A Gaussian mixture autoregressive model for univariate time series*

Leena Kalliovirta
University of Helsinki

Mika Meitz
Koç University

Pentti Saikkonen
University of Helsinki

August 14, 2012

Abstract

This paper presents a general formulation for the univariate nonlinear autoregressive model discussed by Glasbey [*Journal of the Royal Statistical Society: Series C*, 50(2001), 143–154] in the first order case, and provides a more thorough treatment of its theoretical properties and practical usefulness. The model belongs to the family of mixture autoregressive models but it differs from its previous alternatives in several advantageous ways. A major theoretical advantage is that, by the definition of the model, conditions for stationarity and ergodicity are always met and these properties are much more straightforward to establish than is common in nonlinear autoregressive models. Moreover, for a pth order model an explicit expression of the $(p+1)$–dimensional stationary distribution is known and given by a mixture of Gaussian distributions with constant mixing weights. Lower dimensional stationary distributions have a similar form whereas the conditional distribution given the past observations is a Gaussian mixture with time varying mixing weights that depend on p lagged values of the series in a natural way. Due to the known stationary distribution exact maximum likelihood estimation is feasible, and one can assess the applicability of the model in advance by using a nonparametric estimate of the density function. An empirical example with interest rate series illustrates the practical usefulness of the model.

*The first and third authors thank the Academy of Finland and the OP-Pohjola Group Research Foundation for financial support. We thank Esa Nummelin, Antti Ripatti, Timo Teräsvirta, and Howell Tong for useful comments and suggestions. Contact addresses: Leena Kalliovirta, Department of Political and Economic Studies, University of Helsinki, P. O. Box 17, FI–00014 University of Helsinki, Finland; e-mail: leena.kalliovirta@helsinki.fi. Mika Meitz, Department of Economics, Koç University, Rumelifeneri Yolu, 34450 Sariyer, Istanbul, Turkey; e-mail: mmeitz@ku.edu.tr. Pentti Saikkonen, Department of Mathematics and Statistics, University of Helsinki, P. O. Box 68, FI–00014 University of Helsinki, Finland; e-mail: pentti.saikkonen@helsinki.fi.

1 Introduction

During the past two or three decades various nonlinear autoregressive (AR) models have been proposed to model time series data. This paper is confined to univariate parametric models although multivariate models and nonparametric models have also attracted interest. Tong (1990) and Granger and Teräsvirta (1993) provide comprehensive accounts of the early stages of threshold autoregressive (TAR) models and smooth transition autoregressive (STAR) models which have become perhaps the most popular nonlinear AR models (see also the review of Tong (2011)). An up-to-date discussion of TAR and STAR models, as well as other nonlinear time series models, can be found in Teräsvirta, Tjøstheim, and Granger (2010). From a statistical perspective, TAR and STAR models are distinctively models for the conditional expectation of a time series given its past history although they may also include a time varying conditional variance (here, as well as later, a TAR model refers to a self-exciting TAR model or a SETAR model). The conditional expectation is specified as a convex combination of conditional expectations of two or more linear AR models and similarly for the conditional variance if it is assumed time varying. The weights of these convex combinations (typically) depend on a past value of the time series so that different models are obtained by different specifications of the weights.

The specification of TAR and STAR models is focused on the conditional expectation (and possibly conditional variance) and not so much on the conditional distribution which in parameter estimation is typically assumed to be Gaussian. In so-called mixture AR models the focus is more on the specification of the entire conditional distribution. In these models the conditional distribution, not only the conditional expectation (and possibly conditional variance) is specified as a convex combination of (typically) Gaussian conditional distributions of linear AR models. Thus, the conditional distribution is a

mixture of Gaussian distributions and, similarly to TAR and STAR models, different models are obtained by different specifications of the mixing weights, often assumed to be functions of past values of the series. Models of this kind were introduced by Le, Martin, and Raftery (1996) and further developed by Wong and Li (2000, 2001a,b). Further references include Glasbey (2001), Lanne and Saikkonen (2003), Gourieroux and Robert (2006), Dueker, Sola, and Spagnolo (2007), and Bec, Rahbek, and Shephard (2008) (for reasons to be discussed in Section 2.3 we treat the model of Dueker, Sola, and Spagnolo (2007) as a mixture model although the authors call it a STAR model). Markov switching AR models (see, e.g., Hamilton (1994, Ch. 22)) are also related to mixture AR models although the Markov chain structure used in their formulation makes them distinctively different from the mixture AR models we are interested in.

A property that makes the stationary linear Gaussian AR model different from most, if not nearly all, of its nonlinear AR alternatives is that the probability structure of the underlying stochastic process is fully known. In particular, the joint distribution of any finite realization is Gaussian with mean and covariance matrix being simple functions of the parameters of the conditional distribution used to parameterize the model. In nonlinear AR models the situation is typically very different. The conditional distribution is known by construction but what is usually known beyond that is only the existence of a stationary distribution and finiteness of some of its moments. As discussed by Tong (2011, Section 4.2) an explicit expression for the stationary distribution or its density function is only rarely known and usually only in simple special cases. Furthermore, conditions under which the stationary distribution exists may not be fully known. A notable exception is the mixture AR model discussed by Glasbey (2001, Section 3). In his paper Glasbey (2001) explicitly considers the model only in the first order case and applies it to solar radiation data. In this paper, we extend this model to the general pth order case and provide a more detailed discussion of its properties.

In the considered mixture AR model the mixing weights are defined in a specific way which turns out to have very convenient implications from both theoretical and practical point of view. A theoretical consequence is that stationarity of the underlying stochastic process is a simple consequence of the definition of the model and ergodicity can also be established straightforwardly without imposing any additional restrictions on the parameter space of the model. Moreover, in the pth order case, the $(p+1)$–dimensional stationary distribution is known to be a mixture of Gaussian distributions with constant mixing weights and known structure for the mean and covariance matrix of the component distributions. Consequently, all lower dimensional stationary distributions are of the same type. From the specification of the mixing weights it also follows that the conditional distribution is a Gaussian mixture with time varying mixing weights that depend on p lagged values of the series in a way that has a natural interpretation. Thus, similarly to the linear Gaussian AR process, and contrary to (at least most) other nonlinear AR models, the structure of stationary marginal distributions of order $p+1$ or smaller is fully known. Stationary marginal distributions of order higher than $p+1$ are not Gaussian mixtures and for them no explicit expressions are available. This need not be a drawback, however, because a process with all finite dimensional distributions being Gaussian mixtures (with constant mixing weights) cannot be ergodic, as we shall demonstrate in the paper. Despite this fact, the formulation of the model is based on the assumption of Gaussianity, and therefore we call the model a Gaussian Mixture AR (GMAR) model.

A practical convenience of having an explicit expression for the stationary marginal density is that one can use a nonparametric density estimate to examine the suitability of the GMAR model in advance and, after fitting a GMAR model to data, assess the fit by comparing the density implied by the model with the nonparametric estimate. Because the p–dimensional stationary distribution of the process is known the exact likelihood function can be constructed and used to obtain exact maximum likelihood (ML) estimates.

A further advantage, which also stems from the formulation of the model, is the specific form of the time varying mixing weights which appears very flexible. These convenient features are illusrated in our empirical example, which also demonstrates that the GMAR model can be a flexible alternative to previous mixture AR models and TAR and STAR models.

The rest of the paper is organized as follows. After discussing general mixture AR models, Section 2 presents the GMAR model along with a discussion of its properties, and a comparison to previous related models. Section 3 deals with issues of specification and evaluation of GMAR models as well as estimation of parameters by the method of maximum likelihood. Section 4 presents an empirical example with interest rate data, and Section 5 concludes. Two appendices provide some technical derivations and graphical illustrations of the employed mixing weights.

2 Models

2.1 Mixture autoregressive models

Let y_t $(t = 1, 2, \ldots)$ be the real-valued time series of interest, and let $\mathcal{F}_{t-1}$ denote the σ–algebra generated by $\{y_{t-j},\ j > 0\}$. We consider a mixture autoregressive model in which the conditional density function of y_t given its past, $f(\cdot \mid \mathcal{F}_{t-1})$, is of the form

$$f(y_t \mid \mathcal{F}_{t-1}) = \sum_{m=1}^{M} \alpha_{m,t} \frac{1}{\sigma_m} \phi\left(\frac{y_t - \mu_{m,t}}{\sigma_m}\right). \tag{1}$$

Here the (positive) mixing weights $\alpha_{m,t}$ are $\mathcal{F}_{t-1}$–measurable and satisfy $\sum_{m=1}^{M} \alpha_{m,t} = 1$ (for all t). Furthermore, $\phi(\cdot)$ denotes the density function of a standard normal random variable, $\mu_{m,t}$ is defined by

$$\mu_{m,t} = \varphi_{m,0} + \sum_{i=1}^{p} \varphi_{m,i} y_{t-i}, \qquad m = 1, \ldots, M, \tag{2}$$

and $\boldsymbol{\vartheta}_m = (\varphi_{m,0}, \boldsymbol{\varphi}_m, \sigma_m^2)$ with $\boldsymbol{\varphi}_m = (\varphi_{m,1}, \ldots, \varphi_{m,p})$ and $\sigma_m^2 > 0$ $(m = 1, \ldots, M)$ contain the unknown parameters introduced in the above equations. (By replacing p in (2) with p_m, the autoregressive orders in the component models could be allowed to vary; on the other hand, this can also be achieved by restricting some of the $\varphi_{m,i}$–coefficients in (2) to be zero.) As equation (2) indicates, the definition of the model also requires a specification of the initial values $y_{-p+1}, \ldots, y_0$. Different mixture autoregressive models are obtained by different specifications of the mixing weights. Section 2.3 provides a more detailed discussion of the various specifications proposed in the literature.

For further intuition we express the model (1)–(2) in a different format. Let $P_{t-1}(\cdot)$ signify the conditional probability of the indicated event given $\mathcal{F}_{t-1}$, and let ε_t be a sequence of independent standard normal random variables ($\varepsilon_t \sim NID(0,1)$) such that ε_t is independent of $\{y_{t-j},\ j > 0\}$. Furthermore, let $\boldsymbol{s}_t = (s_{t,1}, \ldots, s_{t,M})$ $(t = 1, 2, \ldots)$ be a sequence of (unobserved) M–dimensional random vectors such that, conditional on $\mathcal{F}_{t-1}$, $\boldsymbol{s}_t$ and ε_t are independent. The components of $\boldsymbol{s}_t$ are such that, for each t, exactly one of them takes the value one and others are equal to zero, with conditional probabilities $P_{t-1}(s_{t,m} = 1) = \alpha_{m,t}$, $m = 1, \ldots, M$. Now y_t can be expressed as

$$y_t = \sum_{m=1}^{M} s_{t,m}(\mu_{m,t} + \sigma_m \varepsilon_t) = \sum_{m=1}^{M} s_{t,m} \left(\varphi_{m,0} + \sum_{i=1}^{p} \varphi_{m,i} y_{t-i} + \sigma_m \varepsilon_t \right). \tag{3}$$

This formulation suggests that the mixing weights $\alpha_{m,t}$ can be thought of as probabilities that determine which one of the M autoregressive components of the mixture generates the next observation.

From (1)–(2) or (3) one immediately finds that the conditional mean and variance of y_t given $\mathcal{F}_{t-1}$ are

$$E[y_t \mid \mathcal{F}_{t-1}] = \sum_{m=1}^{M} \alpha_{m,t} \mu_{m,t} = \sum_{m=1}^{M} \alpha_{m,t} \left(\varphi_{m,0} + \sum_{i=1}^{p} \varphi_{m,i} y_{t-i} \right) \tag{4}$$

and

$$Var[y_t \mid \mathcal{F}_{t-1}] = \sum_{m=1}^{M} \alpha_{m,t} \sigma_m^2 + \sum_{m=1}^{M} \alpha_{m,t} \left(\mu_{m,t} - \left(\sum_{m=1}^{M} \alpha_{m,t} \mu_{m,t} \right) \right)^2. \tag{5}$$

These expressions apply for any specification of the mixing weights $\alpha_{m,t}$. The conditional mean is a weighted average of the conditional means of the M autoregressive components with weights generally depending on the past history of the process. The conditional variance also contains a similar weighted average of the conditional (constant) variances of the M autoregressive components but there is an additional additive term which depends on the variability of the conditional means of the component processes. This additional term makes the conditional variance nonconstant even if the mixing weights are nonrandom and constant over time.

2.2 The Gaussian Mixture Autoregressive (GMAR) model

The mixture autoregressive model considered in this paper is based on a particular choice of the mixing weights in (1). Using the parameters $\varphi_{m,0}$, $\boldsymbol{\varphi}_m = (\varphi_{m,1}, \ldots, \varphi_{m,p})$, and σ_m (see equation (1) or (3)) we first define the M auxiliary Gaussian AR(p) processes

$$\nu_{m,t} = \varphi_{m,0} + \sum_{i=1}^{p} \varphi_{m,i}\nu_{m,t-i} + \sigma_m \varepsilon_t, \qquad m = 1, \ldots, M,$$

where the autoregressive coefficients $\boldsymbol{\varphi}_m$ are assumed to satisfy

$$\varphi_m(z) = 1 - \sum_{i=1}^{p} \varphi_{m,i} z^i \neq 0 \quad \text{for } |z| \leq 1, \qquad m = 1, \ldots, M. \tag{6}$$

This condition implies that the processes $\nu_{m,t}$ are stationary and also that each of the component models in (3) satisfies the usual stationarity condition of the conventional linear AR(p) model.

