD models that extend beyond linear, Gaussian dynamics. We study conditions for first-order strict stationarity which allow for different constructions with either continuous or discrete families for the first-order transition densities given a prespecified family for the marginal density, and with general forms for the resulting conditional expectations. Inference and prediction are developed under the Bayesian framework with particular emphasis on flexible, structured priors for the mixture weights. Model properties are investigated both analytically and through synthetic data examples. Finally, Poisson and Lomax examples are illustrated through real data applications. Supplementary files for this article are available online.

ARTICLE HISTORY
Received October 2020
Revised June 2021

KEYWORDS
Bayesian inference;
First-order strict stationarity;
Markov chain Monte Carlo;
Non-Gaussian time series

1. Introduction

Mixture transition distribution (MTD) models describe a time series $\{X_t : t \in \mathbb{N}\}$, where $X_t \in \mathcal{S} \subseteq \mathbb{R}$ for all t, by specifying the distribution of X_t conditional on the past as follows:

$$F(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l \, F_l(x_t|x_{t-l}), \tag{1}$$

for $t > L$, based on initial values for $(x_1, \ldots, x_L)^\top$. In Equation (1), $F(x_t \mid \boldsymbol{x}^{t-1})$ is the conditional cumulative distribution function (cdf) of X_t given that $\boldsymbol{X}^{t-1} = \boldsymbol{x}^{t-1}$, and $F_l(x_t \mid x_{t-l})$ is the conditional cdf of X_t with respect to the lth transition component given that $X_{t-l} = x_{t-l}$, where $\boldsymbol{X}^{t-1} = \{X_i : i \le t-1\}$ and $\boldsymbol{x}^{t-1} = \{x_i : i \le t-1\}$. The parameters $w_l \ge 0$, $l = 1, \ldots, L$, assign weights to the transition components, such that $\sum_{l=1}^{L} w_l = 1$. On a finite state space this model provides a parsimonious approximation of high-order Markov chains (Raftery 1985; Raftery and Tavaré 1994; Berchtold 2001). On a more general space, the model structure can represent time series that depict non-Gaussian features such as burst, outliers, and flat stretches (Le, Martin, and Raftery 1996), or change-points (Raftery 1994). We refer to Berchtold and Raftery (2002) for a review. An MTD model consists of L first-order transition components. The mixture autoregressive model of Wong and Li (2000) is a generalization that allows for each transition component to depend on a different number of lags; Lau and So (2008) consider a Bayesian nonparametric prior for the transition component of such models. There are several related extensions that consider mixtures of autoregressive conditional heteroscedastic terms, including Wong and Li (2001b), Berchtold (2003), Zhu, Li, and Wang (2010), and Li et al. (2017). Other extensions include multivariate model settings (Hassan and Lii 2006; Fong et al. 2007; Kalliovirta, Meitz, and Saikkonen 2016), time-varying mixture weights (Wong and Li 2001a; Bartolucci and Farcomeni 2010; Bolano and Berchtold 2016), nonlinear transition dynamics (Heiner and Kottas 2021a), and order/lag selection (Khalili, Chen, and Stephens 2017; Heiner and Kottas 2021b). Applications of these models appear in many fields such as finance, and the environmental and medical sciences; see, for example, MacDonald and Zucchini (1997); Lanne and Saikkonen (2003); Escarela, Mena, and Castillo-Morales (2006); Cervone et al. (2014).

Stationarity for MTD models, and their extensions, is generally difficult to attain due to the mixture model structure. This limits the choices of parametric families for the transition components for these models. Families considered in the literature include: Gaussian (Le, Martin, and Raftery 1996; Wong and Li 2000; Kalliovirta, Meitz, and Saikkonen 2015); Student-t (Wong, Chan, and Kam 2009; Meitz, Preve, and Saikkonen 2021); Laplace (Nguyen et al. 2016); Weibull (Luo and Qiu 2009); and Poisson (Zhu, Li, and Wang 2010). These models are typically parameterized in ways that result in conditional expectations that are linear functions of the lags. This particular parameterization facilitates the study of stationarity, though only in a weak sense, at the cost of reducing model flexibility. Indeed, the conditional expectation of an MTD model has the general form $\sum_{l=1}^{L} w_l \, \mu_l(x_{t-l})$, where $\mu_l(y) = \int x dF_l(x \mid y)$, allowing for nonlinear dependence of the mean, conditional on past observations.

The primary goal of this article is to develop conditions for first-order strictly stationary MTD models, that is, stationary models with an invariant marginal distribution. We show that a sufficient condition is to assume the same marginal distribution for all the components of the mixture. It turns out that this marginal distribution is also the invariant marginal distribution

CONTACT Xiaotian Zheng xzheng24@ucsc.edu Department of Statistics, University of California, Santa Cruz, CA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

of the time series. Under this condition, the first-order strict stationarity is achieved with respect to any particular parameterization. We thus obtain a rich class of distribution specifications for the model, facilitating the study of component distributions that have not been explored in the literature, and enhancing the modeler's ability to extend beyond high-order linear dependence in the conditional expectation. Although the focus of our methodology is on strictly stationary models, we also study weak stationarity conditions for MTD models with linear conditional expectation.

MTD models are usually built by specifying transition densities $f_{U_l|V_l}$ for each component $l = 1, \ldots, L$. These correspond to conditional densities for random variable U_l given random variable V_l. This specification raises a question of existence of a coherent bivariate density f_{U_l,V_l}. Our second goal is to provide a constructive approach to building MTD models that satisfy our strict stationarity condition under a coherent bivariate density f_{U_l,V_l}. We present two distinct approaches: the bivariate distribution method, which is based on specifying the bivariate distribution of the pair $(U_l, V_l), l = 1, \ldots, L$; and the conditional distribution method, which consists of finding pairs of compatible conditional distributions $f_{U_l|V_l}$ and $f_{V_l|U_l}$ for all (U_l, V_l).

Our final goal is to develop a Bayesian framework for MTD model inference and prediction. We assume that the order of dependence is unknown, but is bounded above by a finite number L. We use an over-specified model with L chosen conservatively, under the expectation that only a few of the lags contribute to the dynamics of the series. We consider two priors for the mixture weights, one based on a truncated stick-breaking process, and the other obtained by discretization of a cdf which is assigned a nonparametric prior. While the former supports stochastically decreasing weights, the latter favors important, but not necessarily consecutive weights.

The rest of the article is organized as follows. In Section 2, we review the issues related to establishing stationarity conditions for MTD models. We then introduce the invariant condition that yields the class of first-order strictly stationary MTD models, and connect it to weak stationarity. Section 3 illustrates two methods to construct such models with many examples. In Section 4, we outline the Bayesian approach for model estimation and prediction, followed in Section 5 by an illustration of the properties of two structured priors for mixture weights on synthetic data, and applications of the models on two real data sets of different nature. Finally, we conclude with a discussion in Section 6. Proofs and details of Markov chain Monte Carlo (MCMC) algorithms are provided in the appendix and the supplementary material.

2. First-Order Strict Stationarity

Consider the conditional density specification of the model in Equation (1)

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l f_l(x_t|x_{t-l}). \tag{2}$$

Under our modeling framework, each transition component is taken to correspond to the distribution for a random vector (U_l, V_l), for $l = 1, ..., L$, where $f_l \equiv f_{U_l|V_l}$ denotes the associated conditional density.

Earlier work has studied necessary and sufficient conditions for constant first and second moments (Le, Martin, and Raftery 1996). In general, such conditions are difficult to establish, especially for the second moment $\int_{\mathcal{S}} x_t^2 g_t(x_t) dx_t$, where $g_t(x_t) = \sum_{l=1}^{L} w_l \int_{\mathcal{S}} f_l(x_t \mid x_{t-l}) g_{t-l}(x_{t-l}) dx_{t-l}$ is the marginal density of the process $\{X_t\}$. This restricts the choices of parametric families for the component transition densities. In particular, those choices result in linear conditional expectations. Even when conditions for time-independent first and second moments can be obtained, the resulting constrained parameter spaces complicate estimation.

The key result for our methodology is given in the following proposition, the proof of which can be found in the Appendix. The result provides the foundation for different constructions of the first-order strictly stationary MTD models. Rather than imposing restrictions on the parameter space, the proposition formulates a substantially easier to implement condition on the marginals of the bivariate distributions that define the transition components.

Proposition 1. Consider a set of bivariate random vectors (U_l, V_l) taking values in $\mathcal{S} \times \mathcal{S}$, $\mathcal{S} \subseteq \mathbb{R}$, with conditional densities $f_{U_l|V_l}$, $f_{V_l|U_l}$ and marginal densities f_{U_l}, f_{V_l}, for $l = 1, \ldots, L$, and let $w_l \geq 0$, for $l = 1, ..., L$, with $\sum_{l=1}^{L} w_l = 1$. Consider a time series $\{X_t : t \in \mathbb{N}\}$, where $X_t \in \mathcal{S}$, generated from

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l f_{U_l|V_l}(x_t|x_{t-l}), \quad t > L, \tag{3}$$

and from

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{t-2} w_l f_{U_l|V_l}(x_t|x_{t-l}) + \left(1 - \sum_{k=1}^{t-2} w_k\right) f_{U_{t-1}|V_{t-1}}(x_t|x_1), \quad 2 \leq t \leq L.$$

This time series is first-order strictly stationary with invariant marginal density f_X if it satisfies the invariant condition: $X_1 \sim f_X$, and $f_X(x) = f_{U_l}(x) = f_{V_l}(x)$, for all $x \in \mathcal{S}$, and for all l.

