Global Research Unit
Working Paper #2018-013

A mixture autoregressive model based on Student's t-distribution

Mika Meitz, University of Helsinki
Daniel Preve, City University of Hong Kong
Pentti Saikkonen, University of Helsinki

© 2018 by Meitz, Preve and Saikkonen. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

A mixture autoregressive model based on Student's t–distribution*

Mika Meitz
University of Helsinki

Daniel Preve
City University of Hong Kong

Pentti Saikkonen
University of Helsinki

Abstract

A new mixture autoregressive model based on Student's t–distribution is proposed. A key feature of our model is that the conditional t–distributions of the component models are based on autoregressions that have multivariate t–distributions as their (low-dimensional) stationary distributions. That autoregressions with such stationary distributions exist is not immediate. Our formulation implies that the conditional mean of each component model is a linear function of past observations and the conditional variance is also time varying. Compared to previous mixture autoregressive models our model may therefore be useful in applications where the data exhibits rather strong conditional heteroskedasticity. Our formulation also has the theoretical advantage that conditions for stationarity and ergodicity are always met and these properties are much more straightforward to establish than is common in nonlinear autoregressive models. An empirical example employing a realized kernel series based on S&P 500 high-frequency data shows that the proposed model performs well in volatility forecasting.

Keywords: Conditional heteroskedasticity; mixture model; regime switching; Student's t–distribution.

*Contact addresses: Mika Meitz, Discipline of Economics, University of Helsinki, P. O. Box 17, FI–00014 University of Helsinki, Finland; e-mail: mika.meitz@helsinki.fi. Daniel Preve, Department of Economics and Finance, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China; e-mail: pdapreve@cityu.edu.hk. 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

Different types of mixture models are in widespread use in various fields. Overviews of mixture models can be found, for example, in the monographs of McLachlan & Peel (2000) and Frühwirth-Schnatter (2006). In this paper, we are concerned with mixture autoregressive models that were introduced by Le et al. (1996) and further developed by Wong & Li (2000, 2001a,b) (for further references, see Kalliovirta et al. (2015)).

In mixture autoregressive models the conditional distribution of the present observation given the past is a mixture distribution where the component distributions are obtained from linear autoregressive models. The specification of a mixture autoregressive model typically requires two choices: choosing a conditional distribution for the component models and choosing a functional form for the mixing weights. In a majority of existing models a Gaussian distribution is assumed whereas, in addition to constants, several different time-varying mixing weights (functions of past observations) have been considered in the literature.

Instead of a Gaussian distribution, Wong et al. (2009) proposed using Student's t–distribution. A major motivation for this comes from the heavier tails of the t–distribution which allow the resulting model to better accommodate for the fat tails encountered in many observed time series, especially in economics and finance. In the model suggested by Wong et al. (2009), the conditional mean and conditional variance of each component model are the same as in the Gaussian case (a linear function of past observations and a constant, respectively), and what changes is the distribution of the independent and identically distributed error term: instead of a standard normal distribution, a Student's t–distribution is used. This is a natural approach to formulate the component models and hence also a mixture autoregressive model based on the t–distribution.

In this paper, we also consider a mixture autoregressive model based on Student's t–distribution, but our specification differs from that used by Wong et al. (2009). Our starting point is the characteristic feature of linear Gaussian autoregressions that stationary distributions (of consecutive observations) as well as conditional distributions are Gaussian. We imitate this feature by using a (multivariate) Student's t–distribution and, as a first step, construct a linear autoregression in which both conditional and (low-dimensional) stationary distributions have Student's t–distributions. This leads to a model where the conditional mean is as in the Gaussian case (a linear function of past observations) whereas the conditional variance is no longer constant but depends on a quadratic form of past observations. These linear models are then used as component models in our new mixture autoregressive model which we call the StMAR model.

Our StMAR model has some very attractive features. Like the model of Wong et al. (2009), it can be useful for modelling time series with regime switching, multimodality, and conditional heteroskedasticity. As the conditional variances of the component models are time-varying, the StMAR model can potentially accommodate for stronger forms of conditional heteroskedasticity than the model of Wong et al. (2009). Our formulation also has the theoretical advantage that, for a pth order model, the stationary distribution of $p+1$ consecutive observations is fully known and is a mixture of particular Student's t–distributions. Moreover, stationarity and ergodicity are simple consequences of the definition of the model and do not require complicated proofs.

Finally, a few notational conventions. All vectors are treated as column vectors and we write $\boldsymbol{x} = (x_1, \ldots, x_n)$ for the vector $\boldsymbol{x}$ where the components x_i may be either scalars or vectors. The

notation $\mathbf{X} \sim n_d(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ signifies that the random vector $\mathbf{X}$ has a d–dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and (positive definite) covariance matrix $\boldsymbol{\Gamma}$. Similarly, by $\mathbf{X} \sim t_d(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \nu)$ we mean that $\mathbf{X}$ has a d–dimensional Student's t–distribution with mean $\boldsymbol{\mu}$, (positive definite) covariance matrix $\boldsymbol{\Gamma}$, and degrees of freedom ν (assumed to satisfy $\nu > 2$); the density function and some properties of the multivariate Student's t–distribution employed are given in an Appendix. The notation $\mathbf{1}_d$ is used for a d–dimensional vector of ones, $\imath_d$ signifies the vector $(1, 0, \ldots, 0)$ of dimension d, and the identity matrix of dimension d is denoted by I_d. The Kronecker product is denoted by $\otimes$, and $vec(A)$ stacks the columns of matrix A on top of one another.