To enhance the flexibility of the model our definition of the mixing weights $\alpha_{m,t}$ also involves a choice of a lag length $q \geq p$. As will be discussed later, setting $q = p$ appears a convenient first choice. Set $\boldsymbol{\nu}_{m,t} = (\nu_{m,t}, \ldots, \nu_{m,t-q+1})$ and $\mathbf{1}_q = (1, \ldots, 1)$ $(q \times 1)$, and let $\mu_m \mathbf{1}_q$ and $\boldsymbol{\Gamma}_{m,q}$ signify the mean vector and covariance matrix of $\boldsymbol{\nu}_{m,t}$ $(m = 1, \ldots, M)$. Here $\mu_m = \varphi_{m,0}/\varphi_m(1)$ and each $\boldsymbol{\Gamma}_{m,q}$, $m = 1, \ldots, M$, has the familiar form of being

a $q \times q$ symmetric Toeplitz matrix with $\gamma_{m,0} = Cov[\nu_{m,t}, \nu_{m,t}]$ along the main diagonal, and $\gamma_{m,i} = Cov[\nu_{m,t}, \nu_{m,t-i}]$, $i = 1, \ldots, q-1$, on the diagonals above and below the main diagonal. For the dependence of the covariance matrix $\boldsymbol{\Gamma}_{m,q}$ on the parameters $\boldsymbol{\varphi}_m$ and σ_m, see Reinsel (1997, Sec. 2.2.3). The random vector $\boldsymbol{\nu}_{m,t}$ follows the q–dimensional multivariate normal distribution with density

$$\mathsf{n}_q\left(\boldsymbol{\nu}_{m,t}; \boldsymbol{\vartheta}_m\right) = (2\pi)^{-q/2} \det(\boldsymbol{\Gamma}_{m,q})^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\nu}_{m,t} - \mu_m \mathbf{1}_q\right)' \boldsymbol{\Gamma}_{m,q}^{-1}\left(\boldsymbol{\nu}_{m,t} - \mu_m \mathbf{1}_q\right)\right\}. \quad (7)$$

Now set $\boldsymbol{y}_{t-1} = (y_{t-1}, \ldots, y_{t-q})$ $(q \times 1)$, and define the mixing weights $\alpha_{m,t}$ as

$$\alpha_{m,t} = \frac{\alpha_m \mathsf{n}_q\left(\boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_m\right)}{\sum_{n=1}^{M} \alpha_n \mathsf{n}_q\left(\boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_n\right)}, \quad (8)$$

where the $\alpha_m \in (0,1)$, $m = 1, \ldots, M$, are unknown parameters satisfying $\sum_{m=1}^{M} \alpha_m = 1$. (Clearly, the coefficients $\alpha_{m,t}$ are measurable functions of $\boldsymbol{y}_{t-1} = (y_{t-1}, \ldots, y_{t-q})$ and satisfy $\sum_{m=1}^{M} \alpha_{m,t} = 1$ for all t.) We collect the unknown parameters to be estimated in the vector $\boldsymbol{\theta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_M, \alpha_1, \ldots, \alpha_{M-1})$ $((M(p+3)-1) \times 1)$; the coefficient α_M is not included due to the restriction $\sum_{m=1}^{M} \alpha_m = 1$. Equations (1), (2), and (8) (or (3) and (8)) define the Gaussian Mixture Autoregressive model or the GMAR model. We use the abbreviation GMAR(p, q, M), or simply GMAR(p, M) when $q = p$, when the autoregressive order and number of component models need to be emphasized.

A major motivation for specifying the mixing weights as in (8) is theoretical attractiveness. We shall discuss this point briefly before providing an intuition behind this particular choice of the mixing weights. First note that the conditional distribution of $\boldsymbol{y}_t$ given $\mathcal{F}_{t-1}$ only depends on $\boldsymbol{y}_{t-1}$, implying that the process $\boldsymbol{y}_t$ is Markovian. This fact is formally stated in the following theorem which shows that there exists a choice of initial values $\boldsymbol{y}_0$ such that $\boldsymbol{y}_t$ is a stationary and ergodic Markov chain. An explicit expression for the stationary distribution is also provided. As will be discussed in more detail shortly, it is quite exceptional among mixture autoregressive models or other related nonlinear autoregressive models such as TAR models or STAR models that the stationary

distribution is fully known. As our empirical examples demonstrate, this result is also practically very convenient.

The proof of the following theorem can be found in Appendix A.

Theorem 1. *Consider the GMAR process y_t generated by (1), (2), and (8) (or, equivalently, (3) and (8)) with condition (6) satisfied and $q \geq p$. Then $\boldsymbol{y}_t = (y_t, \ldots, y_{t-q+1})$ $(t = 1, 2, \ldots)$ is a Markov chain on $\mathbb{R}^q$ with a stationary distribution characterized by the density*

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m \mathsf{n}_q \left(\boldsymbol{y}; \boldsymbol{\vartheta}_m\right). \tag{9}$$

Moreover, $\boldsymbol{y}_t$ is ergodic.

Thus, the stationary distribution of $\boldsymbol{y}_t$ is a mixture of M multinormal distributions with constant mixing weights α_m that appear in the time varying mixing weights $\alpha_{m,t}$ defined in (8). An immediate consequence of this result is that all moments of the stationary distribution exist and are finite. In the proof of Theorem 1 it is also demonstrated that the stationary distribution of the $(q+1)$–dimensional random vector $(y_t, \boldsymbol{y}_{t-1})$ is a Gaussian mixture with density of the same form as in (9) or, specifically, $\sum_{m=1}^{M} \alpha_m \mathsf{n}_{q+1} \left((y, \boldsymbol{y}); \boldsymbol{\vartheta}_m\right)$ with an explicit form of the density function $\mathsf{n}_{q+1} \left((y, \boldsymbol{y}); \boldsymbol{\vartheta}_m\right)$ given in the proof of Theorem 1. It is straightforward to check that the marginal distributions of this Gaussian mixture belong to the same family (this can be seen by integrating the relevant components of $(y, \boldsymbol{y})$ out of the density). It may be worth noting, however, that this does not hold for higher dimensional realizations so that the stationary distribution of $(y_{t+1}, y_t, \boldsymbol{y}_{t-1})$, for example, is not a Gaussian mixture. This fact was already pointed out by Glasbey (2001) who considered a first order version of the same model (i.e., the case $q = p = 1$) by using a slightly different formulation. Glasbey (2001) did not discuss higher order models explicitly and he did not establish ergodicity obtained in Theorem 1. Interestingly, in the discussion section of his paper he mentions that a drawback of his model is that

joint and conditional distributions in higher dimensions are not Gaussian mixtures. It would undoubtedly be convenient in many respects if all finite dimensional distributions of a process were Gaussian mixtures (with constant mixing weights) but an undesirable implication would then be that ergodicity could not hold true. We demonstrate this in Appendix A by using a simple special case.

A property that makes our GMAR model different from most, if not nearly all, previous nonlinear autoregressive models is that its stationary distribution obtained in Theorem 1 is fully known (a few rather simple first order examples, some of which also involve Gaussian mixtures, can be found in Tong (2011, Section 4.2)). As illustrated in Section 4, a nonparametric estimate of the stationary density of y_t can thus be used (as one tool) to assess the need of a mixture model and the fit of a specified GMAR model. It is also worth noting that in order to prove Theorem 1 we are not forced to restrict the parameter space over what is used to define the model and the parameter space is defined by familiar conditions that can readily be checked. This is in contrast with similar previous results where conditions for stationarity and ergodicity are typically only sufficient and restrict the parameter space or, if sharp, cannot be verified without resorting to simulation or numerical methods (see, e.g., Cline (2007)). It is also worth noting that Theorem 1 can be proved in a much more straightforward manner than most of its previous counterparts. In particular, we do not need to apply the so-called drift criterion which has been a standard tool in previous similar proofs (see, e.g., Saikkonen (2007), Bec, Rahbek, and Shephard (2008), and Meyn and Tweedie (2009)). On the other hand, our GMAR model assumes that the components of the mixture satisfy the usual stationarity condition of a linear AR(p) model which is not required in all previous models. For instance, Bec, Rahbek, and Shephard (2008) prove an analog of Theorem 1 with $M = 2$ without any restrictions on the autoregressive parameters of one of the component models (see also Cline (2007)). Note also that the favorable results of Theorem 1 require that $q \geq p$. They

are not obtained if $q < p$ and, therefore, we will not consider this case (the role of the lag length q will be discussed more at the end of this section).

Unless otherwise stated, the rest of this section assumes the stationary version of the process. According to Theorem 1, the parameter α_m $(m = 1, \ldots, M)$ then has an immediate interpretation as the unconditional probability of the random vector $\boldsymbol{y}_t = (y_t, \ldots, y_{t-q+1})$ being generated from a distribution with density $\mathsf{n}_q(\boldsymbol{y}; \boldsymbol{\vartheta}_m)$, that is, from the mth component of the Gaussian mixture characterized in (9). As a direct consequence, α_m $(m = 1, \ldots, M)$ also represents the unconditional probability of the component y_t being generated from a distribution with density $\mathsf{n}_1(y; \boldsymbol{\vartheta}_m)$ which is the mth component of the (univariate) Gaussian mixture density $\sum_{m=1}^{M} \alpha_m \mathsf{n}_1(y; \boldsymbol{\vartheta}_m)$ where $\mathsf{n}_1(y; \boldsymbol{\vartheta}_m)$ is the density function of a normal random variable with mean μ_m and variance $\gamma_{m,0}$. Furthermore, it is straightforward to check that α_m also represents the unconditional probability of (the scalar) y_t being generated from the mth autoregressive component in (3) whereas $\alpha_{m,t}$ represents the corresponding conditional probability $P_{t-1}(s_{t,m} = 1) = \alpha_{m,t}$. This conditional probability depends on the (relative) size of the product $\alpha_m \mathsf{n}_q(\boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_m)$, the numerator of the expression defining $\alpha_{m,t}$ (see (8)). The latter factor of this product, $\mathsf{n}_q(\boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_m)$, can be interpreted as the likelihood of the mth autoregressive component in (3) based on the observation $\boldsymbol{y}_{t-1}$. Thus, the larger this likelihood is the more likely it is to observe y_t from the mth autoregressive component. However, the product $\alpha_m \mathsf{n}_q(\boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_m)$ is also affected by the former factor α_m or the weight of $\mathsf{n}_q(\boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_m)$ in the stationary mixture distribution of $\boldsymbol{y}_{t-1}$ (evaluated at $\boldsymbol{y}_{t-1}$; see (9)). Specifically, even though the likelihood of the mth autoregressive component in (3) is large (small) a small (large) value of α_m attenuates (amplifies) its effect so that the likelihood of observing y_t from the mth autoregressive component can be small (large). This seems intuitively natural because a small (large) weight of $\mathsf{n}_q(\boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_m)$ in the stationary mixture distribution of $\boldsymbol{y}_{t-1}$ means that observations cannot be generated by the mth autoregressive component

too frequently (too infrequently).

It may also be noted that the probabilities $\alpha_{m,t}$ are formally similar to posterior model probabilities commonly used in Bayesian statistics (see, e.g., Sisson (2005) or Del Negro and Schorfheide (2011)). An obvious difference is that in our model the parameters $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_M$ are treated as fixed so that no prior distributions are specified for them. Therefore, the marginal likelihood used in the Bayesian setting equals the density $\mathsf{n}_q(\boldsymbol{y}; \boldsymbol{\vartheta}_m)$ associated with the mth model. However, as α_m only requires knowledge of the stationary distribution of the process, not observed data, it can be thought of as one's prior probability of the observation y_t being generated from the mth autoregressive component in (3). When observed data $\mathcal{F}_{t-1}$ (or $\boldsymbol{y}_{t-1}$) are available one can compute $\alpha_{m,t}$, an analog of the corresponding posterior probability, which provides more accurate information about the likelihood of observing y_t from the mth autoregressive component in (3). Other things being equal a decrease (increase) in the value of α_m decreases (increases) the value of $\alpha_{m,t}$. That the stationary distribution of the process explicitly affects the conditional probability of observing y_t from the mth autoregressive component appears intuitively natural regardless of whether one interprets α_m as a prior probability or a mixing weight in the stationary distribution.

Using the facts that the density of $(y_t, \boldsymbol{y}_{t-1})$ is $\sum_{m=1}^{M} \alpha_m \mathsf{n}_{q+1}\left((y_t, \boldsymbol{y}_{t-1}); \boldsymbol{\vartheta}_m\right)$ and that of y_t is $\sum_{m=1}^{M} \alpha_m \mathsf{n}_1(y; \boldsymbol{\vartheta}_m)$ we can obtain explicit expressions for the mean, variance, and first q autocovariances of the process y_t. With the notation introduced in equation (7) we can express the mean as

$$\mu \overset{def}{=} E[y_t] = \sum_{m=1}^{M} \alpha_m \mu_m$$

and the variance and first q autocovariances as

$$\gamma_j \overset{def}{=} Cov[y_t, y_{t-j}] = \sum_{m=1}^{M} \alpha_m \gamma_{m,j} + \sum_{m=1}^{M} \alpha_m (\mu_m - \mu)^2, \qquad j = 0, 1, \ldots, q.$$

Using these autocovariances and Yule-Walker equations (see, e.g., Box, Jenkins, and Rein-

sel (2008, p. 59)) one can derive the parameters of the linear AR(q) process that best approximates a GMAR(p, q, M) process. As higher dimensional stationary distributions are not Gaussian mixtures and appear difficult to handle no simple expressions are available for autocovariances at lags larger than q.

The preceding discussions also illuminate the role of the lag length q ($\geq p$). The autoregressive order p (together with the other model parameters) determines the dependence structure of the component models as well as the mean, variance, and (the first p) autocovariances of the process y_t. On the other hand, the parameter q determines how many lags of y_t affect $\alpha_{m,t}$, the conditional probability of y_t being generated from the mth autoregressive component. While the case $q = p$ may often be appropriate, choosing $q > p$ allows for the possibility that the autoregressive order is (relatively) small compared with the mechanism governing the choice of the component model that generates the next observation. As already indicated, the case $q < p$ would be possible but not considered because then the convenient theoretical properties in Theorem 1 are not obtained.