The two different expressions for the transition density allow us to establish the stationarity condition for the entire time series. The relevant form for inference is the one in Equation (3), since we work with the likelihood conditional on the first L time series observations. Proposition 1 applies regardless of X_t being a continuous, discrete or mixed random variable.

Regarding strict stationarity, the literature mostly focuses on existence of a stationary distribution. Exceptions are Kalliovirta, Meitz, and Saikkonen (2015) and Meitz, Preve, and Saikkonen (2021), where a stationary marginal distribution for a mixture autoregressive model is obtained, albeit again under constrained parameter spaces, and Mena and Walker (2007) whose approach is the one most closely related to our proposed methods.

Mena and Walker (2007) used the latent variable method proposed in Pitt, Chatfield, and Walker (2002) to construct the conditional density for each transition component of the MTD. More specifically, $f_l(x_t \mid x_{t-l}) = \int h_{X|Z}(x_t \mid z) h_{Z|X}(z \mid x_{t-l}) dz$, where $h_{X|Z}(x \mid z) \propto h_{Z|X}(z \mid x) f_X(x)$, and the integral is replaced by a sum if Z is a discrete variable. Then, provided $X_1 \sim f_X$, the MTD model is first-order strictly stationary with invariant density f_X. Under this construction, the invariant density f_X can be viewed as the prior for likelihood $h_{Z|X}$, which is built through latent variable Z. In practice, this restricts the approach to continuous time series, and the choices for the invariant density to cases where f_X is conjugate to $h_{Z|X}$. Even for such cases, the transition component will typically have a complex form. In particular, the example explored in Mena and Walker (2007) involves a gamma invariant distribution, with $h_{Z|X}$ corresponding to a Poisson distribution. In this case, $f_l(x_t \mid x_{t-l})$ is a countable sum whose evaluation requires modified Bessel functions of the first kind. Moreover, following Pitt, Chatfield, and Walker (2002), Mena and Walker (2007) restricted attention to choices of $h_{Z|X}$ that yield linear conditional expectations for the transition components, and thus also for the MTD models.

The key feature of our approach is that it builds from the bivariate distributions, f_{U_l,V_l}, corresponding to the transition components. In the next section, we discuss two approaches to specifying those bivariate distributions, either directly or via compatible conditionals, $f_{U_l|V_l}$ and $f_{V_l|U_l}$. In conjunction with Proposition 1, we obtain a general framework to constructing first-order strictly stationary MTD models that can be applied to both discrete and continuous time series, while allowing for a wide variety of invariant marginal distributions, as well as for both linear and nonlinear lag dependence in the conditional expectation.

In general, an explicit expression for the autocorrelation function for general MTD models is difficult to derive. However, a recursive equation can be obtained for a class of linear MTD models. We say the MTD model is linear if $E(U_l|V_l = y) = a_l + b_l y$ for some $a_l, b_l \in \mathbb{R}$, $l = 1, \ldots, L$. Consider a linear MTD model that satisfies the invariant condition of Proposition 1, and assume that the first and second moments of the invariant marginal distribution, denoted by μ and $\mu^{(2)}$, exist and are finite. Then, for any L and $h \geq L$, we can derive

$$E(X_{t+h}X_t) = \textstyle\sum_{l=1}^L w_l a_l \mu + \sum_{l=1}^L w_l b_l E(X_{t+h-l}X_t). \quad (4)$$

Assuming that, for any $h \geq 1$, $E(X_{t+h}X_t)$ does not depend on time t, let $r(h)$ be the lag-h autocorrelation function. Then,

$$r(h) = \phi + \sum_{l=1}^{L} w_l b_l r(h-l), \quad h \geq L, \quad (5)$$

where $\phi = (\sum_{l=1}^L w_l a_l \mu - (1 - \sum_{l=1}^L w_l b_l)\mu^2)/(\mu^{(2)} - \mu^2)$ is zero if and only if $\mu = 0$ or $a_l = (1 - b_l)\mu$. When $b_l = \rho$, $\rho \in (0, 1)$ and $a_l = (1 - \rho)\mu$, for all l, Equation (5) reduces to $r(h) = \rho \sum_{l=1}^L w_l r(h-l)$, $h \geq L$, which is the result in Mena and Walker (2007).

In the case of distinct roots, the general solution to Equation (5) is

$$r(h) = c_1 z_1^h + \cdots + c_L z_L^h + \phi\,((1 - z_1) \ldots (1 - z_L))^{-1}, \quad (6)$$

where $c_1, \ldots, c_L$ are determined by the initial conditions $r(0), \ldots, r(L-1)$ and $z_1, \ldots, z_L$ are the roots of the associated polynomial $z^L - w_1 b_1 z^{L-1} - \cdots - w_L b_L = 0$. It follows that, as $h \to \infty$, $r(h) \to 0$ if and only if: (1) $\phi = 0$; (2) $z_1, \ldots, z_L$ all lie inside the unit circle.

The above discussion provides an approach to obtaining a weakly stationary MTD model based on Equation (3), and is summarized in the following proposition the proof of which is included in the supplementary material.

Proposition 2. The time series defined in Equation (3) is weakly stationary if: (i) the invariant condition of Proposition 1 is satisfied with a stationary marginal for which the first two moments exist and are finite; (ii) the conditional expectation with respect to $f_{U_l|V_l}$ is $E(U_l \mid V_l = y) = a_l + b_l y$, for some $a_l, b_l \in \mathbb{R}$, and for all l; (iii) Equation (4) is independent of time t, and the roots of the equation $z^L - w_1 b_1 z^{L-1} - \cdots - w_L b_L = 0$ all lie inside the unit circle.

Proposition 2 illustrates the construction of a weakly stationary MTD model building from the invariant condition of Proposition 1. We focus on the first-order strictly stationary MTD models. Weak stationarity can be further studied if conditions (2) and (3) of Proposition 2 are satisfied.

3. Construction of the First-Order Strictly Stationary MTD Models

Here, we present two methods to develop first-order strictly stationary MTD models. The bivariate distribution method constructs the transition density given a specific marginal distribution. This method may result in analytically intractable transition densities. The second method, consisting of directly specifying the transition component conditional densities, has estimation advantages, although the analytical form of the marginal density may not be readily available. Thus, the selection among these methods depends on the modeling objectives. In fact, there are special cases where both the transition and marginal densities belong to the same family of distributions.

3.1. Bivariate Distribution Method

Under this method, we seek bivariate distributions f_{U_l,V_l} whose marginals f_{U_l} and f_{V_l} are equal to a given f_X, for $l = 1, \ldots, L$. Consequently, the lth transition component density is $f_{U_l|V_l}(u \mid v) = f_{U_l,V_l}(u, v)/f_X(v)$. In contrast to the approach in Mena and Walker (2007), which is practical when the marginal density is a conjugate prior for some likelihood, the bivariate distribution method is applicable to essentially any discrete or continuous marginal invariant density f_X. In fact, for most parametric families, there is a rich literature defining collections of bivariate distributions with a desired marginal distribution, and allowing for a variety of dependence structures. The following examples illustrate the method.

Example 1: Gaussian and continuous mixtures of Gaussians MTD models. Under marginal $f_X(x) = N(x \mid \mu, \sigma^2)$, the Gaussian MTD model can be constructed via the bivariate Gaussian distribution for (U_l, V_l), with mean $(\mu, \mu)^\top$ and covariance matrix $\Sigma = \sigma^2 \left(\begin{smallmatrix} 1 & \rho_l \\ \rho_l & 1 \end{smallmatrix}\right)$, resulting in a Gaussian density for $f_{U_l|V_l}$.

In particular,

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l N\left(x_t|(1-\rho_l)\mu + \rho_l x_{t-l}, \sigma^2(1-\rho_l^2)\right). \tag{7}$$

Let $t(x\,|\,\mu,\sigma,\nu) \propto \left(1+\nu^{-1}((x-\mu)/\sigma)^2\right)^{-(\nu+1)/2}$ denote the Student-t density, where μ,σ and ν are respectively location, scale and tail parameters. To construct as a natural extension of the Gaussian MTD model a stationary Student-t MTD model, consider the bivariate Student-t distribution, which can be defined as a scale mixture of a bivariate Gaussian with mean $(\mu,\mu)^\top$ and covariance matrix $q\Sigma$, with Σ as previously defined, mixing on q with respect to an inverse-gamma, $\mathrm{IG}(\nu/2,\nu/2)$, distribution. Under marginal $f_X(x) = t(x\,|\,\mu,\sigma,\nu)$, the Student-t MTD model is given by

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l\, t(x_t|(1-\rho_l)\mu + \rho_l x_{t-l}, \sigma^2(1-\rho_l^2)(\nu+d_l)/(\nu+1), \nu+1), \tag{8}$$

where $d_l = (x_{t-l}-\mu)^2/\sigma^2$. In both the Gaussian and Student-t MTD examples, the transition component densities and the invariant density belong to the same family of distributions.

The Student-t MTD model is an example for building MTD models through bivariate distributions that admit a location-scale mixture representation. Taking an exponential distribution for the scale q yields the bivariate Laplace distribution of Eltoft, Kim, and Lee (2006), thus producing an MTD model with an invariant Laplace marginal density. Scaling both the mean μ and the covariance Σ of the bivariate Gaussian distribution by a unit rate exponential random variable yields the bivariate asymmetric Laplace distribution of Kotz, Kozubowski, and Podgorski (2012), and thus an MTD model with an asymmetric Laplace distribution as the invariant marginal. We can further elaborate on this approach using appropriate mixing distributions for the Gaussian location and scale parameters to obtain skewed-Gaussian and skewed-t distributions (Azzalini 2013) for the bivariate component distributions, as well as for the invariant marginal distribution.