2 Linear Student's t autoregressions

In order to formulate our new mixture model, this section briefly considers linear pth order autoregressions that have multivariate Student's t–distributions as their stationary distributions. First, for motivation and to develop notation, consider a linear Gaussian autoregression z_t $(t = 1, 2, \ldots)$ generated by

$$z_t = \varphi_0 + \sum_{i=1}^{p} \varphi_j z_{t-i} + \sigma e_t, \tag{1}$$

where the error terms e_t are independent and identically distributed with a standard normal distribution, and the parameters satisfy $\varphi_0 \in \mathbb{R}$, $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p) \in \mathbb{S}^p$, and $\sigma > 0$, where

$$\mathbb{S}^p = \{(\varphi_1, \ldots, \varphi_p) \in \mathbb{R}^p : \varphi(z) = 1 - \sum_{i=1}^{p} \varphi_i z^i \neq 0 \ \text{ for } \ |z| \leq 1\} \tag{2}$$

is the stationarity region of a linear pth order autoregression. Denoting $\boldsymbol{z}_t = (z_t, \ldots, z_{t-p+1})$ and $\boldsymbol{z}_t^+ = (z_t, \boldsymbol{z}_{t-1})$, it is well known that the stationary solution z_t to (1) satisfies

$$\begin{aligned} \boldsymbol{z}_t &\sim n_p(\mu \mathbf{1}_p, \boldsymbol{\Gamma}_p), \\ \boldsymbol{z}_t^+ &\sim n_{p+1}(\mu \mathbf{1}_{p+1}, \boldsymbol{\Gamma}_{p+1}), \\ z_t \mid \boldsymbol{z}_{t-1} &\sim n_1(\varphi_0 + \boldsymbol{\varphi}' \boldsymbol{z}_{t-1}, \sigma^2) = n_1(\mu + \boldsymbol{\gamma}_p' \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{z}_{t-1} - \mu \mathbf{1}_p), \sigma^2), \end{aligned} \tag{3}$$

where the last relation defines the conditional distribution of z_t given $\boldsymbol{z}_{t-1}$ and the quantities $\boldsymbol{\Gamma}_p$, γ_0, $\boldsymbol{\gamma}_p$, μ, and $\boldsymbol{\Gamma}_{p+1}$ are defined via

$$\begin{aligned} &vec(\boldsymbol{\Gamma}_p) = (I_{p^2} - (\Phi \otimes \Phi))^{-1} \imath_{p^2} \sigma^2, \quad \Phi = \begin{bmatrix} \varphi_1 \cdots \varphi_{p-1} & \varphi_p \\ I_{p-1} & \mathbf{0}_{p-1} \end{bmatrix}, \\ &\gamma_0 = \sigma^2 + \boldsymbol{\varphi}' \boldsymbol{\Gamma}_p \boldsymbol{\varphi}, \quad \boldsymbol{\gamma}_p = \boldsymbol{\Gamma}_p \boldsymbol{\varphi}, \quad \mu = \varphi_0 / (1 - \varphi_1 - \cdots - \varphi_p), \quad \boldsymbol{\Gamma}_{p+1} = \begin{bmatrix} \gamma_0 & \boldsymbol{\gamma}_p' \\ \boldsymbol{\gamma}_p & \boldsymbol{\Gamma}_p \end{bmatrix}. \end{aligned} \tag{4}$$

Two essential properties of linear Gaussian autoregressions are that they have the distributional features in (3) and the representation in (1).

It is not immediately obvious that linear autoregressions based on Student's t–distribution with similar properties exist (such models have, however, appeared at least in Spanos (1994), Heracleous & Spanos (2006), and Pitt & Walker (2006)). Suppose that for a random vector in $\mathbb{R}^{p+1}$ it holds that $(z, \boldsymbol{z}) \sim t_{p+1}(\mu \mathbf{1}_{p+1}, \boldsymbol{\Gamma}_{p+1}, \nu)$ where $\nu > 2$ (and other notation is as above in (4)). Then (for details,

see the Appendix) the conditional distribution of z given $\boldsymbol{z}$ is $z \mid \boldsymbol{z} \sim t_1(\mu(\boldsymbol{z}), \sigma^2(\boldsymbol{z}), \nu + p)$, where

$$\mu(\boldsymbol{z}) = \varphi_0 + \boldsymbol{\varphi}'\boldsymbol{z}, \qquad \sigma^2(\boldsymbol{z}) = \frac{\nu - 2 + (\boldsymbol{z} - \mu\mathbf{1}_p)'\boldsymbol{\Gamma}_p^{-1}(\boldsymbol{z} - \mu\mathbf{1}_p)}{\nu - 2 + p}\sigma^2. \tag{5}$$

We now state the following theorem (proofs of all theorems are in the Supplementary Material).

Theorem 1. *Suppose $\varphi_0 \in \mathbb{R}$, $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p) \in \mathbb{S}^p$, $\sigma > 0$, and $\nu > 2$. Then there exists a process $\boldsymbol{z}_t = (z_t, \ldots, z_{t-p+1})$ $(t = 0, 1, 2, \ldots)$ with the following properties.*

(i) The process $\boldsymbol{z}_t$ $(t = 1, 2, \ldots)$ is a Markov chain on $\mathbb{R}^p$ with a stationary distribution characterized by the density function $t_p(\mu\mathbf{1}_p, \boldsymbol{\Gamma}_p, \nu)$. When $\boldsymbol{z}_0 \sim t_p(\mu\mathbf{1}_p, \boldsymbol{\Gamma}_p, \nu)$, we have, for $t = 1, 2, \ldots$, that $\boldsymbol{z}_t^+ \sim t_{p+1}(\mu\mathbf{1}_{p+1}, \boldsymbol{\Gamma}_{p+1}, \nu)$ and the conditional distribution of z_t given $\boldsymbol{z}_{t-1}$ is

$$z_t \mid \boldsymbol{z}_{t-1} \sim t_1(\mu(\boldsymbol{z}_{t-1}), \sigma^2(\boldsymbol{z}_{t-1}), \nu + p). \tag{6}$$

(ii) Furthermore, for $t = 1, 2, \ldots$, the process z_t has the representation

$$z_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i \, z_{t-i} + \sigma_t \, \varepsilon_t \tag{7}$$

with conditional variance $\sigma_t^2 = \sigma^2(\boldsymbol{z}_{t-1})$ (see (5)), where the error terms ε_t form a sequence of independent and identically distributed random variables with a marginal $t_1(0, 1, \nu + p)$ distribution and with ε_t independent of $\{z_s,\ s < t\}$.