Note also that q determines (through $\alpha_{m,t}$) how many lagged observations affect the conditional variance of the process (see (5)). Thus, the possibility $q > p$ may be useful when the autoregressive order is (relatively) small compared with the number of lags needed to allow for conditional heteroskedasticity. For instance, in the extreme case $p = 0$ (but $q > 0$), the GMAR process generates observations that are uncorrelated but with time-varying conditional heteroskedasticity.

2.3 Discussion of models

In this section, we discuss the GMAR model in relation to other nonlinear autoregressive models introduced in the literature. If the mixing weights are assumed constant over time the general mixture autoregressive model (1) reduces to the MAR model studied by Wong and Li (2000). The MAR model, in turn, is a generalization of a model considered

by Le, Martin, and Raftery (1996). Wong and Li (2001b) considered a model with time-varying mixing weights. In their Logistic MAR (LMAR) model, only two regimes are allowed, with a logistic transformation of the two mixing weights, $\log(\alpha_{1,t}/\alpha_{2,t})$, being a linear function of past observed variables. Related two-regime mixture models with time-varying mixing weights were also considered by Gourieroux and Robert (2006) and Bec, Rahbek, and Shephard (2008). Lanne and Saikkonen (2003) considered a mixture AR model in which multiple regimes are allowed (see also Zeevi, Meir, and Adler (2000) and Carvalho and Tanner (2005) in the engineering literature). Lanne and Saikkonen (2003) specify the mixing weights as

$$\alpha_{m,t} = \begin{cases} 1 - \Phi((y_{t-d} - c_1)/\sigma_\eta), & m = 1, \\ \Phi((y_{t-d} - c_{m-1})/\sigma_\eta) - \Phi((y_{t-d} - c_m)/\sigma_\eta), & m = 2, \ldots, M-1, \\ \Phi((y_{t-d} - c_{M-1})/\sigma_\eta), & m = M, \end{cases} \tag{10}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable, $d \in \mathbb{Z}_+$ is a delay parameter, and the real constants $c_1 < \cdots < c_{M-1}$ are location parameters. In their model, the probabilities determining which of the M autoregressive components the next observation is generated from depend on the location of y_{t-d} relative to the location parameters $c_1 < \cdots < c_{M-1}$. Thus, when $p = d = 1$ a similarity between the mixing weights in the model of Lanne and Saikkonen (2003) and in the GMAR model is that the value of y_{t-1} gives indication concerning which regime will generate the next observation. However, even in this case the functional forms of the mixing weights and their interpretation are rather different.

An interesting two-regime mixture model with time-varying mixing weights was recently introduced by Dueker, Sola, and Spagnolo (2007) (see also Dueker, Psaradakis, Sola, and Spagnolo (2011) for a multivariate extension).[1] In their model, the mixing

[1] According to the authors their model belongs to the family of STAR models and this interpretation is indeed consistent with the initial definition of the model which is based on equations (1)–(4) in Dueker,

weights are specified as

$$\alpha_{1,t} = \frac{\Phi\left((c_1 - \varphi_{1,0} - \boldsymbol{\varphi}_1'\boldsymbol{y}_{t-1})/\sigma_1\right)}{\Phi\left((c_1 - \varphi_{1,0} - \boldsymbol{\varphi}_1'\boldsymbol{y}_{t-1})/\sigma_1\right) + \left[1 - \Phi\left((c_1 - \varphi_{2,0} - \boldsymbol{\varphi}_2'\boldsymbol{y}_{t-1})/\sigma_2\right)\right]} \tag{11}$$

and $\alpha_{2,t} = 1 - \alpha_{1,t}$. Here c_1 is interpreted as a location parameter similar to that in the model of Lanne and Saikkonen (2003). However, similarly to our model the mixing weights are determined by lagged values of the observed series and the autoregressive parameters of the component models. The same number of lags is assumed in both the mixing weights and the autoregressive components (or that $q = p$ in the notation of the present paper). Nevertheless, the interpretation of the mixing weights is closer to that of our GMAR model than is the case for the model of Lanne and Saikkonen (2003). The probability that the next observation is generated from the first or second regime is determined by the locations of the conditional means of the two autoregressive components from the location parameter c_1 whereas in the GMAR model this probability is determined by the stationary densities of the two component models and their weights in the stationary mixture distribution. The functional form of the mixing weights of Dueker, Sola, and Spagnolo (2007) is also similar to ours except that instead of the Gaussian density function used in our GMAR model, Dueker, Sola, and Spagnolo (2007) have the Gaussian cumulative distribution function.

The GMAR model is also related to threshold and smooth transition type nonlinear models. In particular, the conditional mean function $E[y_t \mid \mathcal{F}_{t-1}]$ of our GMAR model is similar to those of a TAR or a STAR model (see, e.g., Tong (1990) and Teräsvirta (1994)). In a basic two-regime TAR model, whether a threshold variable (a lagged value of y_t) exceeds a certain threshold or not determines which of the two component models

Sola, and Spagnolo (2007). However, we have chosen to treat the model as a mixture model because the likelihood function used to fit the model to data is determined by conditional density functions that are of the mixture form (1). These conditional density functions are given in equation (7) of Dueker, Sola, and Spagnolo (2007) but their connection to the aforementioned equations (1)–(4) is not clear to us.

describes the generating mechanism of the next observation. The threshold and threshold variable are analogous to the location parameter c_1 and the variable y_{t-d} in the mixing weights used in the two-regime ($M = 2$) mixture model of Lanne and Saikkonen (2003) (see (10)). In a STAR model, one gradually moves from one component model to the other as the threshold (or transition) variable changes its value. In a GMAR model, the mixing weights follow similar smooth patterns. A difference to STAR models is that while the mixing weights of the GMAR model vary smoothly, the next observation is generated from one particular AR component whose choice is governed by these mixing weights. In a STAR model, the generating mechanism of the next observation is described by a convex combination of the two component models. This difference is related to the fact that the conditional distribution of the GMAR model is of a different type than the conditional distribution of the STAR (or TAR) model which is not a mixture distribution. This difference is also reflected in differences between the conditional variances associated with the GMAR model and STAR (or TAR) models.

To illustrate the preceding discussion and the differences between alternative mixture AR models, Figure 6 in Appendix B depicts the mixing weights $\alpha_{1,t}$ of the GMAR model and some of the alternative models with certain parameter combinations. A detailed discussion of this figure is provided in Appendix B, so here we only summarize some of the main points. For presentational clarity, the figure only concerns first-order models with two regimes, and how $\alpha_{1,t}$ changes as a function of y_{t-1}. In this case, the mixture models of Wong and Li (2001b) and Lanne and Saikkonen (2003) can only produce mixing weights with smooth, monotonically increasing patterns (comparable to those of a transition function of a basic logistic two-regime STAR model). In these models, nonmonotonic mixing weights can be obtained when there are more than two regimes. In the model of Dueker, Sola, and Spagnolo (2007), the mixing weights can be nonmonotonic even in the case of two regimes, although the range of available shapes appears rather limited. In

contrast to these previous models, with suitable parameter values the GMAR model can produce both monotonic and nonmonotonic mixing weights of various shapes. Further details can be found in Appendix B, but the overall conclusion is that our GMAR model appears more flexible in terms of the form of mixing weights than the aforementioned previous mixture models.

Finally, we also note that the MAR model with constant mixing weights (Wong and Li, 2000) is a special case of the Markov switching AR model (see, e.g., Hamilton (1994, Ch. 22)). In the context of equation (3) (the basic form of) the Markov switching AR model corresponds to the case where the sequence $\boldsymbol{s}_t$ forms a (time-homogeneous) Markov chain whose transition probabilities correspond to the mixing weights. Thus, the sequence $\boldsymbol{s}_t$ is dependent whereas in the MAR model of Wong and Li (2000) it is independent in time. In time-inhomogeneous versions of the Markov switching AR model (see, e.g., Diebold, Lee, and Weinbach (1994) and Filardo (1994)) the transition probabilities depend on lagged values of the observed time series and are therefore analogs of time-varying mixing weights. However, even in this case the involved Markov chain structure of the sequence s_t makes Markov switching AR models rather different from the mixture AR models considered in this paper.

3 Model specification, estimation, and evaluation

3.1 Specification

We next discuss some general aspects of building a GMAR model. A natural first step is to consider whether a conventional linear Gaussian AR model provides an adequate description of the data generation process. Thus, one finds an AR(p) model that best describes the autocorrelation structure of the time series, and checks whether residual diagnostics show signs of non-Gaussianity and possibly also of conditional heteroskedasticity. At this

point also the graph of the series and a nonparametric estimate of the density function may be useful. The former may indicate the presense of multiple regimes, whereas the latter may show signs of multimodality.

If a linear AR model is found inadequate, specifying a GMAR(p, q, M) model requires the choice of the number of component models M, autoregressive order p, and the lag length q. A nonparametric estimate of the density function of the observed series may give an indication of how many mixture components are needed. One should, however, be conservative with the choice of M, because if the number of component models is chosen too large then some parameters of the model are not identified. Therefore, a two component model ($M = 2$) is a good first choice. If an adequate two component model is not found, only then should one proceed to a three component model and, if needed, consider even more components.

The initial choice of the autoregressive order p can be based on the order chosen for the linear AR model. Also, setting $q = p$ appears a good starting point. Again, one should favor parsimonious models, and initially try a smaller p if the order selected for the linear AR model appears large. One reason for this practice is that if the true model is a GMAR(p, M) model then an overspecified GMAR$(p + 1, M)$ model will be misspecified. (The source of misspecification here is an overly large q, namely if the true model is a GMAR(p, q, M), then an overspecified GMAR$(p, \tilde{q}, M)$ model with $\tilde{q} > q$ will be misspecified.) After finding an adequate GMAR(p, M) model, one may examine for possible simplifications obtained by parameter restrictions. For instance, some of the parameters may be restricted to be equal in each component or evidence for a smaller autoregressive order may be found, leading to a model with $q > p$.

3.2 Estimation

After an initial candidate specification (or specifications) is (are) chosen, the parameters of a GMAR model can be estimated by the method of maximum likelihood. As the stationary distribution of the GMAR process is known it is even possible to make use of initial values and construct the exact likelihood function and obtain exact ML estimates, as already discussed by Glasbey (2001) in the first order case. Assuming the observed data is $y_{-q+1}, \ldots, y_0, y_1, \ldots, y_T$ and stationary initial values the log-likelihood function takes the form

$$\begin{aligned} l_T(\boldsymbol{\theta}) &= \log\left(\sum_{m=1}^{M} \alpha_m \mathsf{n}_q(\boldsymbol{y}_0; \boldsymbol{\vartheta}_m)\right) \\ &\quad + \sum_{t=1}^{T} \log\left(\sum_{m=1}^{M} \alpha_{m,t}(\boldsymbol{\theta}) \left(2\pi\sigma_m^2\right)^{-1/2} \exp\left(-\frac{\left(y_t - \mu_{m,t}(\boldsymbol{\vartheta}_m)\right)^2}{2\sigma_m^2}\right)\right), \end{aligned} \quad (12)$$

where dependence of the mixing weights $\alpha_{m,t}$ and the conditional expectations $\mu_{m,t}$ of the component models on the parameters is made explicit (see (8) and (2)). Maximizing the log-likelihood function $l_T(\boldsymbol{\theta})$ with respect to the parameter vector $\boldsymbol{\theta}$ yields the ML estimate denoted by $\hat{\boldsymbol{\theta}}$ (a similar notation is used for components of $\hat{\boldsymbol{\theta}}$). Here we have assumed that the initial values in the vector $\boldsymbol{y}_0$ are generated by the stationary distribution. If this assumption seems inappropriate one can condition on initial values and drop the first term on the right hand side of (12). For reasons of identification the inequality restrictions $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_M$ are imposed on the parameters α_m ($m = 1, \ldots, M$, $\alpha_M = 1 - \sum_{m=1}^{M-1} \alpha_m$).

In our empirical examples we have used the optimization algorithms in the cmlMT library of Gauss to maximize the likelihood function or its conditional version. Especially the Newton-Raphson algorithm in that library seemed to work quite well but one could alternatively follow Wong and Li (2001b) and use the EM algorithm. As usual in nonlinear optimization, good initial values improve the performance of the estimation algorithm. One way to obtain initial values is to make use of the fact that the

$(q+1)$–dimensional stationary distribution of the observed process is characterized by the density $\sum_{m=1}^{M} \alpha_m \mathsf{n}_{q+1}\left((y_t, \boldsymbol{y}_{t-1}); \boldsymbol{\vartheta}_m\right)$. Rough initial estimates for the parameters of the model can thus be obtained by maximizing the (quasi)likelihood function based on the (incorrect) assumption that the observations $(y_t, \boldsymbol{y}_{t-1})$, $t = 1, \ldots, T$, are independent and identically distributed with density $\sum_{m=1}^{M} \alpha_m \mathsf{n}_{q+1}\left((y_t, \boldsymbol{y}_{t-1}); \boldsymbol{\vartheta}_m\right)$. This maximization requires numerical methods and, although it appears simpler than the maximization of the log-likelihood function (12) or its conditional version, it can be rather demanding if the sample size or the dimension $q+1$ is large. A simpler alternative is to make use of the one dimensional stationary distribution characterized by the density $\sum_{m=1}^{M} \alpha_m \mathsf{n}_1\left(y_t; \boldsymbol{\vartheta}_m\right)$ which depends on the expectations μ_m, variances $\gamma_{m,0}$, and mixing weights α_m $(m = 1, \ldots, M)$. Rough initial estimates for these parameters can thus be obtained by maximizing the (quasi)likelihood function based on the (incorrect) assumption that the observed series y_t, $t = -q+1, \ldots, T$, is independent and identically distributed with density $\sum_{m=1}^{M} \alpha_m \mathsf{n}_1\left(y_t; \boldsymbol{\vartheta}_m\right)$. Our experience on the estimation of GMAR models indicates that it is especially useful to have good initial values for the (unequal) intercept terms $\varphi_{m,0}$ $(m = 1, \ldots, M)$. Once initial values for the expectations μ_m are available one can compute initial values for the intercept terms $\varphi_{m,0}$ by using the formula $\varphi_{m,0} = \varphi_m(1)\mu_m$ with a chosen value of $\varphi_m(1)$. For instance, one can (possibly incorrectly) assume that the autoregressive polynomials $\varphi_m(z)$ are identical for all m and estimate $\varphi_m(1)$ for all m by using the autoregressive polynomial of a linear autoregressive model fitted to the series. Using these initial values for the autoregressive parameters $\varphi_{m,0}$ and $\boldsymbol{\varphi}_m$, one can further obtain rough initial values for the error variances σ_m^2 and thereby for all parameters of the model. Finding out the usefulness of these approaches in initial estimation requires further investigation but, according to our limited experience, they can be helpful.