Example 2: Poisson and Poisson mixture MTD models. To model time series of counts taking countably infinite values, we can construct an MTD model with a Poisson marginal by considering the bivariate Poisson distribution of Holgate (1964) for the transition components. This choice has been discussed in Berchtold and Raftery (2002), without addressing the stationarity condition. In particular, we consider the latent variable representation of Holgate's bivariate Poisson. Given a Poisson marginal $f_X(x) = \mathrm{Pois}(x\,|\,\phi)$, we take $(U_l,V_l) \equiv (U,V) = (Q+Z, W+Z)$, for all l, where Q, W and Z are independent Poisson random variables with means λ, λ and γ, respectively. It follows that both U and V are Poisson random variables with rate parameter $\phi = \lambda+\gamma$. Using the latent variable representation, the lth component transition density of the Poisson MTD model can be sampled through $Q_t \sim \mathrm{Pois}(q_t\,|\,\lambda)$ and $Z_t\,|\,X_{t-l} = x_{t-l} \sim \mathrm{Bin}(z_t\,|\,x_{t-l},\gamma/\phi)$, with $X_t = Q_t + Z_t$ obtained as the realization from the lth component conditional distribution of $X_t\,|\,X_{t-l} = x_{t-l}$. Here, $\mathrm{Bin}(x\,|\,n,p)$ denotes the binomial distribution with n trials and probability of success p.

A common extension of the Poisson to account for counts that have excess zeros is a mixture of Poisson and a distribution that degenerates at zero. A random variable X is zero-inflated Poisson distributed, denoted as $\mathrm{ZIP}(x\,|\,\phi,q)$, if its distribution is a mixture of a point mass at zero and a Poisson distribution with parameter ϕ, with respective probabilities $0<q<1$ and $(1-q)$. Given an invariant marginal $f_X(x) = \mathrm{ZIP}(x\,|\,\phi,q)$, we use the bivariate zero-inflated Poisson distribution of Li et al. (1999) for $(U_l,V_l) \equiv (U,V)$, for all l, given by a mixture of a point mass at $(0,0)$, two univariate Poisson distributions, and a bivariate Poisson distribution; that is $f_{U,V}(u,v) = q_0(0,0) + 0.5q_1(\mathrm{Pois}(u\,|\,\phi),0) + 0.5q_1(0,\mathrm{Pois}(v\,|\,\phi)) + q_2\mathrm{BP}(u,v\,|\,\phi,\phi)$, where $\sum_{j=0}^{2} q_j = 1$, $q_0 + 0.5q_1 = q$, and $\mathrm{BP}(\cdot,\cdot|\phi,\phi)$ denotes Holgate's bivariate Poisson distribution. Although the corresponding component density $f_{U\,|\,V}(u\,|\,v) = f_{U,V}(u,v)/f_X(v)$ is complex, this example provides possibilities for modeling stationary zero-inflated count time series.

Exploiting the latent variable representation of Holgate's bivariate Poisson, we can obtain extensions of the Poisson MTD model that allow for more flexible dependence structure and for overdispersion. Following the earlier notation, replace the means λ and γ of the latent Poisson random variables with $\alpha\lambda$ and $\alpha\gamma$, and mix over α with respect to a $\mathrm{Ga}(\alpha\,|\,k,\eta)$ distribution, where $\mathrm{Ga}(x\,|\,a,b)$ denotes the gamma distribution with mean a/b. Such mixing yields a bivariate negative binomial distribution after α is marginalized out (Kocherlakota and Kocherlakota 2006). The conditional distribution of U given $V = v$ admits a convolution representation. Let Z_1 and Z_2 be conditionally independent, given $V = v$, following a $\mathrm{Bin}\,(z_1\,|\,v,\gamma/(\lambda+\gamma))$ and $\mathrm{NB}\,(z_2\,|\,k+v, 1-\lambda/(2\lambda+\gamma+\eta))$ distribution, respectively, where $\mathrm{NB}(x\,|\,r,p)$ denotes the negative binomial distribution with r number of successes and probability of success p. Then, $U = Z_1+Z_2$ is a realization from the conditional distribution $U\,|\,V = v$. Similar to the Poisson case, we can use this convolution representation to define a stationary MTD model with a negative binomial marginal $f_X(x) = \mathrm{NB}\,(x\,|\,k,\eta/(\lambda+\gamma+\eta))$.

Example 3: Bernoulli and Binomial MTD models. Assume again $(U_l,V_l) \equiv (U,V)$, for all l, and consider the bivariate Bernoulli distribution with probability mass function $p(u,v) = p_1^{uv} p_2^{u(1-v)+(1-u)v}(1-p_1-2p_2)^{(1-u)(1-v)}$, where $p_1>0, p_2>0$ and $p_1+2p_2<1$. Then, marginally U and V are both Bernoulli distributed with probability of success p_1+p_2. The conditional distribution of U given $V = v$ is also Bernoulli (Dai et al. 2013) with probability of success $p(1,v)/\left(p(1,v)+p(0,v)\right)$. Using this bivariate Bernoulli distribution, we define a stationary Bernoulli MTD model

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l\,\mathrm{Ber}\left(x_t|p(1,x_{t-l})/(p(1,x_{t-l})+p(0,x_{t-l}))\right), \tag{9}$$

which has a stationary marginal distribution $f_X(x) = \mathrm{Ber}(x\,|\,p_1+p_2)$.

Sequences of independent bivariate Bernoulli random vectors can be used as building blocks for various bivarate distributions. In particular, a family of bivariate binomial distributions for (U, V) can be constructed by setting $U = \sum_{i=1}^{n} \tilde{U}_i$ and $V = \sum_{i=1}^{n} \tilde{V}_i$, where $(\tilde{U}_i, \tilde{V}_i), i = 1, \ldots, n$, are independent from the bivariate Bernoulli distribution given above (Kocherlakota and Kocherlakota 2006). The conditional distribution of U given $V = v$ can be defined through the convolution of two conditionally independent, given $V = v$, binomial random variables, one with parameters $n - v$ and $p_2/(1 - p_1 - p_2)$ and the other with parameters v and $p_1/(p_1 + p_2)$. Again, this convolution representation can be used to define a stationary binomial MTD model with marginal $f_X(x) = \text{Bin}(x \mid n, p_1 + p_2)$.

Examples 2 and 3 illustrate MTD models for finite/infinite-range discrete-valued time series with high-order dependence, and with stationary marginal distributions belonging to a range of families. These can be used, for example, for classification of time series data, or for time-varying counts that exhibit features such as overdispersion or excess of zero values when compared to a traditional Poisson model. It is worth mentioning that some of our examples induce nonlinear conditional expectations. For example, the conditional expectation of the Bernoulli MTD model is $\sum_{l=1}^{L} w_l p(1, x_{t-l})/(p(1, x_{t-l}) + p(0, x_{t-l}))$. Building MTD models like the ones we have proposed using the existing methods in the MTD literature is a formidable task.

3.2. Conditional Distribution Method

The strategy here is to use compatible conditional densities, $f_{U_l|V_l}$ and $f_{V_l|U_l}$, to specify the bivariate density of (U_l, V_l) for the lth transition component. Conditional densities $f_{U|V}$ and $f_{V|U}$ are said to be compatible if there exists a bivariate density with its conditionals given by $f_{U|V}$ and $f_{V|U}$; see Arnold et al. (1999) for general conditions under which candidate families of two conditionals are compatible.

We begin with the assumption that $f_{U_l|V_l}$ and $f_{V_l|U_l}$ belong to the same family. This assumption is reasonable, since the invariant condition of Proposition 1 requires that all marginals are the same. Once the family of distributions for the conditionals is chosen, we ensure the conditionals are compatible, as well as that both marginals of the corresponding bivariate density are given by the target invariant density f_X. In some special cases, the marginal densities are in the same family as the compatible conditionals. To demonstrate this method, we use a pair of Lomax conditionals and a pair of gamma conditionals; both cases are considered in Arnold et al. (1999) to identify compatibility restrictions for their parameters.

Example 4: Lomax MTD models. The Lomax distribution is a shifted version of the Pareto Type I distribution such that it is supported on $\mathbb{R}^+$. Denote by $P(x \mid \sigma, \alpha) = \alpha\sigma^{-1}\left(1 + x\sigma^{-1}\right)^{-(\alpha+1)}$ the Lomax density, where $\alpha > 0$ is the shape parameter, and $\sigma > 0$ the scale parameter. The corresponding tail distribution function is $\Pr(X > x) = (1 + x\sigma^{-1})^{-\alpha}$, implying a polynomial tail that supports modeling for time series with high levels of skewness. We consider a pair of compatible Lomax densities for $(U_l, V_l) \equiv (U, V)$, for all l, such that $f_{U \mid V}(u \mid v) = P(u \mid (\lambda_0 + \lambda_1 v)/(\lambda_1 + \lambda_2 v), \alpha)$, and $f_{V \mid U}(v \mid u) = P(v \mid (\lambda_0 + \lambda_1 u)/(\lambda_1 + \lambda_2 u), \alpha)$, with the restriction that $\lambda_0, \lambda_1, \lambda_2 > 0$ if $\alpha = 1$, $\lambda_0 \geq 0, \lambda_1, \lambda_2 > 0$ if $0 < \alpha < 1$, and $\lambda_0, \lambda_1 > 0, \lambda_2 \geq 0$ if $\alpha > 1$, to guarantee that these are proper densities. Lomax MTD models specified using the conditional distributions above have an invariant marginal $f_X(x) \propto (\lambda_1 + \lambda_2 x)^{-1}(\lambda_0 + \lambda_1 x)^{-\alpha}$. Taking $\alpha > 1$ and $\lambda_2 = 0$ leads to a special case where both the component transition density and the marginal density are Lomax. This particular Lomax MTD model is

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l\, P(x_t|\phi + x_{t-l}, \alpha), \tag{10}$$

where $\phi = \lambda_0/\lambda_1$, and the invariant marginal is $f_X(x) = P(x \mid \phi, \alpha - 1)$.