Results (i) and (ii) in Theorem 1 are comparable to properties (3) and (1) in the Gaussian case. Part (i) shows that both the stationary and conditional distributions of z_t are t–distributions, whereas part (ii) clarifies the connection to standard AR(p) models. In contrast to linear Gaussian autoregressions, in this t–distributed case z_t is conditionally heteroskedastic and has an 'AR(p)–ARCH(p)' representation (here ARCH refers to autoregressive conditional heteroskedasticity).

3 A mixture autoregressive model based on Student's t–distribution

3.1 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 mixture autoregressive models for which the conditional density function of y_t given its past, $f(\cdot \mid \mathcal{F}_{t-1})$, is of the form

$$f(y_t \mid \mathcal{F}_{t-1}) = \sum_{m=1}^{M} \alpha_{m,t} f_m(y_t \mid \mathcal{F}_{t-1}), \tag{8}$$

where 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), and the $f_m(\cdot \mid \mathcal{F}_{t-1})$, $m = 1, \ldots, M$, describe the conditional densities of M autoregressive component models. Different mixture models are obtained with different specifications of the mixing weights $\alpha_{m,t}$ and the conditional densities $f_m(\cdot \mid \mathcal{F}_{t-1})$.

Starting with the specification of the conditional densities $f_m(\cdot \mid \mathcal{F}_{t-1})$, a common choice has been to assume the component models to be linear Gaussian autoregressions. For the mth component

model ($m = 1, \ldots, M$), denote the parameters of a pth order linear autoregression with $\varphi_{m,0} \in \mathbb{R}$, $\boldsymbol{\varphi}_m = (\varphi_{m,1}, \ldots, \varphi_{m,p}) \in \mathbb{S}^p$, and $\sigma_m > 0$. Also set $\boldsymbol{y}_{t-1} = (y_{t-1}, \ldots, y_{t-p})$. In the Gaussian case, the conditional densities in (8) take the form ($m = 1, \ldots, M$)

$$f_m(y_t \mid \mathcal{F}_{t-1}) = \frac{1}{\sigma_m}\phi\Big(\frac{y_t - \mu_{m,t}}{\sigma_m}\Big),$$

where $\phi(\cdot)$ signifies the density function of a standard normal random variable, $\mu_{m,t} = \varphi_{m,0} + \boldsymbol{\varphi}_m'\boldsymbol{y}_{t-1}$ is the conditional mean function (of component m), and $\sigma_m^2 > 0$ is the conditional variance (of component m), often assumed to be constant. Instead of a Gaussian density, Wong et al. (2009) consider the case where $f_m(\cdot \mid \mathcal{F}_{t-1})$ is the density of Student's t–distribution with conditional mean and variance as above, $\mu_{m,t} = \varphi_{m,0} + \boldsymbol{\varphi}_m'\boldsymbol{y}_{t-1}$ and a constant σ_m^2, respectively.

In this paper, we also consider a mixture autoregressive model based on Student's t–distribution, but our formulation differs from that used by Wong et al. (2009). In Theorem 1 it was seen that linear autoregressions based on Student's t–distribution naturally lead to the conditional distribution $t_1(\mu(\cdot), \sigma^2(\cdot), \nu + p)$ in (6). Motivated by this, we consider a mixture autoregressive model in which the conditional densities $f_m(y_t \mid \mathcal{F}_{t-1})$ in (8) are specified as

$$f_m(y_t \mid \mathcal{F}_{t-1}) = t_1(y_t; \mu_{m,t}, \sigma_{m,t}^2, \nu_m + p), \tag{9}$$

where the expressions for $\mu_{m,t} = \mu_m(\boldsymbol{y}_{t-1})$ and $\sigma_{m,t}^2 = \sigma_m^2(\boldsymbol{y}_{t-1})$ are as in (5) except that $\boldsymbol{z}$ is replaced with $\boldsymbol{y}_{t-1}$ and all the quantities therein are defined using the regime specific parameters $\varphi_{m,0}$, $\boldsymbol{\varphi}_m$, σ_m, and ν_m (whenever appropriate a subscript m is added to previously defined notation, e.g., μ_m or $\boldsymbol{\Gamma}_{m,p}$). A key difference to the model of Wong et al. (2009) is that the conditional variance of component m is not constant but a function of $\boldsymbol{y}_{t-1}$. An explicit expression for the density in (9) can be obtained from the Appendix and is

$$f_m(y_t \mid \mathcal{F}_{t-1}) = C(\nu_m)\sigma_{m,t}^{-1}\Big(1 + (\nu_m + p - 2)^{-1}\Big(\frac{y_t - \mu_{m,t}}{\sigma_{m,t}}\Big)^2\Big)^{-\frac{1+\nu_m+p}{2}}, \tag{10}$$

where $C(\nu) = \frac{\Gamma((1+\nu+p)/2)}{(\pi(\nu+p-2))^{1/2}\Gamma((\nu+p)/2)}$ (and $\Gamma(\cdot)$ signifies the gamma function).