Concerning the asymptotic properties of the ML estimator, Dueker, Sola, and Spagnolo

(2007) show that, under appropriate regularity conditions, the usual results of consistency and asymptotic normality hold in their mixture model. The conditions they use are of general nature and using the ergodicity result of Theorem 1 along with similar "high level" conditions it is undoubtedly possible to show the consistency and asymptotic normality of the ML estimator in our GMAR model as well. However, we prefer to leave a detailed analysis of this issue for future work. In our empirical examples we treat the ML estimator $\hat{\boldsymbol{\theta}}$ as approximately normally distributed with mean vector $\boldsymbol{\theta}$ and covariance matrix the inverse of the Fisher information matrix $E\left[-\partial^2 l_T\left(\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\right]$ that can be estimated by inverting the observed information matrix $-\partial^2 l_T(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. It is worth noting that the aforementioned results require a correct specification of the number of autoregressive components M. In particular, standard likelihood-based tests are not applicable if the number of component models is chosen too large because then some parameters of the model are not identified. This particularly happens when one tests for the number of component models. For further discussion of this issue, see Dueker et al. (2007, 2011) and the references therein. In our model, the situation is similar with respect to the lag length q. If q is chosen too large, the model becomes misspecified, and for this reason standard likelihood-based tests cannot be used to choose q.

3.3 Evaluation

Having estimated a few candidate models, one must check their adequacy and choose the best fitting GMAR(p, q, M) model. As mentioned above, standard likelihood-based tests cannot be used to test for the number of component models M or for the lag length q. Instead of trying to develop proper test procedures for these purposes, we take a pragmatic approach and propose the use of residual-based diagnostics and information criteria (AIC and BIC) to select a model. In practice, this is often how model selection is done in other nonlinear models as well (cf., e.g., Teräsvirta, Tjøstheim, and Granger (2010, Ch. 16); for

instance, often the choice of a lag length to be used in a threshold/transition variable is done in a somewhat informal manner). When M and q are (correctly) chosen, standard likelihod-based inference can be used to choose the autoregressive order p (which can vary from one component model to the other).

In mixture models, care is needed when residual-based diagnostics are used to evaluate fitted models. The reason is that residuals with conventional properties are not readily available. This can be seen from the formulation of the GMAR model in equation (3) which shows that, due to the presence of the unobserved variables $s_{t,m}$, an empirical counterpart of the error term ε_t cannot be straightforwardly computed. A more elaborate discussion of this can be found in Kalliovirta (2012). Making use of ideas put forth by Smith (1985), Dunn and Smyth (1996), Palm and Vlaar (1997), and others, Kalliovirta (2012) proposes to use so-called quantile residuals instead of conventional (Pearson) residuals in mixture models (note that quantile residuals have also been called by other names such as normalized residuals and normal forecast transformed residuals). Quantile residuals are defined by two transformations. Assuming correct specification, the first one (the so-called probability integral transformation) uses the estimated conditional cumulative distribution function implied by the specified model to transform the observations into approximately independent uniformly distributed random variables. In the second transformation the inverse of the cumulative distribution function of the standard normal distribution is used to get variables that are approximately independent with standard normal distribution. Based on these 'two-stage' quantile residuals Kalliovirta (2012) proposes tests that can be used to check for autocorrelation, conditional heteroskedasticity, and non-normality in quantile residuals. These tests correctly allow for the uncertainty caused by parameter estimation so that, under correct specification, the obtained p–values are asymptotically valid. These are the residual-based diagnostic tests we use in our empirical application along with associated graphical tools to evaluate a fitted model.

4 Empirical example

4.1 A GMAR model of the Euro–U.S. interest rate differential

To illustrate how the GMAR model works in practice we present an example with interest rate data. Interest rate series are typically highly persistent and exhibit nonlinear behavior possibly due to regime switching dynamics. Consequently, various regime switching models have previously been used in modelling interest rate series, see for example Garcia and Perron (1996), Enders and Granger (1998), and Dueker et al. (2007, 2011). Our data, retrieved from OECD Statistics, consists of the monthly difference between the Euro area and U.S. long-term government bond yields from January 1989 to December 2009, a period that also contains the recent turbulences of the financial crisis since 2008 (in a small out-of-sample forecasting exercise we also use observations till September 2011).[2] This series, also referred to as the *interest rate differential*, is depicted in Figure 1 (left panel, solid line). Interest rate differential is a variable that has been of great interest in economics, although in empirical applications it has mostly been used in multivariate contexts together with other relevant variables, especially the exchange rate between the considered currencies (see, e.g., Chinn (2006) and the references therein). Our empirical example sheds light on the time series properties of the interest rate differential between the Euro area and U.S. long-term government bond yields, which may be useful if this variable is used in a more demanding multivariate modeling exercise.

Following the model-building strategy described in Section 3, we now consider the interest-rate differential series shown in Figure 1. (Estimation and all other computations

[2]The data series considered is $i_{EUR} - i_{USA}$, where i_{EUR} and i_{USA} are yields of government bonds with 10 years maturity, as calculated by the ECB and the Federal Reserve Board. Prior to 2001, the Euro area data refer to EU11 (Belgium, Germany, Ireland, Spain, France, Italy, Luxembourg, the Netherlands, Austria, Portugal, and Finland), from 2001 to 2006 to EU12 (EU11 plus Greece), and from January 2007 onwards to EU13 (EU12 plus Slovenia).

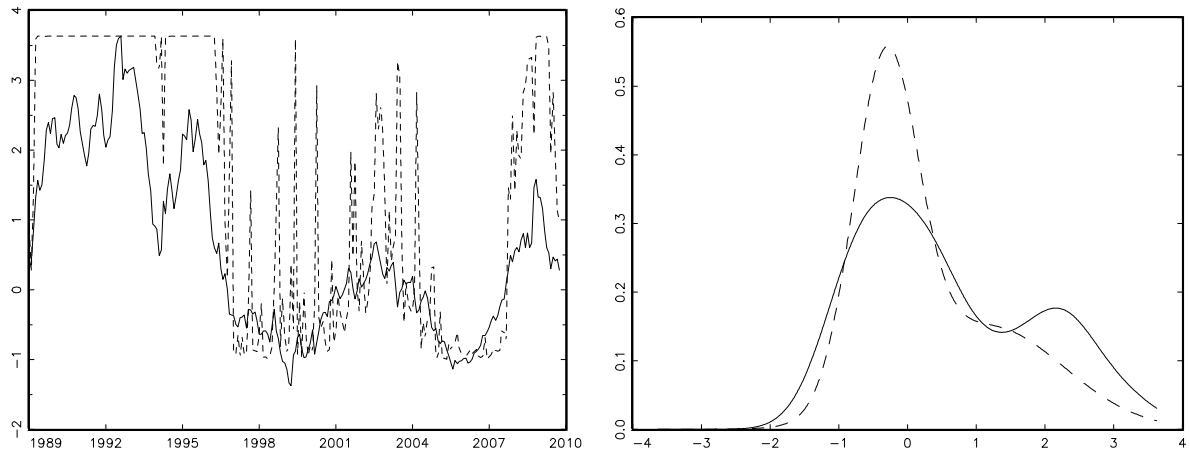

Figure 1: **Left panel:** Interest rate differential between the Euro area and the U.S. (solid line), and scaled mixing weights based on the estimates of the restricted GMAR(2,2) model in Table 1 (dashed line). The scaling is such that $\hat{\alpha}_{1,t} = \max y_t$, when $\hat{\alpha}_{1,t} = 1$, and $\hat{\alpha}_{1,t} = \min y_t$, when $\hat{\alpha}_{1,t} = 0$. **Right panel:** A kernel density estimate of the observations (solid line) and mixture density implied by the same GMAR(2,2) model as in the left panel (dashed line).

are carried out using GAUSS; the program codes are available upon request from the first author.) Of linear AR models, AR(4) was deemed to have the best fit. (The AIC and BIC suggested linear AR(2) and AR(5) models when the considered maximum order was eight; the AR(2) model had remaining autocorrelation in the residuals whereas, in terms of residual diagnostics, the more parsimonious AR(4) appeared equally good as AR(5).) Table 1 (leftmost column) reports parameter estimates for the linear AR(4) model along with the the values of AIC and BIC and (quantile) residual based tests of normality, autocorrelation, and conditional heteroskedasticity (brief descriptions of these tests are provided in the notes under Table 1, for further details see Kalliovirta (2012); for the Gaussian AR(4) model, quantile residuals are identical to conventional residuals). The AR(4) model appears adequate in terms of residual autocorrelation, but the tests for normality and conditional heteroskedasticity clearly reject it. In addition, the kernel density estimate of the original series depicted in Figure 1 (right panel, solid line) similarly suggests clear departures from normality (the estimate is bimodal, with mode -0.18 and a local mode 2.2), indicating that linear Gaussian AR models may be inappropriate.

Having found linear Gaussian AR models inadequate, of the GMAR models we first tried an unrestricted GMAR$(2, 2)$ specification. Two AR components seems to match with the graph of the series, where two major levels can be seen, as well as with the bimodal expression of the kernel density estimate (see Figure 1, right panel, solid line). According to (quantile) residual diagnostics (not reported), the unrestricted GMAR$(2, 2)$ specification turned out to be adequate but, as the AR polynomials in the two components seemed to be very close to each other, we restricted them to be the same (this restriction was not rejected by the LR test, which had p–value 0.61). Estimation results for the restricted GMAR(2,2) model are presented in Table 1 (for this series, all estimation and test results based on the exact likelihood and the conditional likelihood were quite close to each other, and the latter yielded slightly more accurate forecasts (see Section 4.4 below),

Table 1: Estimated AR, GMAR, and LMAR models (left panel) and means and covariances implied by the GMAR(2,2) model (right panel).

	Estimated Models			Means and Covariances	
	AR(4)	GMAR(2,2)	LMAR	Implied by GMAR(2,2)	
$\varphi_{1,0}$	0.010 (0.014)	0.043 (0.024)	0.010 (0.034)	μ_1	1.288
$\varphi_{2,0}$		−0.012 (0.006)	0.006 (0.020)	μ_2	−0.348
φ_1	1.278 (0.062)	1.266 (0.064)	1.257 (0.063)	$\gamma_{1,0}$	1.260
φ_2	−0.419 (0.101)	−0.299 (0.065)	−0.272 (0.065)	$\gamma_{1,1}$	1.228
φ_3	0.309 (0.101)			$\gamma_{1,1}/\gamma_{1,0}$	0.974
φ_4	−0.187 (0.062)			$\gamma_{2,0}$	0.225
σ_1^2	0.037 (0.003)	0.058 (0.008)	0.056 (0.008)	$\gamma_{2,1}$	0.220
σ_2^2		0.010 (0.002)	0.009 (0.003)	$\gamma_{2,1}/\gamma_{2,0}$	0.974
α_1		0.627 (0.197)			
β_0			0.033 (0.602)		
β_2			2.402 (0.674)		
$\max l_T(\theta)$	58.3	78.8	75.5		
AIC	−107	−146	−137		
BIC	−89	−124	−112		
N	0	0.77	0.39		
A_1	0.36	0.85	0.60		
A_4	0.27	0.08	0.07		
H_1	0.003	0.96	0.28		
H_4	0	0.69	0.23		

Notes: Left panel: Parameter estimates (with standard errors calculated using the Hessian in parentheses) of the estimated AR, GMAR, and LMAR models. GMAR(2,2) refers to the restricted model ($\varphi_{m,1} = \varphi_1$, $\varphi_{m,2} = \varphi_2$, $m = 1, 2$), with estimation based on the conditional likelihood. In LMAR model, the same restriction is imposed, and the β's define the mixing weights via $\log(\alpha_{1,t}/\alpha_{2,t}) = \beta_0 + \beta_2 y_{t-2}$. Rows labelled N, ..., H_4 present p–values of diagnostic tests based on quantile residuals. The test statistic for normality, N, is based on moments of quantile residuals and the test statistics for autocorrelation, A_k, and conditional heteroskedasticity, H_k, are based on the first k autocovariances and squared autocovariances of quantile residuals, respectively. Under correct specification, test statistic N is approximately distributed as χ^2_2 (AR(4)) or χ^2_3 (GMAR(2,2) and LMAR) and test statistics A_k and H_k are approximately distributed as χ^2_k. A p–value < 0.001 is denoted by 0. **Right panel:** Estimates derived for the expectations μ_m and elements of the covariance matrix $\boldsymbol{\Gamma}_{m,2}$; see Section 2.2.

so we only present those based on the conditional likelihood).