Example 5: Gamma MTD models. We consider a pair of conditional gamma densities for the random vector $(U_l, V_l) \equiv (U, V)$, for all l, such that $f_{U|V}(u \mid v) = \text{Ga}(u \mid m_0, m_1 + m_2 v)$, and $f_{V|U}(v \mid u) = \text{Ga}(v \mid m_0, m_1 + m_2 u)$, where $m_0, m_1, m_2 > 0$. This pair of conditionals is one of six choices discussed in Arnold et al. (1999) in the context of conditional gamma distributions that produce proper bivariate densities for (U, V). The resulting transition density is

$$f(x_t|\boldsymbol{x}^{t-1}) = \sum_{l=1}^{L} w_l\, \text{Ga}(x_t|m_0, m_1 + m_2 x_{t-l}), \tag{11}$$

and the invariant marginal is $f_X(x) \propto x^{m_0-1}\exp(-m_1 x)(m_1 + m_2 x)^{-m_0}$.

Examples 4 and 5 present two stationary MTD models with, respectively, polynomial and exponential tail behaviors. They provide alternatives to the existing MTD model literature for positive-valued time series, where the only model that has received attention is based on the Weibull distribution. In addition, the general Lomax MTD model with $\lambda_2 \neq 0$ and the gamma MTD model have nonlinear conditional expectations.

4. Bayesian Implementation

4.1. Hierarchical Model Formulation

Here, we outline an approach to perform posterior inference for the general MTD model, using a likelihood that is conditional on the first L observations of the time series realization $\{x_t\}_{t=1}^{n}$. We introduce a set of latent variables $\{Z_t\}_{t=L+1}^{n}$ with Z_t taking values in $\{1, \ldots, L\}$ such that $p(z_t \mid \boldsymbol{w}) = \sum_{l=1}^{L} w_l \delta_l(z_t)$, where $\boldsymbol{w} = (w_1, \ldots, w_L)^\top$, and $\delta_l(z_t) = 1$ if $z_t = l$ and 0 otherwise. Conditioning on the set of latent variables and the first L observations, the hierarchical representation of the model is

$$\begin{aligned} x_t|z_t, \boldsymbol{\theta} &\overset{\text{ind}}{\sim} f_{z_t}(x_t|x_{t-z_t}, \boldsymbol{\theta}_{z_t}), \quad t = L+1, \ldots, n, \\ z_t|\boldsymbol{w} &\overset{\text{iid}}{\sim} \sum_{l=1}^{L} w_l \delta_l(z_t), \quad t = L+1, \ldots, n, \\ \boldsymbol{w} &\sim \pi_w(\cdot), \quad \boldsymbol{\theta}_l \overset{\text{ind}}{\sim} \pi_l(\cdot), \quad l = 1, \ldots, L, \end{aligned} \tag{12}$$

where $\boldsymbol{\theta}_l$ denotes the transition component parameters, and $\boldsymbol{\theta}$ collects all $\boldsymbol{\theta}_l$. Any MCMC algorithm for finite mixture models is readily adoptable. If the transition density of the model is sampled via a latent process, such as for Example 2 of Section 3, then an additional step to sample the latent variables needs to be added in Equation (12).

A key component of the Bayesian model formulation is the choice of the prior distribution for the mixture weights. As a point of reference, we consider a uniform Dirichlet prior that assumes equal contribution from each lag, denoted by $\mathrm{Dir}(\cdot|\mathbf{1}_L/L)$, where $\mathbf{1}_L$ is a unit vector of length L. We discuss next two priors that assume more structure.

The first prior is a truncated version of the stick-breaking prior, which characterizes the weights for random discrete distributions generated by the Dirichlet process (Sethuraman 1994). More specifically, the weights are constructed as follows: $w_1 = \zeta_1$, $w_l = \zeta_l \prod_{r=1}^{l-1}(1-\zeta_r)$, $l = 2, \ldots, L-1$, and $w_L = \prod_{l=1}^{L-1}(1-\zeta_l)$, where $\zeta_l \overset{i.i.d.}{\sim} \mathrm{Beta}(1, \alpha_s)$, for $l = 1, \ldots, L-1$. The resulting joint distribution for the mixture weights is a special case of the generalized Dirichlet distribution (Connor and Mosimann 1969). We denote the truncated stick-breaking prior as $\mathrm{SB}(\cdot \mid \alpha_s)$. For $l = 1, \ldots, L-1$, $E(w_l) = \alpha_s^*(1-\alpha_s^*)^{l-1}$, where $\alpha_s^* = (1+\alpha_s)^{-1}$. Hence, on average, this prior implies geometrically decreasing weights, with smaller α_s values favoring stronger contributions from recent lags. In certain applications, it may be natural to expect some directionality in the relevance of the weights implied by time, and this prior provides one option to incorporate into the model such a property.

An alternative prior is obtained by assuming that the weights are increments of a cdf G with support on $[0, 1]$; that is, $w_l = G(l/L) - G((l-1)/L)$, for $l = 1, \ldots, L$. We place a Dirichlet process prior on G, denoted as $\mathrm{DP}(\alpha_0, G_0)$, where $G_0 = \mathrm{Beta}(a_0, b_0)$ and $\alpha_0 > 0$ is the precision parameter. From the Dirichlet process definition (Ferguson 1973), given α_0 and G_0, the vector of mixture weights follows a Dirichlet distribution with shape parameter vector $\alpha_0(a_1, \ldots, a_L)^\top$, where $a_l = G_0(l/L) - G_0((l-1)/L)$, for $l = 1, \ldots, L$. We refer to this prior as the cdf-based prior, and denote it as $\mathrm{CDP}(\cdot \mid \alpha_0, a_0, b_0)$. Under this prior, we have that $E(\boldsymbol{w}) = (a_1, \ldots, a_L)^\top$. The nonparametric prior for G supports general distributional shapes, and thus allows for flexibility in the estimation of the mixture weights. In particular, multimodal distributions G can produce sparse weight vectors, with some/several entries near zero. Hence, this prior may be suitable for scenarios where there are inactive lags between influential lags and the influential lags are not necessarily the most recent lags. Heiner, Kottas, and Munch (2019) proposed a different prior for sparse probability vectors, which generally requires a larger number of prior hyperparameters.

Overall, the properties of both structured priors support flexible inference for the mixture weights, enabling our strategy to specify a large value of L, assigning a priori small probabilities to distant lags. The contribution of each lag will be induced by the mixing, with important lags being assigned large weights a posteriori.

4.2. Estimation, Model Checking and Prediction

The posterior distribution of the model parameters, based on the conditional likelihood, is

$$p(\boldsymbol{w}, \boldsymbol{\theta}, \{z_t\}_{t=L+1}^n | D_n) \propto \pi_w(\boldsymbol{w}) \prod_{l=1}^{L} \pi_l(\boldsymbol{\theta}_l) \prod_{t=L+1}^{n} \left\{ f_{z_t}(x_t | x_{t-z_t}, \boldsymbol{\theta}_{z_t}) \sum_{l=1}^{L} w_l \delta_l(z_t) \right\}, \tag{13}$$

where $D_n = \{x_t\}_{t=L+1}^n$, and it can be explored using MCMC posterior simulation.

Conditional on $\boldsymbol{\theta}$ and $\boldsymbol{w}$, the posterior full conditional of each Z_t is a discrete distribution on $\{1, ..., L\}$ with probabilities proportional to $w_l f_l(x_t \mid x_{t-l}, \boldsymbol{\theta}_l)$. Conditional on the latent variables and $\boldsymbol{w}$, the sampling for each $\boldsymbol{\theta}_l$ depends on the particular choice of the transition component distributions. Details for the models implemented are given in the supplementary material. The sampling for $\boldsymbol{w}$, conditional on $\{z_t\}_{t=L+1}^n$ and $\boldsymbol{\theta}$, depends only on $M_l = |\{t : z_t = l\}|$, for $l = 1, ..., L$, where $|\{\cdot\}|$ is the cardinality of the set $\{\cdot\}$. Both priors for the mixture weights result in ready updates. The posterior full conditional of $\boldsymbol{w}$ under the truncated stick-breaking prior can be sampled through latent variables ζ_l^*, which are conditionally independent $\mathrm{Beta}(1 + M_l, \alpha_s + \sum_{r=l+1}^{L} M_r)$, for $l = 1, \ldots, L-1$, such that $w_1 = \zeta_1^*$, $w_l = \zeta_l^* \prod_{r=1}^{l-1}(1-\zeta_r^*)$, for $l = 2, \ldots, L-1$, and $w_L = \prod_{l=1}^{L-1}(1-\zeta_l^*)$. Under the cdf-based prior, the posterior full conditional of $\boldsymbol{w}$ is Dirichlet with parameter vector $(\alpha_0 a_1 + M_1, \ldots, \alpha_0 a_L + M_L)^\top$.