Now consider the choice of the mixing weights $\alpha_{m,t}$ in (8). The most basic choice is to use constant mixing weights as in Wong & Li (2000) and Wong et al. (2009). Several different time-varying mixing weights have also been suggested, see, e.g., Wong & Li (2001a), Glasbey (2001), Lanne & Saikkonen (2003), Dueker et al. (2007), and Kalliovirta et al. (2015, 2016).

In this paper, we propose mixing weights that are similar to those used by Glasbey (2001) and Kalliovirta et al. (2015). Specifically, we set

$$\alpha_{m,t} = \frac{\alpha_m t_p(\boldsymbol{y}_{t-1}; \mu_m \mathbf{1}_p, \boldsymbol{\Gamma}_{m,p}, \nu_m)}{\sum_{n=1}^M \alpha_n t_p(\boldsymbol{y}_{t-1}; \mu_n \mathbf{1}_p, \boldsymbol{\Gamma}_{n,p}, \nu_n)}, \tag{11}$$

where the $\alpha_m \in (0, 1)$, $m = 1, \ldots, M$, are unknown parameters satisfying $\sum_{m=1}^M \alpha_m = 1$. Note that the Student's t density appearing in (11) corresponds to the stationary distribution in Theorem 1(i): If the y_t's were generated by a linear Student's t autoregression described in Section 2 (with a subscript m added to all the notation therein), the stationary distribution of $\boldsymbol{y}_{t-1}$ would be characterized by $t_p(\boldsymbol{y}_{t-1}; \mu_m \mathbf{1}_p, \boldsymbol{\Gamma}_{m,p}, \nu_m)$. Our definition of the mixing weights in (11) is different from that used in

Glasbey (2001) and Kalliovirta et al. (2015) in that these authors employed the $n_p(\boldsymbol{y}_{t-1}; \mu_m \mathbf{1}_p, \boldsymbol{\Gamma}_{m,p})$ density (corresponding to the stationary distribution of a linear Gaussian autoregression) instead of the Student's t density $t_p(\boldsymbol{y}_{t-1}; \mu_m \mathbf{1}_p, \boldsymbol{\Gamma}_{m,p}, \nu_m)$ we use.

3.2 The Student's t mixture autoregressive model

Equations (8), (9), and (11) define a model we call the Student's t mixture autoregressive, or StMAR, model. When the autoregressive order p or the number of mixture components M need to be emphasized we refer to an StMAR(p,M) model. We collect the unknown parameters of an StMAR model in the vector $\boldsymbol{\theta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_M, \alpha_1, \ldots, \alpha_{M-1})$ ($(M(p+4)-1) \times 1$), where $\boldsymbol{\vartheta}_m = (\varphi_{m,0}, \boldsymbol{\varphi}_m, \sigma_m^2, \nu_m)$ (with $\boldsymbol{\varphi}_m \in \mathbb{S}^p$, $\sigma_m^2 > 0$, and $\nu_m > 2$) contains the parameters of each component model ($m = 1, \ldots, M$) and the α_m's are the parameters appearing in the mixing weights (11); the parameter α_M is not included due to the restriction $\sum_{m=1}^{M} \alpha_m = 1$.

The StMAR model can also be presented in an alternative (but equivalent) form. To this end, let $P_{t-1}(\cdot)$ signify the conditional probability of the indicated event given $\mathcal{F}_{t-1}$, and let $\varepsilon_{m,t}$ be a sequence of independent and identically distributed random variables with a $t_1(0, 1, \nu_m + p)$ distribution such that $\varepsilon_{m,t}$ is independent of $\{y_{t-j},\ j > 0\}$ ($m = 1, \ldots, M$). Furthermore, let $\boldsymbol{s}_t = (s_{1,t}, \ldots, s_{M,t})$ be a sequence of (unobserved) M–dimensional random vectors such that, conditional on $\mathcal{F}_{t-1}$, $\boldsymbol{s}_t$ and $\varepsilon_{m,t}$ are independent (for all m). 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 $P_{t-1}(s_{m,t} = 1) = \alpha_{m,t}$, $m = 1, \ldots, M$. Now y_t can be expressed as

$$y_t = \sum_{m=1}^{M} s_{m,t}(\mu_{m,t} + \sigma_{m,t}\varepsilon_{m,t}) = \sum_{m=1}^{M} s_{m,t}(\varphi_{m,0} + \boldsymbol{\varphi}_m' \boldsymbol{y}_{t-1} + \sigma_{m,t}\varepsilon_{m,t}), \tag{12}$$

where $\sigma_{m,t}$ is as in (9). This formulation suggests that the mixing weights $\alpha_{m,t}$ can be thought of as (conditional) probabilities that determine which one of the M autoregressive components of the mixture generates the observation y_t.

It turns out that the StMAR model has some very attractive theoretical properties; the carefully chosen conditional densities in (9) and the mixing weights in (11) are crucial in obtaining these properties. The following theorem shows that there exists a choice of initial values $\boldsymbol{y}_0$ such that $\boldsymbol{y}_t$ is a stationary and ergodic Markov chain. Importantly, an explicit expression for the stationary distribution is also provided.