According to the diagnostic tests based on quantile residuals (see Table 1), the restricted GMAR$(2, 2)$ model provides a good fit to the data. To further investigate the properties of the quantile residuals, Figure 2 depicts time series and QQ–plots of the quantile residuals as well as the first ten standardized sample autocovariances of quantile residuals and their squares (the employed standardization is such that, under correct specification, the distribution of the sample autocovariances is approximately standard normal). The time series of quantile residuals computed from a correctly specified model should resemble a realization from an independent standard normal sequence. The graph of quantile residuals and the related QQ–plot give no obvious reason to suspect this, although some large positive quantile residuals occur. According to the approximate 99% critical bounds only two somewhat larger autocovariances are seen but even they are found at larger lags (we use 99% critical bounds because, from the viewpoint of statistical testing, several tests are performed). It is particularly encouraging that the GMAR model has been able to accommodate for the conditional heteroskedasticity in the data (see the bottom right panel of Figure 2), unlike the considered linear AR models (see the diagnostic tests for AR(4) model in Table 1). Thus, unlike the linear AR models, the GMAR$(2, 2)$ model seems to provide an adequate description for the interest rate differential series. Moreover, also according to the AIC and BIC, it outperforms the chosen linear AR(4) model by a wide margin (this also holds for the more parsimonious linear AR(2) model suggested by BIC).

Parameter estimates of the restricted GMAR$(2, 2)$ model are presented in Table 1, along with estimates derived for the expectations μ_m and elements of the covariance matrix $\mathbf{\Gamma}_{m,2}$ (see Section 2.2). The estimated sum of the AR coefficients is 0.967 which is slightly less than the corresponding sum 0.982 obtained in the linear AR(4) model. The reduction is presumably related to the differences in the intercept terms of the two AR

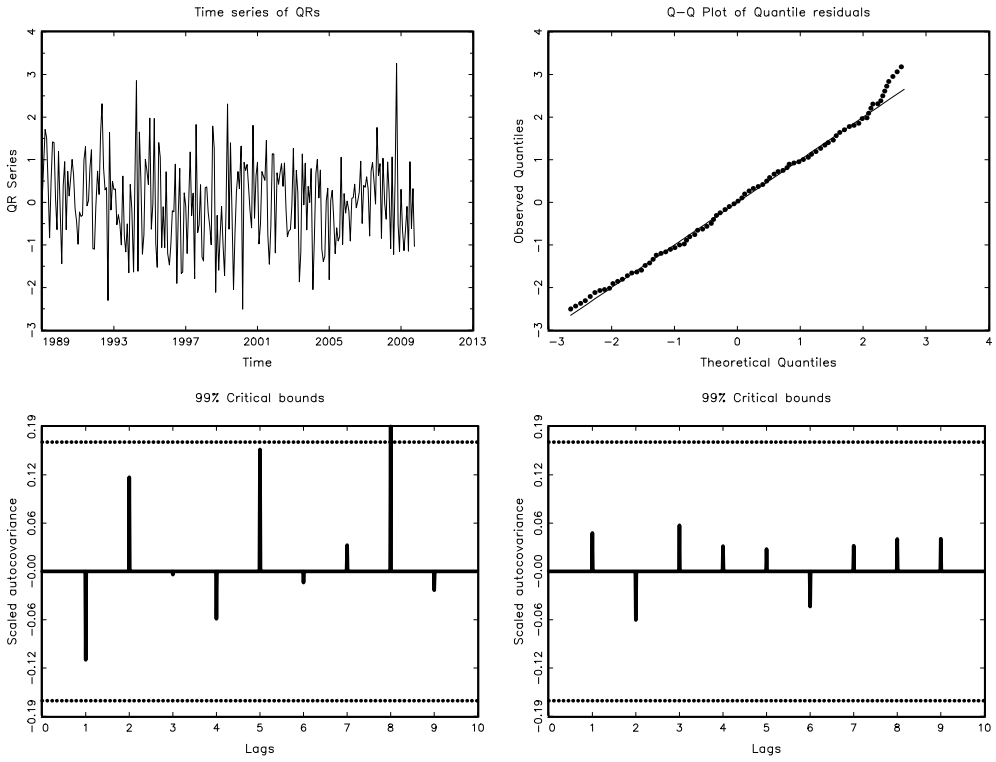

Figure 2: Diagnostics of the restricted GMAR(2,2) model described in Table 1: Time series of quantile residuals (top left panel), QQ-plot of quantile residuals (top right panel), and ten first scaled autocovariances of quantile residuals and squared quantile residuals (bottom left and right panels, respectively). The lines in the bottom panels show approximate 99% critical bounds.

components which is directly reflected as different means in the two regimes, with point estimates 1.288 and -0.348. The estimated error variances of the AR components are also very different and, consequently, the same is true for the variances of the two regimes, with point estimates 1.260 and 0.225. This feature is of course related to the above-mentioned fact that the model has been able to remove the conditional heteroskedasticity observed in linear modeling. According to the approximate standard errors in Table 1, the estimation accuracy appears quite reasonable except for the parameter α_1, the weight of the first component in the stationary distribution of the GMAR$(2, 2)$ process. The point estimate of this parameter is 0.627 with approximate standard error 0.197. A possible

explanation for this rather imprecise estimate is that the series is not sufficiently long to reveal the nature of the stationary distribution to which the parameter α_1 is directly related. (The parameter α_1 is also the one for which estimates based on the conditional and exact likelihoods differ the most, with the estimate based on the latter being 0.586.)

4.2 Mixture distribution and mixing weights

To further illustrate how the GMAR model can describe regime-switching behavior, we next discuss how the estimated mixture distribution and mixing weights may be interpreted. Based on the estimates of Table 1, Figure 3 shows the estimate of the two dimensional stationary mixture density $\sum_{m=1}^{2} \alpha_m \mathsf{n}_2\left(\boldsymbol{y}; \boldsymbol{\vartheta}_m\right)$ along with a related contour plot. A figure of the one dimensional mixture density $\sum_{m=1}^{2} \alpha_m \mathsf{n}_1\left(y; \boldsymbol{\vartheta}_m\right)$ and its two components is also included. These figures clearly illustrate the large differences between the shapes of the two component densities already apparent in the estimates of Table 1. The one dimensional mixture density is also drawn in Figure 1 (right panel, dashed line) and, as can be seen, there are rather large departures between the density implied by the model and the nonparametric kernel density estimate. The density implied by the model is more peaked and more concentrated than the kernel density estimate. The kernel density estimate may not be too reliable, however, because in some parts of the empirical distribution the number of observations seems to be rather small and the choice of the bandwidth parameter has a noticeable effect on the shape of the kernel density (the estimate in Figure 1 is based on the bandwidth suggested by Silverman (1984)).

Figure 1 (left panel, dashed line) depicts the time series of the estimated mixing weight $\hat{\alpha}_{1,t}$ scaled so that $\hat{\alpha}_{1,t} = \max y_t$ when $\hat{\alpha}_{1,t} = 1$, and $\hat{\alpha}_{1,t} = \min y_t$ when $\hat{\alpha}_{1,t} = 0$. During the period before 1996 or 1997 the first regime (with higher mean, $\hat{\mu}_1 = 1.288$) is clearly dominating. Except for only a few exceptional months the mixing weights $\hat{\alpha}_{1,t}$ are practically unity. This period corresponds to a high level regime or regime where

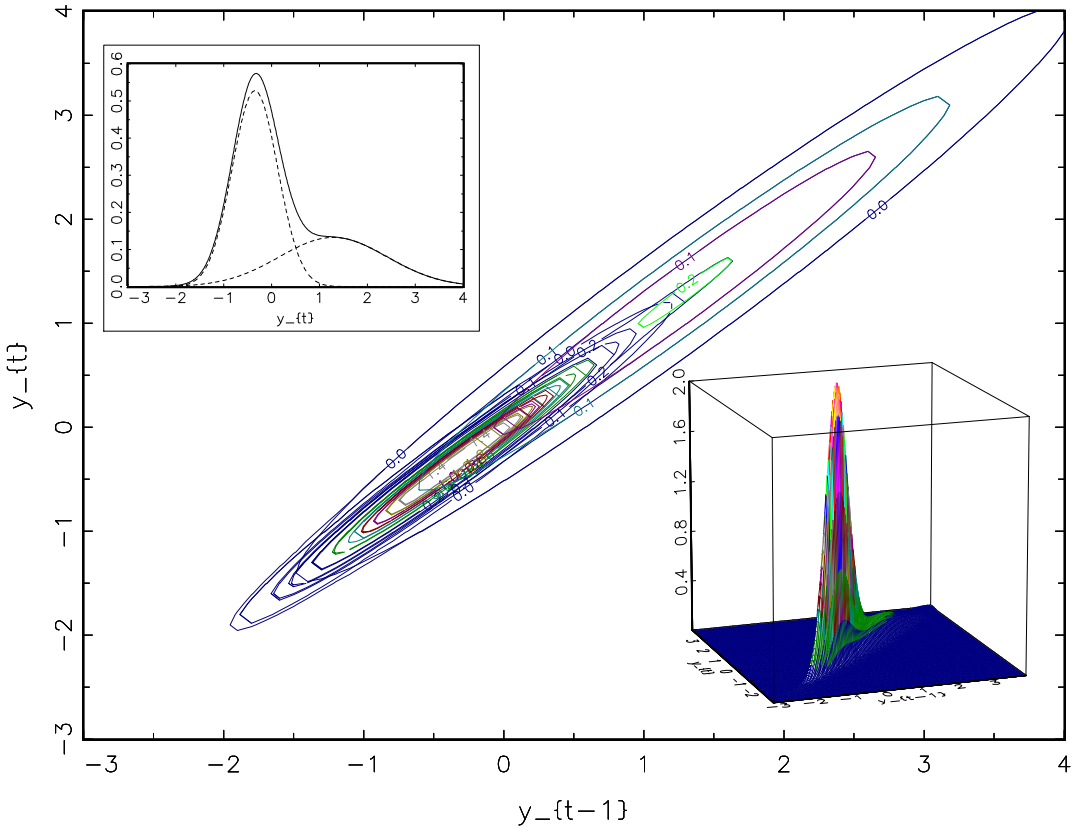

Figure 3: Estimate of the two dimensional stationary mixture density implied by the GMAR(2,2) model described in Table 1 (bottom-right picture), its contour plots (middle), and the corresponding one dimensional marginal density and its two components (top-left).

U.S. bond yields are smaller than Euro area bond yields. After this period a low level regime, where U.S. bond yields are larger than Euro Area bond yields, prevails until 2008 or the early stages of the most recent financial crisis. Interestingly, the period between (roughly) 1997 and 2004 is 'restless' in that several narrow spikes in the estimated mixing weights occur. Because no marked increases appear in the level of the series it seems reasonable to relate these spikes to the rather large differences between the variances in the two regimes. Although the second (low-level) regime is here dominating, observations are occasionally generated by the first AR component whose estimated error variance is over five times the estimated error variance of the second AR component (see Table 1).

However, despite these large shocks from the first AR component, the level of the series has remained rather low.

To discuss this point in more detail, recall that the mixing weights $\alpha_{1,t}$ and $\alpha_{2,t}$ depend on the density functions $\mathsf{n}_2(\boldsymbol{y}_{t-1};\boldsymbol{\vartheta}_1)$ and $\mathsf{n}_2(\boldsymbol{y}_{t-1};\boldsymbol{\vartheta}_2)$ where $\boldsymbol{y}_{t-1} = (y_{t-1}, y_{t-2})$. As Figure 3 indicates, the density function $\mathsf{n}_2(\boldsymbol{y}_{t-1};\boldsymbol{\vartheta}_2)$ ('low-level regime') is concentrated on the lower tail of $\mathsf{n}_2(\boldsymbol{y}_{t-1};\boldsymbol{\vartheta}_1)$ ('high-level regime'; see also the estimates in Table 1). Consequently, it is possible for the process to be in either of these two regimes and at the same time not far from the mean of $\mathsf{n}_2(\boldsymbol{y}_{t-1};\boldsymbol{\vartheta}_2)$. Switching from the second ('low-level') regime to the first ('high-level') one can then happen without much increase in the level of the series. This seems to have happened between 1997 and 2004 when (based on the time series of estimated mixing weights $\hat{\alpha}_{1,t}$) the series appears to have mostly evolved in the second regime and the process has only occasionally paid short visits to the (lower tail of the) first regime. The domination of the second regime has been clearer from 2005 until the early stages of the 2008 financial crisis, after which the first regime becomes dominating. During the last couple of years the estimated mixing weights $\hat{\alpha}_{1,t}$ have part of the time been very high but the level of the series has still remained rather moderate. Again, it seems reasonable to think that the dominance of the first regime is mainly caused by its large variance. This interpretation, as well as the one related to the narrow spikes between 1997 and 2004, is supported by the time series graph of the conditional variance implied by the estimated model. Without showing this graph we just note that its shape is more or less indistinguishable from the time series graph of the estimated mixing weight $\hat{\alpha}_{1,t}$ in the left panel of Figure 1.