We assess the model's validity using randomized quantile residuals (Dunn and Smyth 1996; Escarela, Mena, and Castillo-Morales 2006). Such residuals are calculated by inverting the fitted conditional cdf for the time series. Posterior samples of these quantile sets can then be compared with the standard Gaussian distribution, providing a measure of goodness-of-fit with uncertainty quantification. Specifically, the randomized quantile residual for continuous x_t is defined as $r_t = \Phi^{-1}\left(F(x_t \mid \boldsymbol{x}^{t-1})\right)$ where $\Phi(\cdot)$ is the cdf of the standard Gaussian distribution. If x_t is discrete, $r_t = \Phi^{-1}(u_t)$, where u_t is generated from a uniform distribution on the interval (a_t, b_t) with $a_t = F(x_t - 1 \mid \boldsymbol{x}^{t-1})$ and $b_t = F(x_t \mid \boldsymbol{x}^{t-1})$. If F is correctly specified, the residuals r_t, $t = L+1, \ldots, n$, will be independently and identically distributed as a standard Gaussian distribution.

Finally, we consider prediction for future observations. The posterior predictive density of X_{n+1}, corresponding to the first out-of-sample observation, is obtained by marginalizing the transition density with respect to the posterior distribution of model parameters:

$$p(x_{n+1}|D_n) = \int \int \left\{ \sum_{l=1}^{L} w_l f_l(x_{n+1}|x_{n+1-l}, \boldsymbol{\theta}_l) \right\} p(\boldsymbol{\theta}, \boldsymbol{w}|D_n)\, d\boldsymbol{\theta}\, d\boldsymbol{w}. \tag{14}$$

Exploiting the structure of the conditional distributions of the MTD model, we can sample from the k-step-ahead posterior predictive density using a straightforward extension of Equation (14). Note that the k-step-ahead posterior predictive uncertainty incorporates both the uncertainty from the parameter estimation, and the uncertainty from the predictions of the previous $(k-1)$ out-of-sample observations.

5. Data illustrations

5.1. Simulation Example

We generated 2000 observations from the Gaussian MTD model specified in Equation (7) with $\mu = 10, \sigma^2 = 100$, under two scenarios for the mixture weights, one with exponentially decreasing weights and the other one with an uneven arrangement of the relevant lags. In Scenario 1, we took $\boldsymbol{\rho} = (0.7, 0.3, 0.1, 0.05, 0.05)^\top$ and $w_i \propto \exp(-i), i = 1, \ldots, 5$. In Scenario 2, we took $\boldsymbol{\rho} = (0.4, 0.1, 0.7, 0.1, 0.5)^\top$ and $\boldsymbol{w} = (0.2, 0.05, 0.45, 0.05, 0.25)^\top$. We consider these two scenarios to examine the effectiveness of structured priors for the mixture weights.

We applied the Gaussian MTD model with three different orders $L = 5, 15, 25$. In each case, we considered three priors for the weights: the Dirichlet prior, the truncated stick-breaking prior, and the cdf-based prior. The shape parameter of the Dirichlet prior was $\mathbf{1}_L/L$ for each L. The precision parameter α_s for the truncated stick-breaking prior was taken to be 1, 2, 3, corresponding to the three L values. For the cdf-based prior, we chose $\alpha_0 = 5$ as the precision parameter, and used as base distribution a beta with shape parameter $a_0 = 1$, and $b_0 = 3, 6, 7$, respectively, for the three orders considered. Thus, this prior elicited a decreasing pattern similar to the truncated stick-breaking prior. For all models, the mean μ and the variance σ^2 received conjugate priors $N(\mu \mid 0, 100)$ and $\mathrm{IG}(\sigma^2 \mid 2, 0.1)$, respectively, and the component-specific correlation coefficient ρ_l was assigned a uniform prior Unif$(-1, 1)$ independently for all l.

We ran the Gibbs sampler for 165,000 iterations, discarding the first 5000 samples as burn-in, and collected samples every 20 iterations. Focusing on inference results for the mixture weights, when the order was correctly specified, that is, $L = 5$, all three models provided good estimates. Figure 1 provides a visual inspection on the posterior estimates for the mixture weights when $L = 15$ (the weight patterns estimated from the three models were similar when $L = 25$). In Scenario 1, all models underestimated the weight for lag 2. Models with the proposed priors produced accurate estimates for the rest of the lags, while the model that used the Dirichlet prior systematically overestimated the weight for the first lag, and underestimated all other weights. In Scenario 2, all models underestimated the weight for the first lag. For the other nonzero weights, the model with the Dirichlet prior tended to underestimate the weights for lags 2, 4 and overestimated the weight for lag 5, while the other two models estimated the weights quite well. In both scenarios, the proposed priors had a parsimonious behavior in that, given the data, distant lags were assigned almost zero probability mass with low posterior uncertainty. Overall, we note that, under an over-specified order L, the proposed priors offer inferential advantages when compared to the Dirichlet prior.

We conducted an additional simulation to demonstrate the ability of the negative binomial MTD model to accommodate over-dispersed count data, including comparison with the Poisson MTD model. Details of this simulation example are presented in the supplementary material.

5.2. Chicago Crime Data

The first real data example involves the 1090 daily reported incidents of domestic-related theft that have occurred in Chicago from 2015 to 2017, extracted online from the Chicago Data Portal (https://data.cityofchicago.org/). The data exhibits some flat stretches, without evidence of overdispersion. The empirical mean and variance are 6.05 and 6.39.

We applied the Poisson MTD model discussed in Example 2 of Section 3, with order $L = 20$, selected based on the autocorrelation and partial autocorrelation functions. We reparameterize the model in terms of rate parameter λ, and binomial probability $\theta = \gamma/\phi$ for $Z_t \mid X_{t-l}$. This allows updates for λ and θ with posterior full conditionals available in closed form. The prior for (λ, θ) was taken to be $\mathrm{Ga}(\lambda \mid 2, 1)\mathrm{Beta}(\theta \mid 2, 2)$, implying a $\mathrm{Ga}(4, 1)$ prior for ϕ. Two priors, $\mathrm{SB}(w \mid 2)$ and $\mathrm{CDP}(w \mid 5, 1, 8)$, were considered for the mixture weights. Both models were fitted to the entire dataset. After fitting the model, we obtained the one-step posterior predictive distribution at each time t and the corresponding posterior predictive intervals.

We obtained a thinned sample retaining every 10th iteration, from a total of 85000 samples with the first 5000 as burn-in. The posterior mean and 95% interval for ϕ are 6.04 (5.79, 6.30) and 6.05 (5.82, 6.29) for models with $\mathrm{SB}(\boldsymbol{w} \mid 2)$ and $\mathrm{CDP}(\boldsymbol{w} \mid 5, 1, 8)$ priors. This indicates an average of around six incidents of domestic-related theft per day. Multiple influential lags, with gaps in between, are suggested by the results in Figures 2(b) and 2(c). Both models agree on the pattern for the weights, as well as on lags 1, 4, 6 being the most relevant ones. Compared to the truncated stick-breaking prior, the cdf-based prior suggests a weight pattern that decreases slightly faster, and it assigns relatively larger weights to important lags, albeit with higher uncertainty. Figure 2(a) shows that both models produce similar one-step predictive intervals.

5.3. Tunkhannock Creek Precipitation Data

Our second example involves 22 years of rainfall data from January 1982 to December 2003. The data consist of 1149 mean areal precipitation amounts ranging from 0.01 to 128.87 millimeters, aggregated to a weekly time scale from the daily data for the Tunkhannock Creek near Tunkhannock, Pennsylvania. The data were extracted through R package hddtools (Vitolo 2017).

We consider a multiplicative model $y_t = \mu_t \epsilon_t$, where μ_t is a seasonal factor and ϵ_t is generated by a Lomax MTD model specified in Equation (10), with polynomial tails that can accommodate large precipitation events. More specifically, the model is given by

$$y_t = \mu_t \epsilon_t, \quad \mu_t = \exp(\boldsymbol{x}_t^\top \boldsymbol{\beta}), \quad t = 1, \ldots, n,$$

$$\epsilon_t | \boldsymbol{\epsilon}^{t-1}, \boldsymbol{w}, \phi, \alpha \sim \sum_{l=1}^{L} w_l \, P(\epsilon_t | \phi + \epsilon_{t-l}, \alpha), \tag{15}$$