Theorem 2. *Consider the StMAR process y_t generated by (8), (9), and (11) (or (12) and (11)) with the conditions $\boldsymbol{\varphi}_m \in \mathbb{S}^p$ and $\nu_m > 2$ satisfied for all $m = 1, \ldots, M$. Then $\boldsymbol{y}_t = (y_t, \ldots, y_{t-p+1})$ ($t = 1, 2, \ldots$) is a Markov chain on $\mathbb{R}^p$ with a stationary distribution characterized by the density*

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m t_p(\boldsymbol{y}; \mu_m \mathbf{1}_p, \boldsymbol{\Gamma}_{m,p}, \nu_m).$$

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

The stationary distribution of $\boldsymbol{y}_t$ is a mixture of M p–dimensional t–distributions with constant mixing weights α_m. Hence, moments of the stationary distribution of order smaller than $\min(\nu_1, \ldots, \nu_M)$ exist and are finite. As can be seen from the proof of Theorem 2 (in the Supplementary Material), the

stationary distribution of the vector $(y_t, \boldsymbol{y}_{t-1})$ is also a mixture of M t–distributions with density of the same form, $\sum_{m=1}^{M} \alpha_m t_{p+1}(\mu_m \mathbf{1}_{p+1}, \boldsymbol{\Gamma}_{m,p+1}, \nu_m)$. Thus the mean, variance, and first p autocovariances of y_t are (here the connection between $\gamma_{m,j}$ and $\boldsymbol{\Gamma}_{m,p+1}$ is as in (4))

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

Subvectors of $(y_t, \boldsymbol{y}_{t-1})$ also have stationary distributions that belong to the same family (but this does not hold for higher dimensional vectors such as $(y_{t+1}, y_t, \boldsymbol{y}_{t-1})$).

The fact that an explicit expression for the stationary (marginal) distribution of the StMAR model is available is not only convenient but also quite exceptional among mixture autoregressive models or other related nonlinear autoregressive models (such as threshold or smooth transition models). Previously, similar results have been obtained by Glasbey (2001) and Kalliovirta et al. (2015) in the context of mixture autoregressive models that are of the same form but based on the Gaussian distribution (for a few rather simple first order examples involving other models, see Tong (2011, Section 4.2)).

From the definition of the model, the conditional mean and variance of y_t are obtained as

$$E[y_t \mid \mathcal{F}_{t-1}] = \sum_{m=1}^{M} \alpha_{m,t} \mu_{m,t}, \quad Var[y_t \mid \mathcal{F}_{t-1}] = \sum_{m=1}^{M} \alpha_{m,t} \sigma_{m,t}^2 + \sum_{m=1}^{M} \alpha_{m,t} \left(\mu_{m,t} - \sum_{n=1}^{M} \alpha_{n,t} \mu_{n,t} \right)^2. \quad (13)$$

Except for the different definition of the mixing weights, the conditional mean is as in the Gaussian mixture autoregressive model of Kalliovirta et al. (2015). This is due to the well-known fact that in the multivariate t–distribution the conditional mean is of the same linear form as in the multivariate Gaussian distribution. However, unlike in the Gaussian case, the conditional variance of the multivariate t–distribution is not constant. Therefore, in (13) we have the time-varying variance component $\sigma_{m,t}^2$ which in the models of Kalliovirta et al. (2015) and Wong et al. (2009) is constant (in the latter model the mixing weights are also constants). In (13) both the mixing weights $\alpha_{m,t}$ and the variance components $\sigma_{m,t}^2$ are functions of $\boldsymbol{y}_{t-1}$, implying that the conditional variance exhibits nonlinear autoregressive conditional heteroskedasticity. Compared to the aforementioned previous models our model may therefore be useful in applications where the data exhibits rather strong conditional heteroskedasticity.

4 Estimation

The parameters of an StMAR model can be estimated by the method of maximum likelihood (details of the numerical optimization methods employed and of simulation experiments are available in the Supplementary Material). As the stationary distribution of the StMAR process is known it is even possible to make use of initial values and construct the exact likelihood function and obtain exact maximum likelihood estimates. Assuming the observed data $y_{-p+1}, \ldots, y_0, y_1, \ldots, y_T$ and stationary initial values, the log-likelihood function takes the form

$$L_T(\boldsymbol{\theta}) = \log \left(\sum_{m=1}^{M} \alpha_m t_p(\boldsymbol{y}_0; \mu_m \mathbf{1}_p, \boldsymbol{\Gamma}_{m,p}, \nu_m) \right) + \sum_{t=1}^{T} l_t(\boldsymbol{\theta}), \quad (14)$$

where

$$l_t(\boldsymbol{\theta}) = \log\Bigg(\sum_{m=1}^{M} \alpha_{m,t} t_1(y_t; \mu_{m,t}, \sigma^2_{m,t}, \nu_m + p)\Bigg). \tag{15}$$

An explicit expression for the density appearing in (15) is given in (10), and the notation for $\mu_{m,t}$ and $\sigma^2_{m,t}$ is explained after (9). Although not made explicit, $\alpha_{m,t}$, $\mu_{m,t}$, and $\sigma^2_{m,t}$, as well as the quantities μ_m, $\boldsymbol{\gamma}_{m,p}$, and $\boldsymbol{\Gamma}_{m,p}$, depend on the parameter vector $\boldsymbol{\theta}$.

In (14) it has been assumed that the initial values $\boldsymbol{y}_0$ 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 (14). In what follows we assume that estimation is based on this conditional log-likelihood, namely $L_T^{(c)}(\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T} l_t(\boldsymbol{\theta})$ which we, for convenience, have also scaled with the sample size. Maximizing $L_T^{(c)}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ yields the maximum likelihood estimator denoted by $\hat{\boldsymbol{\theta}}_T$.