To gain further insight into the preceding discussion Figure 4 depicts the estimated mixing weight $\hat{\alpha}_{1,t}$ as a function of y_{t-1} and y_{t-2}. The functional form is similar to an overturned version of the estimated density function $\mathsf{n}_2(\boldsymbol{y}_{t-1};\hat{\boldsymbol{\vartheta}}_2)$. Outside an ellipse, roughly corresponding to an ellipse where the estimated density $\mathsf{n}_2(\boldsymbol{y}_{t-1};\hat{\boldsymbol{\vartheta}}_2)$ has nonnegligible

mass, the estimated mixing weight $\hat{\alpha}_{1,t}$ is nearly unity. On the other hand, in the center of this ellipse, or close to the point where $y_{t-1} = y_{t-2} \approx -0.5$, the estimated mixing weight $\hat{\alpha}_{1,t}$ attains its minimum value. The closer the series is to this minimum point, the clearer it evolves in the lower regime; when it approaches the border of the aforementioned ellipse, the probability of switching to the upper regime increases. The spikes in the time series graph of $\hat{\alpha}_{1,t}$ in Figure 1 (left panel) between 1997 and 2004 have apparently occurred when the series has been close to the border of this ellipse. It is interesting to note that the spikes before 2001 have occurred when the level of the series is quite low so that the series has evolved in a way which has increased the (conditional) probability of obtaining an observation from the upper regime but without much increase in the level of the series. As Figure 4 illustrates, this is possible. A feature like this may be difficult to capture by previous mixture AR models as well as by TAR and STAR models in which regime switches are typically determined by the level of the series. For instance, in the model of Lanne and Saikkonen (2003) the probability of a regime switch is determined by the level of a single lagged value of the series and similarly for (the most typically used) TAR and STAR models (see Tong (1990), Teräsvirta (1994), and Teräsvirta, Tjøstheim, and Granger (2010)). The models of Wong and Li (2001b) and Dueker, Sola, and Spagnolo (2007) are more general in that regime switches are determined by the level of a linear combination of a few past values of the series and, for comparison, we next discuss estimation results based on the model of Wong and Li (2001b).

4.3 Comparison to the LMAR model of Wong and Li (2001b)

For comparison, we also fitted the LMAR model of Wong and Li (2001b) to the interest-rate differential series. Similarly to the GMAR model, our starting point was a LMAR model with two lags in the autoregressive polynomial and in the mixing weights (that is, $\log(\alpha_{1,t}/\alpha_{2,t}) = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2}$). The best fit was obtained with a specification where

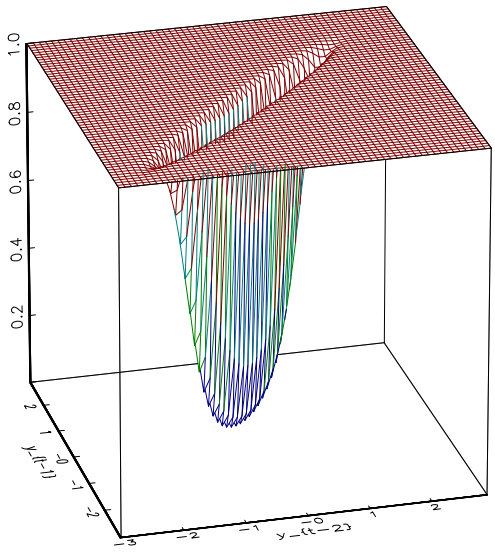

Figure 4: Estimated mixing weight $\hat{\alpha}_{1,t}$ of the restricted GMAR(2,2) model described in Table 1.

the autoregressive coeffients were restricted equal in the two regimes (like in our GMAR model) and the mixing weights were specified as $\log(\alpha_{1,t}/\alpha_{2,t}) = \beta_0 + \beta_2 y_{t-2}$. Table 1 presents the estimated parameters of this model, along with diagnostic tests based on quantile residuals. In terms of parameter estimates of the autoregressive components, the results for the LMAR model are very similar to those of the GMAR model, although the sum of the autoregressive coefficients, 0.986, is slightly larger being close to that obtained with the AR(4) model. The mixing weights implied by the estimated LMAR model (not shown) are also comparable to the ones obtained from the GMAR model (see Figure 1, left panel, dashed line). The most noticeable difference occurs during the 'restless' period between (roughly) 1997 and 2004 where the mixing weights of the LMAR model evolve rather smoothly without large spikes similar to those obtained from the GMAR model. A similar difference occurs in the time series graphs of the conditional variances of the two models (not shown). The LMAR model also passes all the diagnostic tests performed (see Table 1), and a graphical analysis of the quantile residuals (not shown, but comparable to that in Figure 2) indicates no obvious deviations from them forming an (approximately)

independent standard normal sequence. Therefore, the LMAR model appears a viable alternative to the GMAR model although, according to information criteria, the GMAR model provides a better fit.

4.4 Forecasting performance

According to the estimation results presented in Table 1, both the GMAR model and the LMAR model provide a significant improvement over the linear AR model in terms of in-sample statistical fit. We next evaluate their performance in a small out-of-sample forecasting exercise. We consider four forecasting models, namely the GMAR model with estimation based on the conditional likelihood ('GMAR conditional' for brevity), the GMAR model based on the exact likelihood ('GMAR exact'; for this model the estimation results are not shown), the LMAR model, and the linear AR(4) model. Assuming correct specification, optimal one-step-ahead forecasts (in mean squared sense and ignoring estimation errors) are straightforward to compute with each model because explicit formulas are available for the conditional expectation (see (4)). As is well known, computing multi-step-ahead forecasts is simple for linear AR models as well but for mixture models the situation is complicated in that explicit formulas are very difficult to obtain (cf. Dueker, Sola, and Spagnolo (2007, Sec. 4.2)). For mixture models, as well as for TAR and STAR models, a simple and widely used approach to obtain multi-step-ahead forecasts is based on simulation (see, e.g., Dueker, Sola, and Spagnolo (2007, Sec. 4.2), Teräsvirta, Tjøstheim, and Granger (2010, Ch. 14), and the references therein).

The simulation scheme we use is as follows, for each of the considered mixture models. The date of forecasting (up until which observations are used) ranges from December 2009 till October 2011, and for each date of forecasting, forecasts are computed for all the subsequent periods up until September 2011. As estimation in mixture models requires numerical optimization, we do not update estimates when the date of forecasting changes

so that all forecasts are based on a model whose parameters are estimated by using data from January 1989 to December 2009. Using initial values known at the date of forecasting, we simulate 500,000 realizations and treat the mean of these realizations as a point forecast, and repeat this for all forecast horizons up until September 2011. This results in a total of 21 one-step forecasts, 20 two-step forecasts, ..., nine 12-step forecasts (as well as a few forecasts for longer horizons which we discard). For each of the forecast horizons 1, ..., 12, we measure forecast accuracy by the mean squared prediction error (MSPE) and mean absolute prediction error (MAPE), with the mean computed across the 21, ..., nine forecasts available. Due to the small number of prediction errors used to compute these measures the results to be discussed below should be treated as only indicative.

As expected, using both the MSPE measure and the MAPE measure, forecast accuracy was best in one-step-ahead prediction and steadily detoriated with the forecast horizon. A perhaps less expected fact was that the relative ranking of the four forecasting models remained more or less the same regardless of the forecast horizon or the accuracy measure used: The most precise forecasts were always delivered by the GMAR conditional, the next best by the GMAR exact, and followed by the LMAR and AR(4) models whose ranking changed depending on the forecast horizon and accuracy measure used. The results are presented in Figure 5, the two subfigures corresponding to a differect forecast accuracy measure (MSPE, MAPE), and with the forecast horizon (1, ..., 12) on the horizontal axis. For clarity of exposition, these figures present the forecast accuracy of the models relative to the most precise forecasting model, the GMAR conditional. Therefore, in each figure, the straight line (at 100) represents the GMAR conditional, whereas the other three lines represent the size of the forecast error made relative to the GMAR conditional; for instance, a value of 110 in the rightmost figure is to be interpreted as a MAPE 10% larger than for the GMAR conditional.

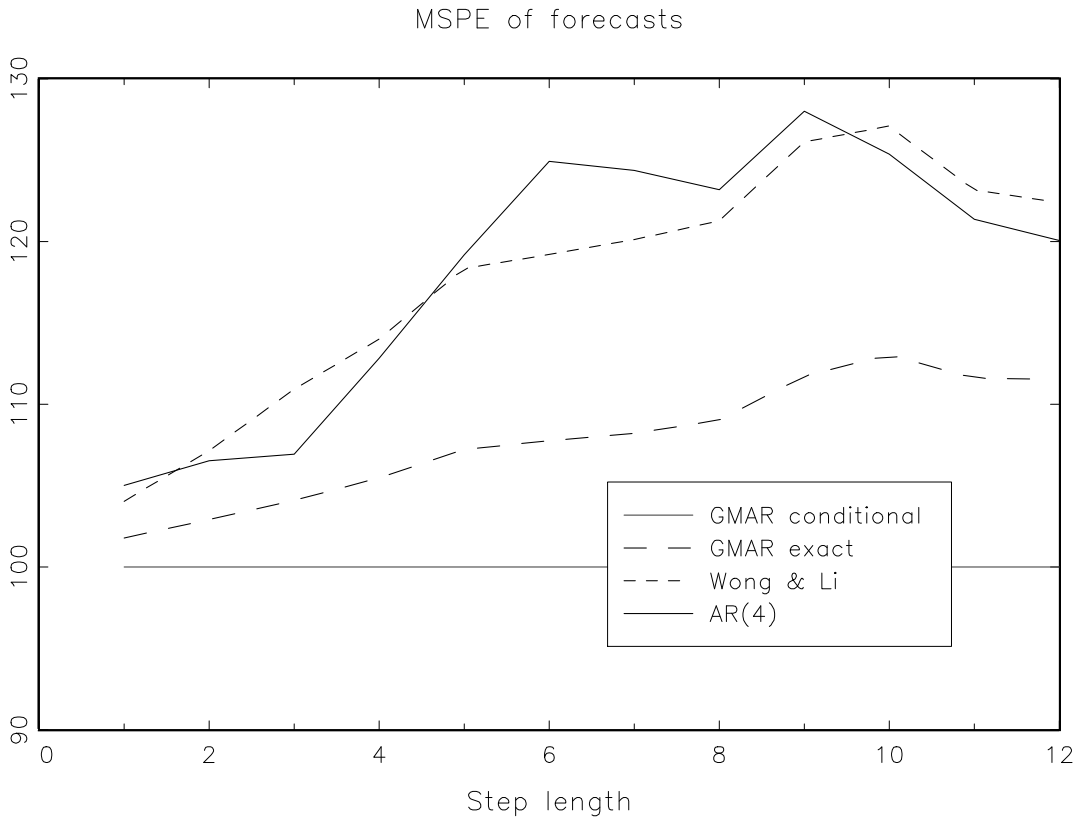

Figure 5: Relative forecast accuracies measured using mean squared prediction error (MSPE). The four lines in each figure represent different forecasting models.

The above-mentioned dominance of the GMAR conditional and GMAR exact over the other two forecasting models is immediate from Figure 5. Although the amount by which their forecasts are more accurate varies depending on the forecast horizon and accuracy measure employed, it is noteworthy that the GMAR model consistently provides the best forecasts. A possible explanation for this outcome lies in the way the mixing weights are defined in the GMAR model. (Note that the two autoregressive components in the GMAR and LMAR models are very similar, suggesting that the difference in forecasting performance is related to the definition of the mixing mechanism.) As the discussion in Section 4.2 already pointed out, the regime from which the next observation is generated is (essentially) determined according to the entire (stationary) distribution of the (in this case) two most recent observations of the series, and not merely their levels or linear combinations. In addition to being intuitively reasonable, this may be advantageous also in forecasting. Moreover, the systematically better forecast accuracy of GMAR conditional over GMAR exact may also be due to a more successful estimation of the mixing weights: although the parameter estimates based on conditional and exact ML are very similar, the greatest difference occurs in the estimate of the parameter α_1 which directly affects the mixing weight $\alpha_{1,t}$ (the exact and conditional ML estimates of α_1 are 0.586 and 0.627, respectively).

Another possible explanation for the dominance of the GMAR conditional and GMAR exact over the other two forecasting models lies in the estimated autoregressive polynomials of the models. Although the estimated autoregressive polynomials of the LMAR and GMAR (both conditional and exact) models are very similar, the sum of the estimated autoregressive coefficients in the LMAR and AR(4) models (0.986 and 0.982, respectively) is larger and closer to unity than the corresponding sum in the GMAR conditional and GMAR exact (0.967 and 0.966, respectively). Even though the difference looks small, it does indicate stronger persistence in the LMAR and AR(4) models, and its impact may

not be so small because it occurs in the vicinity of the boundary value unity where a (linear) autoregressive process becomes a nonstationary unit root process.

5 Conclusion

This paper provides a more detailed discussion of the mixture AR model considered by Glasbey (2001) in the first order case. This model, referred to as the GMAR model, has several attractive properties. Unlike most other nonlinear AR models, the GMAR model has a clear probability structure which translates into simple conditions for stationarity and ergodicity. These theoretical features are due to the definition of the mixing weights which have a natural interpretation. In our empirical example the GMAR model appeared flexible, being able to describe features in the data that may be difficult to capture by alternative (nonlinear) AR models, and it also showed promising forecasting performance.

In this paper we have only considered a univariate version of the GMAR model. In the future we plan to explore a multivariate extension. Providing a detailed analysis of the asymptotic theory of estimation and statistical inference is another topic left for future work. In this context, the problem of developing statistical tests that can be used to test for the number of AR components is of special interest. Due to its nonstandard nature this testing problem may be quite challenging, however. Applications of the GMAR model to different data sets will also be presented. Finally, it would be of interest to examine the forecasting performance of the GMAR model in greater detail than was done here.