$$t = L+1, \ldots, n,$$

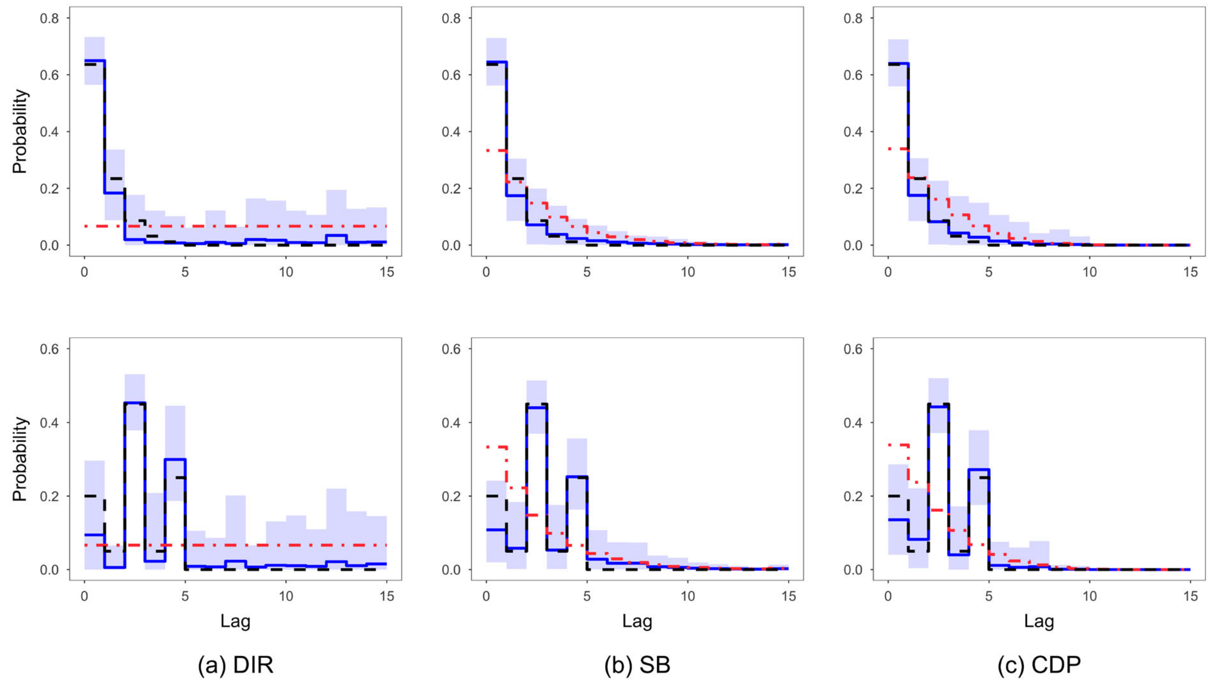

Figure 1. Simulation study. Inference results for the weights under Scenarios 1 (top) and 2 (bottom), based on the Gaussian MTD model ($L = 15$) with the Dirichlet (column (a)), the truncated stick-breaking (column (b)), and the cdf-based (column (c)) priors. Dashed lines are true weights, dot-dashed lines are prior means, solid lines are posterior means, and polygons are 95% posterior credible intervals.

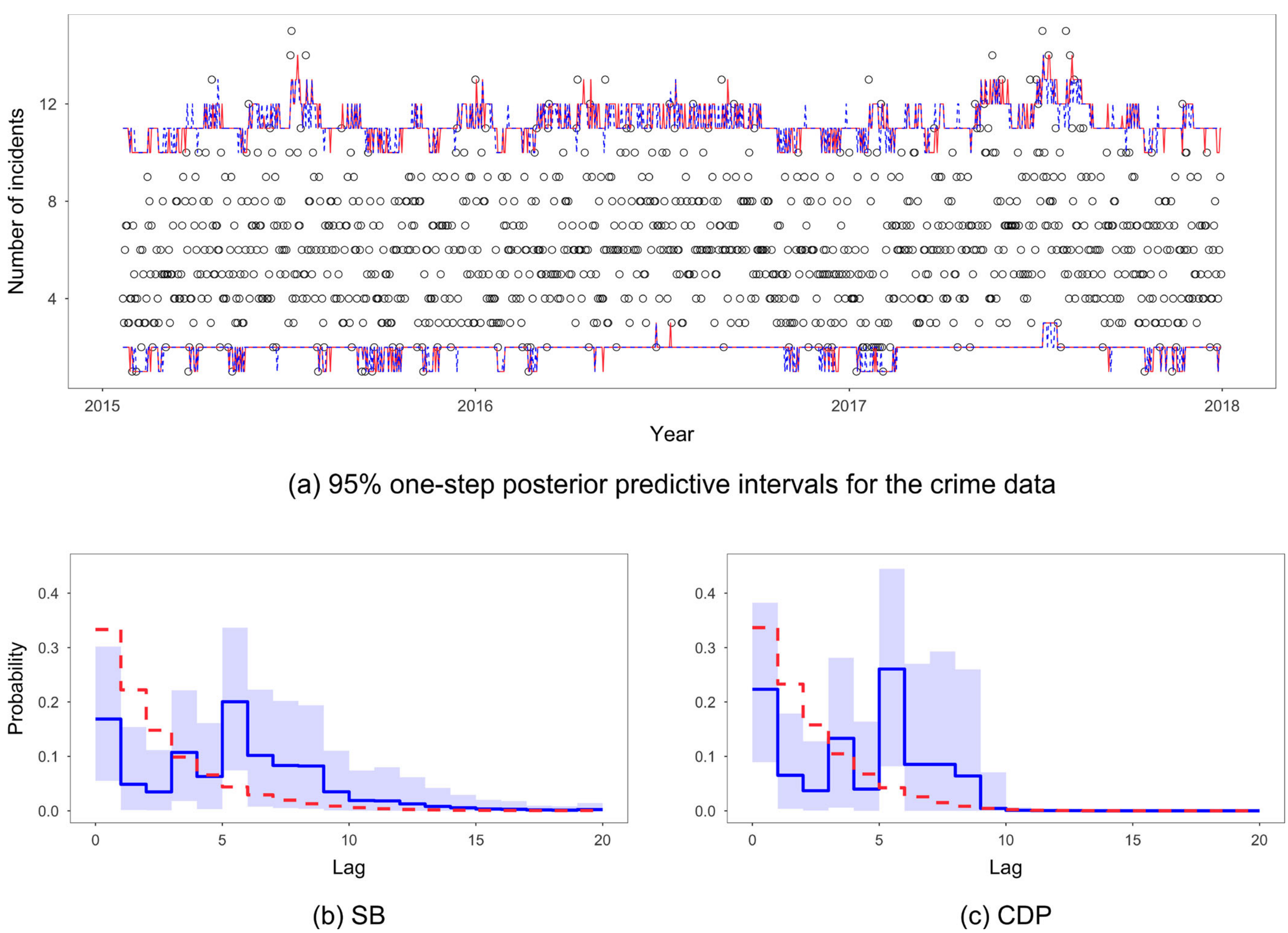

Figure 2. Chicago crime data analysis. In panel (a), the circles denote the data, and solid and dashed lines correspond to the model with the SB and CDP priors, respectively. Panels (b) and (c): prior means (dashed line), posterior means (solid line) and 95% credible intervals (polygon) of the weights under the SB and CDP priors, respectively.

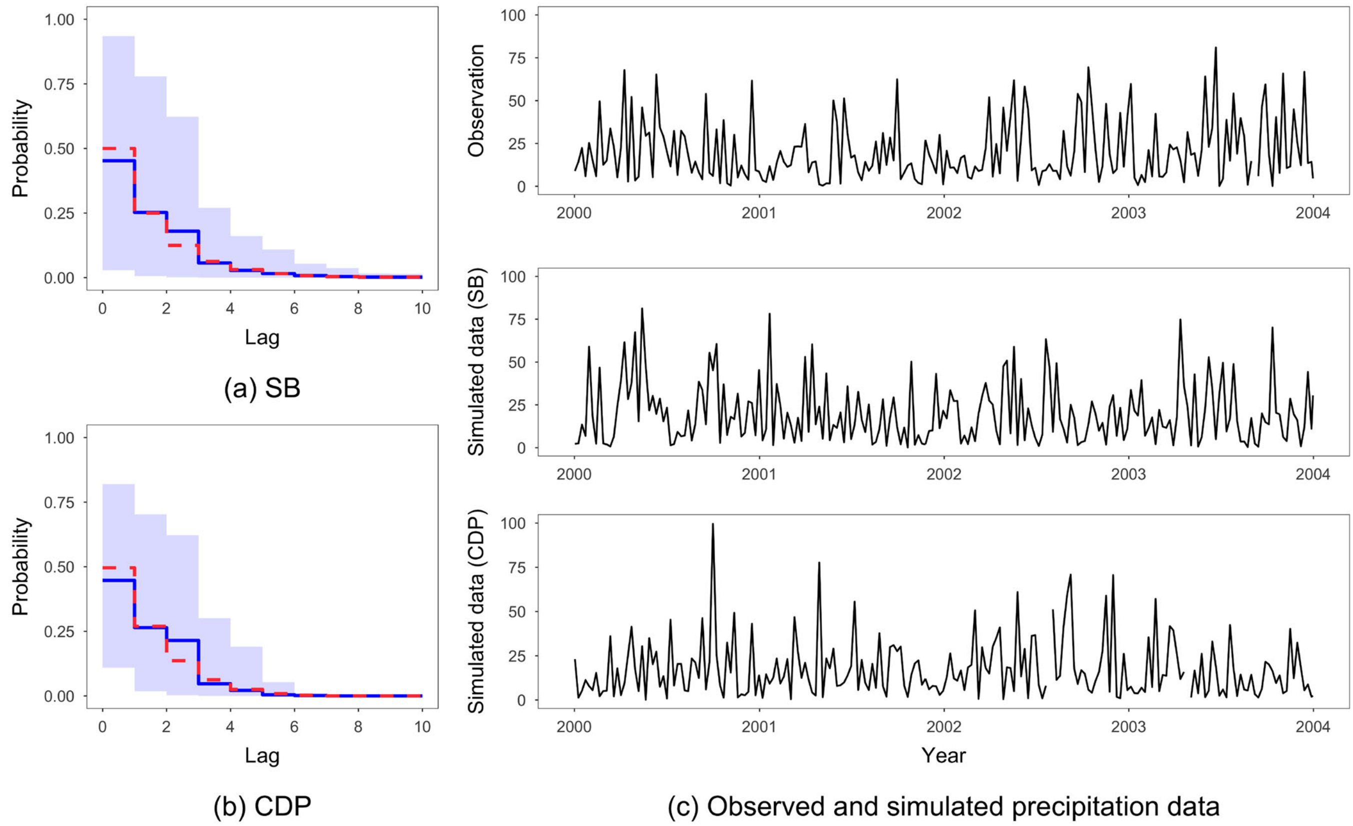

Figure 3. Precipitation data analysis. Panels (a) and (b): prior means (dashed line), posterior means (solid line) and 95% intervals (polygons) of the weights under two priors. The top row of panel (c) plots the observed precipitation amounts from 2000 to 2004, and the middle and bottom rows show sample paths generated from the fitted models with SB and CDP priors, respectively.