The permissible parameter space of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\Theta}$, needs to be constrained in various ways. The stationarity conditions $\boldsymbol{\varphi}_m \in \mathbb{S}^p$, the positivity of the variances σ^2_m, and the conditions $\nu_m > 2$ ensuring existence of second moments are all assumed to hold (for $m = 1, \ldots, M$). Throughout we assume that the number of mixture components M is known, and this also entails the requirement that the parameters α_m ($m = 1, \ldots, M$) are strictly positive (and strictly less than unity whenever $M > 1$). Further restrictions are required to ensure identification. Denoting the true parameter value by $\boldsymbol{\theta}_0$ and assuming stationary initial values, the condition needed is that $l_t(\boldsymbol{\theta}) = l_t(\boldsymbol{\theta}_0)$ almost surely only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. An additional assumption needed for this is

$$\alpha_1 > \cdots > \alpha_M > 0 \quad \text{and} \quad \boldsymbol{\vartheta}_i = \boldsymbol{\vartheta}_j \text{ only if } 1 \leq i = j \leq M. \tag{16}$$

From a practical point of view this assumption is not restrictive because what it essentially requires is that the M component models cannot be 'relabeled' and the same StMAR model obtained. We summarize the restrictions imposed on the parameter space as follows.

Assumption 1. *The true parameter value $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a compact subset of $\{\boldsymbol{\theta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_M, \alpha_1, \ldots, \alpha_{M-1}) \in \mathbb{R}^{M(p+3)} \times (0,1)^{M-1} : \boldsymbol{\varphi}_m \in \mathbb{S}^p, \sigma^2_m > 0,$ and $\nu_m > 2$ for all $m = 1, \ldots, M,$ and (16) holds$\}$.*

Asymptotic properties of the maximum likelihood estimator can now be established under conventional high-level conditions. Denote $\mathcal{I}(\boldsymbol{\theta}) = E\big[\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\big]$ and $\mathcal{J}(\boldsymbol{\theta}) = E\big[\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\big]$.

Theorem 3. *Suppose y_t is generated by the stationary and ergodic StMAR process of Theorem 2 and that Assumption 1 holds. Then $\hat{\boldsymbol{\theta}}_T$ is strongly consistent, i.e., $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}_0$ almost surely. Suppose further that (i) $T^{1/2}\frac{\partial}{\partial \boldsymbol{\theta}} L_T^{(c)}(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathcal{I}(\boldsymbol{\theta}_0))$ with $\mathcal{I}(\boldsymbol{\theta}_0)$ finite and positive definite, (ii) $\mathcal{J}(\boldsymbol{\theta}_0) = -\mathcal{I}(\boldsymbol{\theta}_0)$, and (iii) $E\big[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \big|\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\big|\big] < \infty$ for some $\boldsymbol{\Theta}_0$, a compact convex set contained in the interior of $\boldsymbol{\Theta}$ that has $\boldsymbol{\theta}_0$ as an interior point. Then $T^{1/2}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} N\big(0, -\mathcal{J}(\boldsymbol{\theta}_0)^{-1}\big)$.*

Of the conditions in this theorem, (i) states that a central limit theorem holds for the score vector (evaluated at $\boldsymbol{\theta}_0$) and that the information matrix is positive definite, (ii) is the information matrix equality, and (iii) ensures the uniform convergence of the Hessian matrix (in some neighbourhood of $\boldsymbol{\theta}_0$). These conditions are standard but their verification may be tedious.

Theorem 3 shows that the conventional limiting distribution applies to the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_T$ which implies the applicability of standard likelihood-based tests. It is worth noting,

however, that here a correct specification of the number of autoregressive components M is required. In particular, if the number of component models is chosen too large then some parameters of the model are not identified and, consequently, the result of Theorem 3 and the validity of the related tests break down. This particularly happens when one tests for the number of component models. Such tests for mixture autoregressive models with Gaussian conditional densities (see (8)) are developed by Meitz & Saikkonen (2017). The testing problem is highly nonstandard and extending their results to the present case is beyond the scope of this paper.

Instead of formal tests, in our empirical application we use information criteria to infer which model fits the data best. Similar approaches have also been used by Wong et al. (2009) and others. Note that once the number of regimes is (correctly) chosen, standard likelihood-based inference can be used to choose regime-wise autoregressive orders and to test other hypotheses of interest.

5 Empirical example

Modeling and forecasting financial market volatility is key to manage risk. In this application we use the realized kernel of Barndorff-Nielsen et al. (2008) as a proxy for latent volatility. We obtained daily realized kernel data over the period 3 January 2000 through 20 May 2016 for the S&P 500 index from the Oxford-Man Institute's Realized Library v0.2 (Heber et al., 2009). Figure 1 shows the in-sample period (Jan 3, 2000–June 3, 2014; 3597 observations) for the S&P 500 realized kernel data (RK_t), which is nonnegative with a distribution exhibiting substantial skewness and excess kurtosis (sample skewness 14.3, sample kurtosis 380.8). We follow the related literature which frequently use logarithmic realized kernel ($\log(\text{RK}_t)$), to avoid imposing additional parameter constraints, and to obtain a more symmetric distribution, often taken to be approximately Gaussian. The $\log(\text{RK}_t)$ data, also shown in Figure 1, has a sample skewness of 0.5 and kurtosis of 3.5. Visual inspection of the time series plots of the RK_t and $\log(\text{RK}_t)$ data suggests that the two series exhibit changes at least in levels and potentially also in variability. A kernel estimate of the density function of the $\log(\text{RK}_t)$ series also suggest the potential presence of multiple regimes.

Table 1 reports estimation results for three selected StMAR models (for further details, see the Supplementary Material). Following Wong & Li (2001a), Wong et al. (2009), and Li et al. (2015), we use information criteria for model comparison. For the $\log(\text{RK}_t)$ data in-sample period the Akaike information criterion (AIC) favours the StMAR(4,3) model, the Hannan-Quinn information criterion (HQC) the StMAR(4,2) model, and the Bayesian information criterion (BIC) the simpler StMAR(4,1) model. In view of the approximate standard errors in Table 1, the estimation accuracy appears quite reasonable except for the degrees of freedom parameters. Taking the sum of the autoregressive parameters as a measure of persistence, we find that the estimated persistence for the first regime of the StMAR(4,2) is 0.909 and 0.489 for the second regime, suggesting that persistence is rather strong in the first regime and moderate in the second regime.