Appendix A: Technical details

Proof of Theorem 1. We first note some properties of the stationary auxiliary autoregressions $\nu_{m,t}$. Denoting $\boldsymbol{\nu}_{m,t}^{+} = (\nu_{m,t}, \boldsymbol{\nu}_{m,t-1})$ $((q+1)\times 1)$, it is seen that $\boldsymbol{\nu}_{m,t}^{+}$ follows the

$(q+1)$–dimensional multivariate normal distribution with density

$$\begin{aligned}\mathsf{n}_{q+1}\left(\boldsymbol{\nu}_{m,t}^{+};\boldsymbol{\vartheta}_m\right) &= (2\pi)^{-(q+1)/2}\det(\boldsymbol{\Gamma}_{m,q+1})^{-1/2} \\ &\quad\times\exp\left\{-\frac{1}{2}\left(\boldsymbol{\nu}_{m,t}^{+}-\mu_m\mathbf{1}_{q+1}\right)'\boldsymbol{\Gamma}_{m,q+1}^{-1}\left(\boldsymbol{\nu}_{m,t}^{+}-\mu_m\mathbf{1}_{q+1}\right)\right\},\end{aligned}$$

where $\mathbf{1}_{q+1}=(1,\ldots,1)$ $((q+1)\times 1)$ and the matrices $\boldsymbol{\Gamma}_{m,q+1}$, $m=1,\ldots,M$, have the usual symmetric Toeplitz form similar to their counterparts in (7) with each $\boldsymbol{\Gamma}_{m,q+1}$ depending on the parameters $\boldsymbol{\varphi}_m$ and σ_m (see, e.g., Reinsel (1997, Sec. 2.2.3)). This joint density can be decomposed as

$$\mathsf{n}_{q+1}\left(\boldsymbol{\nu}_{m,t}^{+};\boldsymbol{\vartheta}_m\right)=\mathsf{n}_1\left(\nu_{m,t}\mid\boldsymbol{\nu}_{m,t-1};\boldsymbol{\vartheta}_m\right)\mathsf{n}_q\left(\boldsymbol{\nu}_{m,t-1};\boldsymbol{\vartheta}_m\right), \tag{13}$$

where the normality of the two densities on the right-hand side follows from properties of the multivariate normal distribution (see, e.g., Anderson (2003, Theorems 2.4.3 and 2.5.1)). Moreover, $\mathsf{n}_q(\cdot;\boldsymbol{\vartheta}_m)$ clearly has the form given in (7), and making use of the Yule-Walker equations (see, e.g., Box, Jenkins, and Reinsel (2008, p. 59)), it can be seen that (here $\mathbf{0}_{q-p}$ denotes a vector of zeros with dimension $q-p$)

$$\begin{aligned}&\mathsf{n}_1\left(\nu_{m,t}\mid\boldsymbol{\nu}_{m,t-1};\boldsymbol{\vartheta}_m\right)\\ &= \left(2\pi\sigma_m^2\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma_m^2}(\nu_{m,t}-\mu_m-(\boldsymbol{\varphi}_m,\mathbf{0}_{q-p})'(\boldsymbol{\nu}_{m,t-1}-\mu_m\mathbf{1}_q))^2\right\}\\ &= \left(2\pi\sigma_m^2\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma_m^2}(\nu_{m,t}-\varphi_{m,0}-\boldsymbol{\varphi}_m'\left(\nu_{m,t-1},\ldots,\nu_{m,t-p}\right))^2\right\}.\end{aligned} \tag{14}$$

The rest of the proof makes use of the theory of Markov chains (for the employed concepts, see Meyn and Tweedie (2009)). To make the Markov chain representation of $\boldsymbol{y}_t$ explicit we denote $\imath_q=(1,0,\ldots,0)$ $(q\times 1)$, and for $m=1,\ldots,M$,

$$\Phi_m=\begin{bmatrix}\varphi_{m,1} & \cdots & \cdots & \varphi_{m,p} & \\ 1 & \cdots & 0 & 0 & \\ \vdots & \ddots & \vdots & \vdots & \mathbf{0}_{p,q-p}\\ 0 & \cdots & 1 & 0 & \\ & \mathbf{0}_{q-p,p} & & & I_{q-p}\end{bmatrix}\qquad(q\times q),$$

where $\mathbf{0}_{p,q-p}$ and $\mathbf{0}_{q-p,p}$ denote zero matrices with the indicated dimensions. Then (3) can be written as

$$\boldsymbol{y}_t = \sum_{m=1}^{M} \boldsymbol{s}_{t,m} \left(\varphi_{m,0} \imath_q + \Phi_m \boldsymbol{y}_{t-1} + \sigma_m \varepsilon_t \imath_q \right),$$

making clear that $\boldsymbol{y}_t$ is a Markov chain on $\mathbb{R}^q$.

Now, let $\boldsymbol{y}_0 = (y_0, \ldots, y_{-q+1})$ be a random vector whose distribution has the density

$$f(\boldsymbol{y}_0; \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m \mathsf{n}_q \left(\boldsymbol{y}_0; \boldsymbol{\vartheta}_m \right).$$

According to (1) and (14), the conditional density of y_1 given $\boldsymbol{y}_0$ is

$$\begin{aligned} f(y_1 \mid \boldsymbol{y}_0; \boldsymbol{\theta}) &= \sum_{m=1}^{M} \alpha_{m,1} \mathsf{n}_1 \left(y_1 \mid \boldsymbol{y}_0; \boldsymbol{\vartheta}_m \right) \\ &= \sum_{m=1}^{M} \frac{\alpha_m}{\sum_{n=1}^{M} \alpha_n \mathsf{n}_q \left(\boldsymbol{y}_0; \boldsymbol{\vartheta}_n \right)} \mathsf{n}_q \left(\boldsymbol{y}_0; \boldsymbol{\vartheta}_m \right) \mathsf{n}_1 \left(y_1 \mid \boldsymbol{y}_0; \boldsymbol{\vartheta}_m \right) \\ &= \sum_{m=1}^{M} \frac{\alpha_m}{\sum_{n=1}^{M} \alpha_n \mathsf{n}_q \left(\boldsymbol{y}_0; \boldsymbol{\vartheta}_n \right)} \mathsf{n}_{q+1} \left((y_1, \boldsymbol{y}_0); \boldsymbol{\vartheta}_m \right), \end{aligned}$$

where the second and third equalities are due to (8) and (13). It thus follows that the density of $(y_1, \boldsymbol{y}_0) = (y_1, y_0, \ldots, y_{-q+1})$ is

$$f((y_1, \boldsymbol{y}_0); \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m \mathsf{n}_{q+1} \left((y_1, \boldsymbol{y}_0); \boldsymbol{\vartheta}_m \right).$$

Integrating y_{-q+1} out it follows that the density of $\boldsymbol{y}_1$ is

$$f(\boldsymbol{y}_1; \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m \mathsf{n}_q \left(\boldsymbol{y}_1; \boldsymbol{\vartheta}_m \right)$$

Therefore, $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are identically distributed. As $\{\boldsymbol{y}_t\}_{t=1}^{\infty}$ is a (time homogeneous) Markov chain, we can thus conclude that $\{\boldsymbol{y}_t\}_{t=1}^{\infty}$ has a stationary distribution $\pi_{\boldsymbol{y}}(\cdot)$, say, characterized by the density

$$f(\cdot; \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m \mathsf{n}_q \left(\cdot; \boldsymbol{\vartheta}_m \right)$$

(cf. Meyn and Tweedie (2009, pp. 230–231)). Being a mixture of multivariate normal distributions, all moments of the stationary distribution are finite.

It remains to establish ergodicity. To this end, let $P^q_{\boldsymbol{y}}(\boldsymbol{y}, \cdot) = \Pr(\boldsymbol{y}_q \mid \boldsymbol{y}_0 = \boldsymbol{y})$ signify the q-step transition probability measure of $\boldsymbol{y}_t$. It is straightforward to check that $P^q_{\boldsymbol{y}}(\boldsymbol{y}, \cdot)$ has a density given by

$$f(\boldsymbol{y}_q \mid \boldsymbol{y}_0; \boldsymbol{\theta}) = \prod_{t=1}^{q} f(y_t \mid \boldsymbol{y}_{t-1}; \boldsymbol{\theta}) = \prod_{t=1}^{q} \sum_{m=1}^{M} \alpha_{m,t} \mathsf{n}_1 \left(y_t \mid \boldsymbol{y}_{t-1}; \boldsymbol{\vartheta}_m\right).$$

The last expression makes clear that $f(\boldsymbol{y}_q \mid \boldsymbol{y}_0; \boldsymbol{\theta}) > 0$ for all $\boldsymbol{y}_q \in \mathbb{R}^q$ and all $\boldsymbol{y}_0 \in \mathbb{R}^q$ so that, from every initial state $\boldsymbol{y}_0 = \boldsymbol{y}$ $(\in \mathbb{R}^q)$, the chain $\boldsymbol{y}_t$ can in q steps reach any set of the state space $\mathbb{R}^q$ with positive Lebesgue measure. Using the definitions of irreducibility and aperiodicity we can therefore conclude that the chain $\boldsymbol{y}_t$ is irreducible and aperiodic (see Meyn and Tweedie (2009, Chapters 4.2 and 5.4)). Moreover, also the q-step transition probability measure $P^q_{\boldsymbol{y}}(\boldsymbol{y}, \cdot)$ is irreducible, aperiodic, and has $\pi_{\boldsymbol{y}}$ as its invariant distribution (Meyn and Tweedie, 2009, Theorem 10.4.5).

A further consequence of the preceding discussion is that the q-step transition probability measure $P^q_{\boldsymbol{y}}(\boldsymbol{y}, \cdot)$ is equivalent to the Lebesgue measure on $\mathbb{R}^q$ for all $\boldsymbol{y} \in \mathbb{R}^q$. As the stationary probability measure $\pi_{\boldsymbol{y}}(\cdot)$ also has a (Lebesgue) density positive everywhere in $\mathbb{R}^q$ it is likewise equivalent with the Lebesgue measure on $\mathbb{R}^q$. Consequently, the q-step transition probability measure $P^q_{\boldsymbol{y}}(\boldsymbol{y}, \cdot)$ is absolutely continuous with respect to the stationary probability measure $\pi_{\boldsymbol{y}}(\cdot)$ for all $\boldsymbol{y} \in \mathbb{R}^q$.

To complete the proof, we now use the preceding facts and conclude from Theorem 1 and Corollary 1 of Tierney (1994) that $\left\|P^{qn}_{\boldsymbol{y}}(\boldsymbol{y}, \cdot) - \pi_{\boldsymbol{y}}(\cdot)\right\| \to 0$ as $n \to \infty$ for all $\boldsymbol{y} \in \mathbb{R}^q$, where $\|\cdot\|$ signifies the total variation norm of probability measures. Now, by Proposition 13.3.2 of Meyn and Tweedie (2009), also $\left\|P^{n}_{\boldsymbol{y}}(\boldsymbol{y}, \cdot) - \pi_{\boldsymbol{y}}(\cdot)\right\| \to 0$ as $n \to \infty$ for all $\boldsymbol{y} \in \mathbb{R}^q$ (as the total variation norm is non-increasing in n). Hence, $\boldsymbol{y}_t$ is ergodic in the sense of Meyn and Tweedie (2009, Ch. 13). ■

Remark. In the discussion following Theorem 1 it was pointed out that non-ergodicity would be an undesirable implication for a process having all finite dimensional distribu-

tions being Gaussian mixtures. To see that this holds in a particular special case, suppose all finite dimensional distributions of a process x_t, say, are Gaussian mixtures of the form (9) so that, for any $T \geq 1$, the distribution of a realization $(x_1, \ldots, x_T)$ is

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m \mathsf{n}_T(\boldsymbol{x}; \boldsymbol{\vartheta}_m),$$

where the density function $\mathsf{n}_T(\boldsymbol{x}; \boldsymbol{\vartheta}_m)$ is a T–dimensional analog of that in (7). The process x_t is clearly stationary. For simplicity, consider the special case where $M = 2$, $\varphi_{m,i} = 0$ $(i = 1, \ldots, p,\ m = 1, 2)$, $\sigma_1 = \sigma_2 = \sigma$, and $\varphi_{1,0} \neq \varphi_{2,0}$. Then $\mathsf{n}_T(\boldsymbol{x}; \boldsymbol{\vartheta}_i)$ is the joint density of T independent Gaussian random variables with mean $\varphi_{i,0}$ and variance σ^2 $(i = 1, 2)$. This means that, for every T,

$$(x_1, \ldots, x_T) \sim \begin{cases} \mathsf{n}_T(\boldsymbol{x}; \boldsymbol{\vartheta}_1), & \text{with probability } \alpha_1 \\ \mathsf{n}_T(\boldsymbol{x}; \boldsymbol{\vartheta}_2), & \text{with probability } 1 - \alpha_1. \end{cases}$$

This implies that, for every T, the sample mean $\bar{X}_T = T^{-1} \sum_{t=1}^{T} x_t$ is distributed as $N(\varphi_{1,0}, \sigma^2/T)$ with probability α_1 and as $N(\varphi_{2,0}, \sigma^2/T)$ with probability $1 - \alpha_1$. As $\varphi_{1,0} \neq \varphi_{2,0}$ and $0 < \alpha_1 < 1$ is assumed, it is therefore immediate that no law of large numbers holds, and consequently the process x_t cannot be ergodic. Indeed, it is not difficult to check that $\bar{X}_T$ converges in distribution to a random variable taking the values $\varphi_{1,0}$ and $\varphi_{2,0}$ with probability α_1 and $1 - \alpha_1$, respectively.