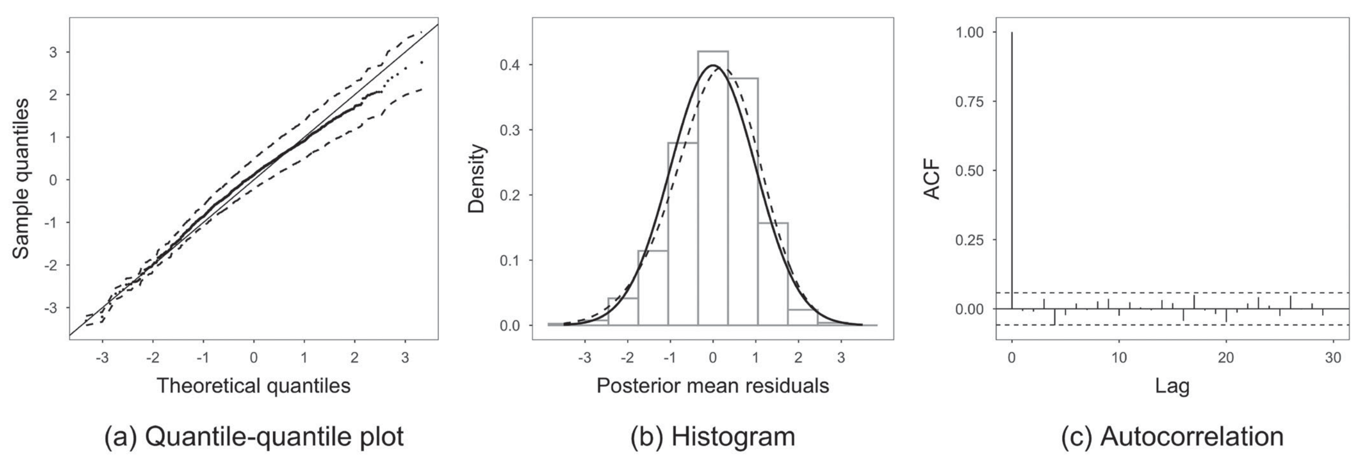

Figure 4. Precipitation data analysis. Randomized quantile residual analysis for the fitted model with the $\mathrm{SB}(w \mid 1)$ prior. In panel (a), the circles and dashed lines correspond to the posterior mean and 95% interval bands, respectively. In panel (b), the solid and dashed line are the standard Gaussian density and the kernel density estimate of the posterior means of the residuals, respectively. Panel (c) is based on the posterior means of the residuals.

with $\boldsymbol{x}_t = (\cos(\omega t), \sin(\omega t), \cos(2\omega t), \sin(2\omega t), \cos(3\omega t), \sin(3\omega t))^\top$, and $\omega = 2\pi/T$ where $T = 52$ is the period for weekly data. On the basis of the autocorrelation and partial autocorrelation functions, we chose model order $L = 10$. The regression coefficients vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_6)^\top$ was assigned a flat prior. The shape parameter α was assigned a $\mathrm{Ga}(\alpha \mid 6, 1)$ prior, and the scale parameter ϕ an $\mathrm{IG}(\phi \mid 3, 20)$ prior. Note that the invariant marginal of the process $\{\epsilon_t\}$ is $P(\epsilon \mid \phi, \alpha - 1)$ and its tail distribution function is $(1 + \epsilon/\phi)^{-(\alpha-1)}$. A small value of α indicates a heavy tail, while a large value of α ensures the existence of finite high moments. Under the priors above, $E(\alpha) = 6$, implying the expectation that the first four moments are finite with respect to both the component and marginal distributions of the Lomax MTD for $\{\epsilon_t\}$. We fit the model with $\mathrm{SB}(\boldsymbol{w} \mid 1)$ and $\mathrm{CDP}(\boldsymbol{w} \mid 5, 1, 6.5)$ priors for the weights.

We ran the algorithm for 85,000 iterations and collected samples every 10 iterations after the first 5000 was discarded. The inference results were almost the same for the two models. Here we report the ones under the $\mathrm{SB}(\boldsymbol{w} \mid 1)$ prior. The posterior mean and 95% credible interval of the shape parameter α are 14.80 (10.30, 20.91), indicating a moderately heavy tail. The corresponding estimates for the scale parameter ϕ are 254.33 (166.36, 370.04), indicating substantial dispersion. Among the harmonic component coefficients, the first and the fourth have 95% posterior credible intervals that indicate statistical significance; the estimates are −0.14 (−0.23, −0.05) for β_1, and −0.13 (−0.22, −0.03) for β_4, implying the presence of semiannual and annual seasonality in the data. Figures 3(a) and 3(b) shows that both models suggest a decreasing weight pattern, with the first three lags being the most influential. As shown in Figure 3(c), the sample paths generated from the models resemble the observed precipitation time series.

Randomized quantile residual analysis results were similar for both models; in Figure 4, we present the ones under the SB($\boldsymbol{w} \mid 1$) prior. The figure shows posterior mean and interval estimates for the Gaussian quantile–quantile plot, and the histogram and autocorrelation function for the posterior means of the residuals. The results suggest reasonably good model fit, providing an illustration of the flexibility of the proposed MTD model to capture non-Gaussian tails.

6. Discussion

We have developed a broad class of stationary MTD models focusing on attaining stationarity from the perspective of a distributional formulation. The advantage of our proposed approach over more traditional methods is that no constraints on the parameter space are needed. This facilitates inference for model parameters, as the need for constrained optimization or sampling is avoided. We further proposed structured priors to support flexible inference on the weights, which accommodate nonstandard scenarios that a model with a Dirichlet prior may fail to capture.

The proposed constructive framework brings several options for alternative parametric families that were formidable to tackle for the MTD model and its extensions, when stationarity is a desirable property. A limitation of our approach is that, if the stationary marginal distribution shares all the parameters with the bivariate component distribution, the resulting transition component lacks component-varying parameters. One solution is to specify the bivariate distribution using a copula (Joe 2014), which we regard as a special case of the bivariate distribution method. Given a prespecified marginal, the construction boils down to the selection of a copula. The copula function, which brings additional component parameters, allows specifying dependence in the bivariate distribution, separately from modeling the marginal distribution. On the other hand, some properties of the resulting model, including the conditional expectation, may be intractable, and the computational cost may increase, especially in the discrete case.

The class of models proposed in this article can be easily extended for nonstationary time series that exhibit trends and seasonality, by incorporating corresponding factors into the model, either multiplicatively or additively. This is illustrated in our second real data example. A similar approach can be applied to incorporate covariates. Therefore, this class of models is quite general, and is useful as an alternative to the existing time series models, especially when traditional models fail to capture non-Gaussian features suggested by the data.

Acknowledgments

This research was supported in part by the National Science Foundation under awards SES 1631963, and SES 2050012. The authors wish to thank three reviewers and an Associate Editor for useful comments.

ORCID

Xiaotian Zheng http://orcid.org/0000-0001-9284-5411

Supplementary Material

The supplementary material includes the proof for Proposition 2, sampling algorithm details, additional simulation results, model checking for the data examples, and R code to reproduce the results in Section 5.

Appendix

Proof of Proposition 1

Proof. Without loss of generality, we consider the case where X_t has a continuous distribution for all t. Moreover, for the argument that follows to apply to any $t \geq 2$, we express the transition density as $f(x_t|x^{t-1}) = \sum_{l=1}^{t_L} w_l^* f_{U_l|V_l}(x_t|x_{t-l})$, for $t \geq 2$, where $t_L = \min\{t-1, L\}$. When $t > L$, $w_l^* \equiv w_l$, for $l = 1, ..., L$, whereas for $2 \leq t \leq L$, $w_l^* = w_l$, for $l = 1, ..., t_L - 1$, and $w_{t_L}^* = 1 - \sum_{k=1}^{t_L-1} w_k$. With this notational convention, we have $\sum_{l=1}^{t_L} w_l^* = 1$.