Numerous alternative models for volatility proxies have been proposed. We employ Corsi's (2009) heterogeneous autoregressive (HAR) model as it is arguably the most popular reference model for forecasting proxies such as the realized kernel. We also consider a pth-order autoregression as the AR(p) often performs well in volatility proxy forecasting. The StMAR models are estimated using maximum likelihood, and the reference AR and HAR models by ordinary least squares. We use a fixed scheme, where the parameters of our volatility models are estimated just once using data from

Table 1: Parameter estimates for three selected StMAR models and the $\log(\mathrm{RK}_t)$ data over the period 3 January 2000 – 3 June 2014. Numbers in parentheses are standard errors based on a numerical Hessian.

	StMAR$(4,1)$		StMAR$(4,2)$		StMAR$(4,3)$	
$\varphi_{1,0}$	-0.746	(0.089)	-0.851	(0.112)	-3.667	(0.727)
$\varphi_{1,1}$	0.428	(0.017)	0.432	(0.024)	0.331	(0.035)
$\varphi_{1,2}$	0.224	(0.019)	0.221	(0.025)	0.169	(0.034)
$\varphi_{1,3}$	0.121	(0.019)	0.122	(0.025)	0.055	(0.033)
$\varphi_{1,4}$	0.150	(0.017)	0.134	(0.024)	0.093	(0.033)
σ_1^2	0.298	(0.011)	0.285	(0.015)	0.293	(0.016)
ν_1	11.999	(1.109)	10.510	(1.426)	18.328	(1.814)
$\varphi_{2,0}$			-5.381	(1.007)	-1.013	(0.341)
$\varphi_{2,1}$			0.289	(0.046)	0.509	(0.038)
$\varphi_{2,2}$			0.129	(0.048)	0.179	(0.042)
$\varphi_{2,3}$			0.023	(0.046)	0.043	(0.045)
$\varphi_{2,4}$			0.047	(0.052)	0.153	(0.036)
σ_2^2			0.287	(0.022)	0.327	(0.024)
ν_2			29.031	(1.595)	12.977	(2.200)
$\varphi_{3,0}$					-3.639	(1.243)
$\varphi_{3,1}$					0.208	(0.072)
$\varphi_{3,2}$					0.198	(0.082)
$\varphi_{3,3}$					0.219	(0.067)
$\varphi_{3,4}$					-0.010	(0.079)
σ_3^2					0.167	(0.025)
ν_3					22.008	(2.697)
α_1			0.724	(0.064)	0.459	(0.088)
α_2					0.342	(0.099)
$TL_T^{(c)}(\hat{\boldsymbol{\theta}}_T)$	-2854.153		-2832.665		-2820.077	
AIC	5722.306		5695.330		5686.154	
HQC	5737.741		5728.406		5736.870	
BIC	5765.613		5788.131		5828.449	

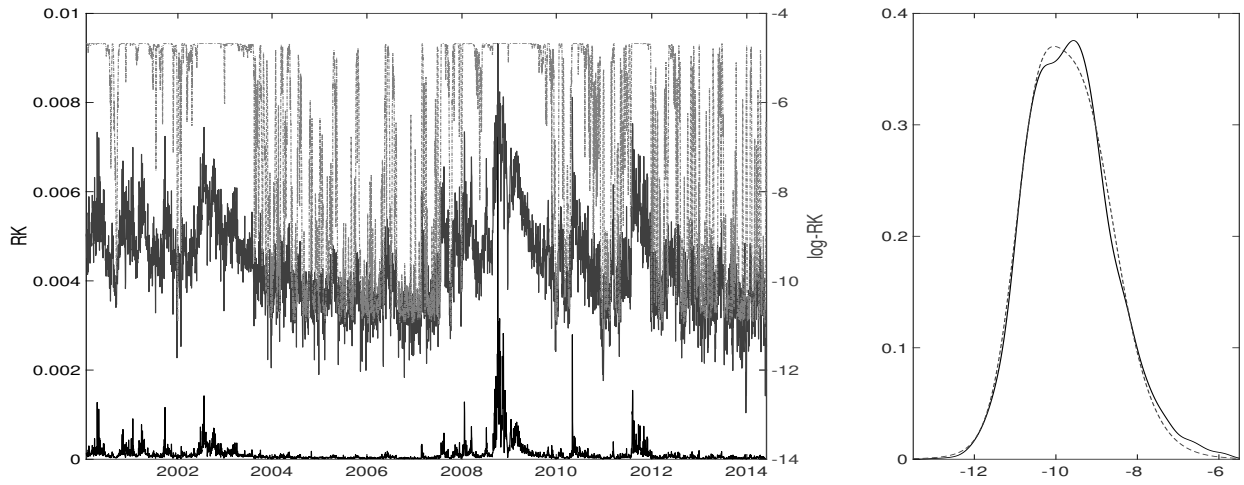

Figure 1: Left panel: Daily RK_t (lower solid) and $\log(\mathrm{RK}_t)$ (upper solid), and mixing weights based on the estimates of the StMAR(4,2) model in Table 1 (dot-dash) for the $\log(\mathrm{RK}_t)$ series. The mixing weights $\hat{\alpha}_{1,t}$ are scaled from (0, 1) to $(\min \log(\mathrm{RK}_t), \max \log(\mathrm{RK}_t))$. Right panel: A kernel density estimate of the $\log(\mathrm{RK}_t)$ observations (solid), and the mixture density (dashes) implied by the same StMAR model as in the left panel.

Table 2: The percentage shares of cumulative realized kernel observations that belong to the 99%, 95% and 90% one-sided upper prediction intervals based on the distribution of 500,000 simulated conditional sample paths.