Appendix B: Comparison of different mixing weights

In this Appendix, we provide a graphical illustration comparing the different mixture AR models discussed in Section 2.3. In the top panels of Figure 6 below, we plot the mixing weight $\alpha_{1,t}$ of the GMAR model as a function of $y_{t-1} = y$ in the case $M = 2$, $p = q = 1$, with certain parameter combinations. The bottom left panel shows $\alpha_{1,t}$ in some cases for the LMAR model of Wong and Li (2001b); in the model of Lanne and Saikkonen (2003) $\alpha_{1,t}$ behaves in a comparable way (no picture presented). The two pictures on the

left illustrate that the three models can produce mixing weights of similar monotonically increasing patterns. The figure in the top left panel also illustrates that, other things being equal, a decrease in the value of α_m decreases the value of $\alpha_{m,t}$. In the conditional expectation of a basic logistic two-regime STAR model, referred to as the LSTAR1 model in Teräsvirta, Tjøstheim, and Granger (2010, Sec. 3.4.1), the transition function, which is the counterpart of the mixing weight $\alpha_{1,t}$, also behaves in a similar monotonically increasing way. Given these observations it is interesting that with suitable parameter values our GMAR model can produce nonmonotonic mixing weights even in the case $M = 2$. The top right panel illustrates this. The considered (first-order) model of Wong and Li (2001b) can produce mixing weights of this form only when $M > 2$ (the same holds for the model of Lanne and Saikkonen (2003)). Similarly, in LSTAR models a transition function of this form cannot be obtained with a LSTAR1 model. For that one needs an LSTAR2 model (see Teräsvirta, Tjøstheim, and Granger (2010, Sec. 3.4.1)) or some other similar model such as the exponential autoregressive model of Haggan and Ozaki (1981). Thus, once the number of component models is specified our GMAR model appears more flexible in terms of the form of mixing weights than the aforementioned previous mixture models and the same is true when the mixing weights of our GMAR model are compared to the transition functions of LSTAR models.

As far as the mixing weights of the model of Dueker, Sola, and Spagnolo (2007) are concerned they can be nonmonotonic, as illustrated in the bottom right panel of Figure 6. After trying a number of different parameter combinations it seems, however, that (at least in the case $p = 1$) nonmonotonic mixing weights are rather special for this model. The first four (monotonic) graphs in the bottom right panel correspond to parameter configurations in Table 2 of Dueker, Sola, and Spagnolo (2007). The fourth one is interesting in that it produces a nearly constant graph (the graph would be constant if the values of the standard deviations σ_1 and σ_2 were changed to be equal). Finally, note that a common

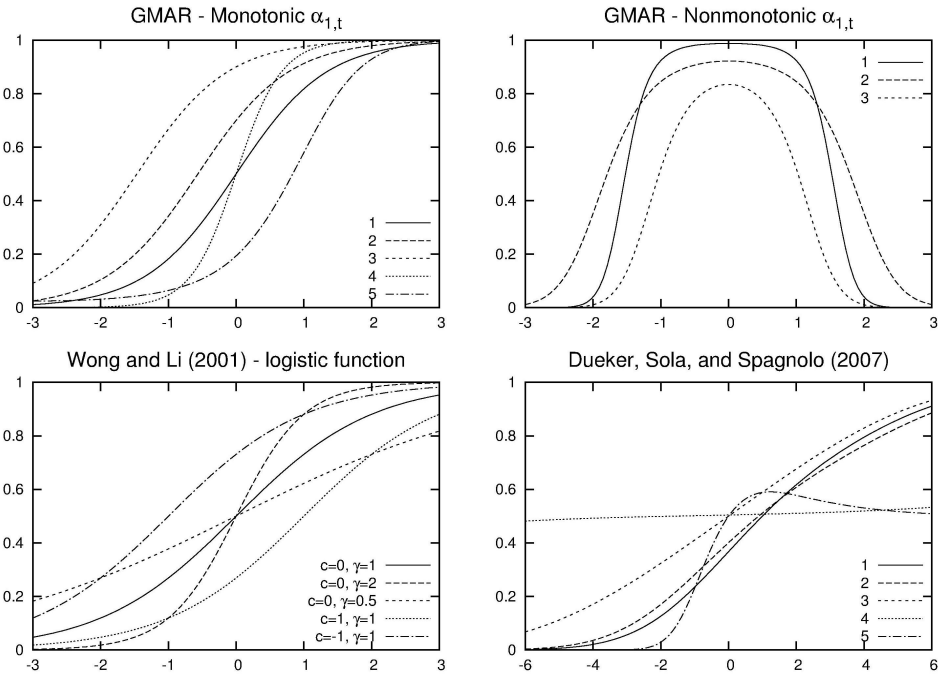

Figure 6: **Top left panel:** $\alpha_{1,t}$ in the GMAR model ($p = 1$) as a function of y_{t-1}. Parameter values used: model 1: $\varphi_{1,0} = 0.5$, $\varphi_{2,0} = -0.5$, $\varphi_{1,1} = \varphi_{2,1} = 0.5$, $\sigma_1^2 = \sigma_2^2 = 1$, $\alpha_1 = 0.5$; model 2: same as model 1 but $\alpha_1 = 0.7$; model 3: same as model 1 but $\alpha_1 = 0.9$; model 4: $\varphi_{1,0} = 1$, $\varphi_{2,0} = -1$, $\varphi_{1,1} = \varphi_{2,1} = 0.5$, $\sigma_1^2 = \sigma_2^2 = 1$, $\alpha_1 = 0.5$; model 5: $\varphi_{1,0} = \varphi_{2,0} = 0.5$, $\varphi_{1,1}0.75$, $\varphi_{2,1} = 0.25$, $\sigma_1^2 = \sigma_2^2 = 1$, $\alpha_1 = 0.5$. **Top rigth panel:** $\alpha_{1,t}$ in the GMAR model ($p = 1$) as a function of y_{t-1}. Parameter values used: model 1: $\varphi_{1,0} = \varphi_{2,0} = 0$, $\varphi_{1,1} = 0.2$, $\varphi_{2,1} = 0.9$, $\sigma_1^2 = 0.25$, $\sigma_2^2 = 4$, $\alpha_1 = 0.5$; model 2: same as model 1 but $\sigma_1^2 = \sigma_2^2 = 0.5$, $\alpha_1 = 0.7$; model 3: same as model 1 but $\sigma_2^2 = 0.25$. **Bottom left panel:** $\alpha_{1,t}$ in the model of Wong and Li (2001b) as a function of y_{t-1}. Logistic equation assumed to be of the form $\log(\alpha_{1,t}/\alpha_{2,t}) = \gamma(y_{t-1} - c)$, or equivalently, $\alpha_{1,t} = \frac{1}{1+e^{-\gamma(y_{t-1}-c)}}$. Note that this is exactly the standard form of the logistic function. Curves correspond to different values of c and γ. **Bottom right panel:** $\alpha_{2,t} = 1-\alpha_{1,t}$ in the model of Dueker, Sola, and Spagnolo (2007) as a function of y_{t-1}. Parameter values used: model 1: $c_1 = 1$, $\varphi_{1,0} = -0.5$, $\varphi_{2,0} = 0.5$, $\varphi_{1,1} = \varphi_{2,1} = 0.9$, $\sigma_1 = 3$, $\sigma_2 = 2$; model 2: $c_1 = 1$, $\varphi_{1,0} = -1$, $\varphi_{2,0} = 1$, $\varphi_{1,1} = \varphi_{2,1} = 0.9$, $\sigma_1 = 3$, $\sigma_2 = 2$; model 3: $c_1 = 0$, $\varphi_{1,0} = -1$, $\varphi_{2,0} = 1$, $\varphi_{1,1} = \varphi_{2,1} = 0.9$, $\sigma_1 = \sigma_2 = 3$; model 4: $c_1 = 0$, $\varphi_{1,0} = -10$, $\varphi_{2,0} = 10$, $\varphi_{1,1} = \varphi_{2,1} = 0.7$, $\sigma_1 = 5$, $\sigma_2 = 4$; model 5: $c_1 = 0$, $\varphi_{1,0} = \varphi_{2,0} = 0$, $\varphi_{1,1} = -0.3$, $\varphi_{2,1} = 0.3$, $\sigma_1 = 1$, $\sigma_2 = 0.25$.

convenience of the GMAR model as well as of the models of Wong and Li (2001b) and Dueker, Sola, and Spagnolo (2007) is that there is no need to choose a threshold variable (typically y_{t-d}) as in the model of Lanne and Saikkonen (2003) (or in TAR and STAR models).

References

Anderson, T. W. (2003): *An Introduction to Multivariate Statistical Analysis*, 3rd edn. Wiley, Hoboken NJ.

Bec, F., A. Rahbek, and N. Shephard (2008): "The ACR model: a multivariate dynamic mixture autoregression," *Oxford Bulletin of Economics and Statistics*, 70, 583–618.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (2008): *Time Series Analysis: Forecasting and Control*, 4th edn. Wiley, Hoboken NJ.

Carvalho, A. X., and M. A. Tanner (2005): "Mixtures-of-experts of autoregressive time series: asymptotic normality and model specification," *IEEE Transactions on Neural Networks*, 16, 39–56.

Chinn, M. D. (2006): "The (partial) rehabilitation of interest rate parity in the floating rate era: Longer horizons, alternative expectations, and emerging markets," *Journal of International Money and Finance*, 25, 7–21.

Cline, D. B. H. (2007): "Evaluating the Lyapounov exponent and existence of moments for threshold AR–ARCH models," *Journal of Time Series Analysis*, 28, 241–260.

Del Negro, M. and F. Schorfheide (2011): Bayesian Macroeconometrics, Chapter 7 in the *Handbook of Bayesian Econometrics*, ed. by J. F. Geweke, G. Koop, and H. K. van Dijk, Oxford University Press, Oxford.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society: Series B*, 39, 1–38.

Diebold, F. X., J.-H. Lee and G. C. Weinbach (1994): "Regime switching with time-varying transition probabilities," in C. Hargreaves (ed.), Nonstationary Time Series Analysis and Cointegration, pp. 283–302, Oxford University Press, Oxford.

Dueker, M. J., Z. Psaradakis, M. Sola and F. Spagnolo (2011): "Multivariate contemporaneous-threshold autoregressive models," *Journal of Econometrics*, 160, 311–325.

Dueker, M. J., M. Sola and F. Spagnolo (2007): "Contemporaneous threshold autoregressive models: estimation, testing and forecasting," *Journal of Econometrics*, 141, 517–547.

Dunn, P. K., and G. K. Smyth (1996): "Randomized quantile residuals," *Journal of Computational and Graphical Statistics*, 5, 236–244.

Enders, W., and C. W. J. Granger (1998): "Unit-root tests and asymmetric adjustment with an example using the term structure of interest rates," *Journal of Business & Economic Statistics*, 16, 304–11.

Filardo, J. F. (1994): "Business-cycle phases and their transitional dynamics," *Journal of Business and Economic Statistics*, 12, 299–308.

Garcia, R., and P. Perron (1996): "An analysis of the real interest rate under regime shifts," *Review of Economics and Statistics*, 78, 111–125.

Glasbey, C. A. (2001): "Non-linear autoregressive time series with multivariate Gaussian mixtures as marginal distributions," *Journal of the Royal Statistical Society: Series C*, 50, 143–154.

Gourieroux, C., and C. Y. Robert (2006): "Stochastic unit root models," *Econometric Theory*, 22, 1052–1090.

GRANGER, C. W. J. AND T. TERÄSVIRTA (1993): *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford.

HAGGAN, V., AND T. OZAKI (1981): "Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model," *Biometrika*, 68, 189–196.

HAMILTON, J, D. (1994): *Time Series Analysis*. Princeton University Press, Princeton.

KALLIOVIRTA, L. (2012): "Misspecification tests based on quantile residuals," *The Econometrics Journal*, 15, 358-393.

LANNE, M., AND P. SAIKKONEN (2003): "Modeling the U.S. short-term interest rate by mixture autoregressive processes," *Journal of Financial Econometrics*, 1, 96–125.

LE, N. D., R. D. MARTIN, AND A. E. RAFTERY (1996): "Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models," *Journal of the American Statistical Association*, 91, 1504–1515.

MEYN, S., AND R. L. TWEEDIE (2009): *Markov Chains and Stochastic Stability*, 2nd edn. Cambridge University Press, Cambridge.

PALM, F. C., AND P. J. G. VLAAR (1997): "Simple diagnostic procedures for modelling financial time series," *Allgemeines Statistisches Archiv*, 81, 85–101.

REINSEL, G. C. (1997): *Elements of Multivariate Time Series Analysis*, 2nd edn. Springer, New York.

SAIKKONEN, P. (2007): "Stability of mixtures of vector autoregressions with autoregressive conditional heteroskedasticity," *Statistica Sinica*, 17, 221–239.

SILVERMAN, B. W. (1984): "Spline smoothing: The equivalent variable kernel method," *Annals of Statistics*, 12, 898–916.

Sisson, S. A. (2005): "Trans-dimensional Markov chains: A decade of progress and future perspectives," *Journal of the American Statistical Association*, 100, 1077–1089.

Smith, J. Q. (1985): "Diagnostic checks of non-standard time series models," *Journal of Forecasting*, 4, 283–291.

Teräsvirta, T. (1994): "Specification, estimation, and evaluation of smooth transition autoregressive models," *Journal of the American Statistical Association*, 89, 208–218.

Teräsvirta, T., D. Tjøstheim and C. W. J. Granger (2010): *Modelling Nonlinear Economic Time Series*. Oxford University Press, Oxford.

Tierney, L. (1994): "Markov chains for exploring posterior distributions," *Annals of Statistics*, 22, 1701–1762.

Tong, H. (1990): *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.

Tong, H. (2011): "Threshold models in time series analysis – 30 years on," *Statistics and Its Interface*, 4, 107–118.

Wong, C. S., and W. K. Li (2000): "On a mixture autoregressive model," *Journal of the Royal Statistical Society: Series B*, 62, 95–115.

Wong, C. S., and W. K. Li (2001a): "On a mixture autoregressive conditional heteroscedastic model," *Journal of the American Statistical Association*, 96, 982–995.

Wong, C. S., and W. K. Li (2001b): "On a logistic mixture autoregressive model," *Biometrika*, 88, 833–846.

Zeevi, A., R. Meir, and R. J. Adler (2000): "Non-linear models for time series using mixtures of autoregressive models," technical report, Technion and Stanford University, 2000.