Using the proposition assumptions,

$$g_2(x_2) = \int_{\mathcal{S}} f(x_2|x_1) f_X(x_1) dx_1 = \int_{\mathcal{S}} f_{U_1|V_1}(x_2|x_1) f_{V_1}(x_1) dx_1 \\ = f_{U_1}(x_2) = f_X(x_2)$$

and thus the result is valid for $t = 2$. To prove the proposition by induction, assume the result holds true for generic $t-1$, that is, $g_{t'}(x_{t'}) = f_X(x_{t'})$, for all $x_{t'} \in \mathcal{S}$, and for all $t' \leq t-1$. Denote by $p(x_1, \ldots, x_{t-1})$ and $p(x_{t-t_L}, \ldots, x_{t-1})$ the joint density for random vector $(X_1, \ldots, X_{t-1})$ and $(X_{t-t_L}, \ldots, X_{t-1})$, respectively. Then, the marginal density for X_t can be derived as follows:

$$\begin{aligned} g_t(x_t) &= \int_{\mathcal{S}^{t-1}} f(x_t|x^{t-1})\, p(x_1, \ldots, x_{t-1})\, dx_1 \ldots dx_{t-1} \\ &= \sum_{l=1}^{t_L} w_l^* \int_{\mathcal{S}^{t_L}} f_{U_l|V_l}(x_t|x_{t-l})\, p(x_{t-t_L}, \ldots, x_{t-1}) \\ &\qquad dx_{t-t_L} \ldots dx_{t-1} \\ &= \sum_{l=1}^{t_L} w_l^* \int_{\mathcal{S}} f_{U_l|V_l}(x_t|x_{t-l})\, g_{t-l}(x_{t-l})\, dx_{t-l} \\ &= \sum_{l=1}^{t_L} w_l^* \int_{\mathcal{S}} f_{U_l|V_l}(x_t|x_{t-l}) f_{V_l}(x_{t-l})\, dx_{t-l} \\ &= f_X(x_t), \end{aligned}$$

where for the second-to-last equation we used $g_{t-l} = f_X$, for $l = 1, ..., t_L$, obtained from the induction argument, as well as the proposition assumption, $f_X = f_{V_l}$, for all l. Finally, the last equation is based on the proposition assumption that $f_{U_l} = f_X$, for all l. □

References

Arnold, B. C., Castillo, E., Sarabia, J.-M., and Sarabia, J. M. (1999), *Conditional Specification of Statistical Models*, Springer Science & Business Media. [5]

Azzalini, A. (2013), *The Skew-Normal and Related Families*, Vol. 3, Cambridge: Cambridge University Press. [4]

Bartolucci, F. and Farcomeni, A. (2010), "A Note on the Mixture Transition Distribution and Hidden Markov Models," *Journal of Time Series Analysis*, 31, 132–138. [1]

Berchtold, A. (2001), "Estimation in the Mixture Transition Distribution Model," *Journal of Time Series Analysis*, 22, 379–397. [1]

——— (2003), "Mixture Transition Distribution (MTD) Modeling of Heteroscedastic Time Series," *Computational Statistics & Data Analysis*, 41, 399–411. [1]

Berchtold, A., and Raftery, A. (2002), "The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series," *Statistical Science*, 17, 328–356. [1,4]

Bolano, D., and Berchtold, A. (2016), "General Framework and Model Building in the Class of Hidden Mixture Transition Distribution Models," *Computational Statistics & Data Analysis*, 93, 131–145. [1]

Cervone, D., Pillai, N. S., Pati, D., Berbeco, R., and Lewis, J. H. (2014), "A Location-Mixture Autoregressive Model for Online Forecasting of Lung Tumor Motion," *The Annals of Applied Statistics*, 8, 1341–1371. [1]

Connor, R. J., and Mosimann, J. E. (1969), "Concepts of Independence for Proportions With a Generalization of the Dirichlet Distribution," *Journal of the American Statistical Association*, 64, 194–206. [6]

Dai, B., Ding, S., and Wahba, G. (2013), "Multivariate Bernoulli Distribution," *Bernoulli*, 19, 1465–1483. [4]

Dunn, P. K., and Smyth, G. K. (1996), "Randomized Quantile Residuals," *Journal of Computational and Graphical Statistics*, 5, 236–244. [6]

Eltoft, T., Kim, T., and Lee, T.-W. (2006), "On the Multivariate Laplace Distribution," *IEEE Signal Processing Letters*, 13, 300–303. [4]

Escarela, G., Mena, R. H., and Castillo-Morales, A. (2006), "A Flexible Class of Parametric Transition Regression Models Based on Copulas: Application to Poliomyelitis Incidence," *Statistical Methods in Medical Research*, 15, 593–609. [1,6]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [6]

Fong, P. W., Li, W. K., Yau, C., and Wong, C. (2007), "On a Mixture Vector Autoregressive Model," *Canadian Journal of Statistics*, 35, 135–150. [1]

Hassan, M. Y., and Lii, K.-S. (2006), "Modeling Marked Point Processes Via Bivariate Mixture Transition Distribution Models," *Journal of the American Statistical Association*, 101, 1241–1252. [1]

Heiner, M., and Kottas, A. (2021a), "Autoregressive Density Modeling With the Gaussian Process Mixture Transition Distribution," *Journal of Time Series Analysis*, to appear. [1]

——— (2021b), "Estimation and Selection for High-Order Markov Chains With Bayesian Mixture Transition Distribution Models," *Journal of Computational and Graphical Statistics*, to appear. [1]

Heiner, M., Kottas, A., and Munch, S. (2019), "Structured Priors for Sparse Probability Vectors With Application to Model Selection in Markov Chains," *Statistics and Computing*, 29, 1077–1093. [6]

Holgate, P. (1964), "Estimation for the Bivariate Poisson Distribution," *Biometrika*, 51, 241–287. [4]

Joe, H. (2014), *Dependence Modeling With Copulas*, Boca Raton, FL: CRC Press. [10]

Kalliovirta, L., Meitz, M., and Saikkonen, P. (2015), "A Gaussian Mixture Autoregressive Model for Univariate Time Series," *Journal of Time Series Analysis*, 36, 247–266. [1,2]

——— (2016), "Gaussian Mixture Vector Autoregression," *Journal of Econometrics*, 192, 485–498. [1]

Khalili, A., Chen, J., and Stephens, D. A. (2017), "Regularization and Selection in Gaussian Mixture of Autoregressive Models," *Canadian Journal of Statistics*, 45, 356–374. [1]

Kocherlakota, S., and Kocherlakota, K. (2006), *Bivariate Discrete Distributions*, *Encyclopedia of Statistical Sciences*. Hoboken, NJ: Wiley. [4,5]

Kotz, S., Kozubowski, T., and Podgorski, K. (2012), *The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Economics, Engineering, and Finance*, Springer Science & Business Media. [4]

Lanne, M., and Saikkonen, P. (2003), "Modeling the US short-term Interest Rate by Mixture Autoregressive Processes," *Journal of Financial Econometrics*, 1, 96–125. [1]

Lau, J. W., and So, M. K. (2008), "Bayesian Mixture of Autoregressive Models," *Computational Statistics & Data Analysis*, 53, 38–60. [1]

Le, N. D., Martin, R. D., and Raftery, A. E. (1996), "Modeling Flat Stretches, Bursts Outliers in Time Series Using Mixture Transition Distribution Models," *Journal of the American Statistical Association*, 91, 1504–1515. [1,2]

Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P. A., and Peterson, J. P. (1999), "Multivariate Zero-Inflated Poisson Models and Their Applications," *Technometrics*, 41, 29–38. [4]

Li, G., Zhu, Q., Liu, Z., and Li, W. K. (2017), "On Mixture Double Autoregressive Time Series Models," *Journal of Business & Economic Statistics*, 35, 306–317. [1]

Luo, J. and Qiu, H.-b. (2009), "Parameter Estimation of the WMTD Model," *Applied Mathematics-A Journal of Chinese Universities*, 24, 379. [1]

MacDonald, I. L. and Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete-Valued Time Series*, Vol. 110, Boca Raton, FL: CRC Press. [1]

Meitz, M., Preve, D., and Saikkonen, P. (2021), "A Mixture Autoregressive Model Based on Student's t-Distribution," *Communications in Statistics-Theory and Methods*, 1–76. [1,2]

Mena, R. H., and Walker, S. G. (2007), "Stationary Mixture Transition Distribution (MTD) Models Via Predictive Distributions," *Journal of Statistical Planning and Inference*, 137, 3103–3112. [2,3]

Nguyen, H. D., McLachlan, G. J., Ullmann, J. F., and Janke, A. L. (2016), "Laplace Mixture Autoregressive Models," *Statistics & Probability Letters*, 110, 18–24. [1]

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002), "Constructing First Order Stationary Autoregressive Models Via Latent Processes," *Scandinavian Journal of Statistics*, 29, 657–663. [2,3]

Raftery, A. and Tavaré, S. (1994), "Estimation and Modelling Repeated Patterns in High Order Markov Chains With the Mixture Transition Distribution Model," *Journal of the Royal Statistical Society*, Series C, 43, 179–199. [1]

Raftery, A. E. (1985), "A Model for High-Order Markov Chains," *Journal of the Royal Statistical Society*, Series B, 47, 528–539. [1]

——— (1994), "Change Point and Change Curve Modeling in Stochastic Processes and Spatial Statistics," *Journal of Applied Statistical Science*, 1, 403–423. [1]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [6]

Vitolo, C. (2017), "hddtools: Hydrological Data Discovery Tools," *The Journal of Open Source Software*, 2, 56. [7]

Wong, C., Chan, W., and Kam, P. (2009), "A Student t-Mixture Autoregressive Model With Applications to Heavy-Tailed Financial Data," *Biometrika*, 96, 751–760. [1]

Wong, C. S., and Li, W. K. (2000), "On a Mixture Autoregressive Model," *Journal of the Royal Statistical Society*, Series B, 62, 95–115. [1]

——— (2001a), "On a Logistic Mixture Autoregressive Model," *Biometrika*, 88, 833–846. [1]

——— (2001b), "On a Mixture Autoregressive Conditional Heteroscedastic Model," *Journal of the American Statistical Association*, 96, 982–995. [1]

Zhu, F., Li, Q., and Wang, D. (2010), "A Mixture Integer-Valued ARCH Model," *Journal of Statistical Planning and inference*, 140, 2025–2036. [1]