	Daily			*Weekly*		
	99%	95%	90%	99%	95%	90%
AR(11)	98.99	95.97	90.52	96.54	91.26	86.18
HAR	98.59	94.76	90.52	96.14	91.06	86.99
StMAR(4,1)	98.99	95.97	92.14	98.17	95.12	90.24
StMAR(4,2)	99.19	95.97	92.54	97.97	94.92	90.65
StMAR(4,3)	99.19	96.37	92.94	98.37	94.72	90.65
	Biweekly			*Monthly*		
	99%	95%	90%	99%	95%	90%
AR(11)	94.05	89.53	85.63	94.11	88.63	85.47
HAR	93.63	88.71	84.80	91.79	87.37	83.79
StMAR(4,1)	97.33	93.22	90.76	97.89	93.89	91.79
StMAR(4,2)	97.33	93.22	90.76	97.26	94.11	91.16
StMAR(4,3)	97.54	93.22	90.97	97.89	94.32	91.37

Jan 3, 2000–June 3, 2014. These estimates are then used to generate all forecasts. The remaining 496 observations of our sample are used to compare the forecasts from the alternative models. As discussed in Kalliovirta et al. (2016), computing multi-step-ahead forecasts for mixture models like the StMAR is rather complicated. For this reason we use computer driven forecasts to predict future volatility: For each out-of-sample date T, and for each alternative model, we simulate 500,000 sample paths. Each path is of length 22 (representing one trading month) and conditional on the information available at date T. In these simulations unknown parameters are replaced by their estimates. As the simulated paths are for $\log(\mathrm{RK}_t)$, and our object of interest is RK_t, an exponential transformation is applied.

We examine daily, weekly (5 day), biweekly (10 day), and monthly (22 day) volatility forecasts generated by the alternative models; for instance, the weekly volatility forecast at date T is the forecast for $\mathrm{RK}_{T+1} + \cdots + \mathrm{RK}_{T+5}$ (the 5-day-ahead *cumulative* realized kernel). Table 2 reports the percentage shares of (1, 5, 10, and 22-day) cumulative RK_t out-of-sample observations that belong to the 99%, 95%, and 90% one-sided upper prediction intervals based on the distribution of the simulated sample paths; these upper prediction intervals for volatility are related to higher levels of risk in financial markets. Overall, it is seen that the empirical coverage rates of the StMAR based prediction intervals are closer to the nominal levels than the ones obtained with the reference models. By comparison, the accuracy of the prediction intervals obtained with the popular HAR model quickly degrade as the forecast period increases. The StMAR model performs well also when two-sided prediction intervals and point forecast accuracy are considered (for details, see the Supplementary Material).

Acknowledgement

The authors thank the Academy of Finland for financial support.

Supplementary material

The supplementary material includes proofs of Theorems 1–3, information on the numerical optimization methods employed for maximum likelihood estimation, simulation experiments, and further details of the empirical example.

Appendix

Properties of the multivariate Student's t–distribution

The standard form of the density function of the multivariate Student's t–distribution with ν degrees of freedom and dimension d is (see, e.g., Kotz & Nadarajah (2004, p. 1))

$$f(\boldsymbol{x}) = \frac{\Gamma\left((d+\nu)/2\right)}{(\pi\nu)^{d/2}\,\Gamma(\nu/2)} \det(\boldsymbol{\Sigma})^{-1/2} \left(1 + \nu^{-1}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-\frac{d+\nu}{2}},$$

where $\Gamma(\cdot)$ is the gamma function and $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}$ $(d \times d)$, a symmetric positive definite matrix, are parameters. For a random vector $\mathbf{X}$ possessing this density, the mean and covariance are $E[\mathbf{X}] = \boldsymbol{\mu}$

and $Cov[\mathbf{X}] = \mathbf{\Gamma} = \frac{\nu}{\nu-2}\mathbf{\Sigma}$ (assuming $\nu > 2$). The density can be expressed in terms of $\boldsymbol{\mu}$ and $\mathbf{\Gamma}$ as

$$f(\boldsymbol{x}) = \frac{\Gamma\left((d+\nu)/2\right)}{(\pi(\nu-2))^{d/2}\,\Gamma(\nu/2)} \det(\mathbf{\Gamma})^{-1/2} \left(1 + (\nu-2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})'\mathbf{\Gamma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{-\frac{d+\nu}{2}}.$$

This form of the density function, denoted by $t_d(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Gamma}, \nu)$, is used in this paper, and the notation $\mathbf{X} \sim t_d(\boldsymbol{\mu}, \mathbf{\Gamma}, \nu)$ is used for a random vector $\mathbf{X}$ possessing this density. Condition $\nu > 2$ and positive definiteness of $\mathbf{\Gamma}$ will be tacitly assumed.

For marginal and conditional distributions, partition $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where the components have dimensions d_1 and d_2 ($d_1 + d_2 = d$). Conformably partition $\boldsymbol{\mu}$ and $\mathbf{\Gamma}$ as $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{12}' & \mathbf{\Gamma}_{22} \end{bmatrix}.$$

Then the marginal distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ are $t_{d_1}(\boldsymbol{\mu}_1, \mathbf{\Gamma}_{11}, \nu)$ and $t_{d_2}(\boldsymbol{\mu}_2, \mathbf{\Gamma}_{22}, \nu)$, respectively. The conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is also a t–distribution, namely (see Ding (2016, Sec. 2))

$$\mathbf{X}_1 \mid (\mathbf{X}_2 = \boldsymbol{x}_2) \sim t_{d_1}(\boldsymbol{\mu}_{1|2}(\boldsymbol{x}_2), \mathbf{\Gamma}_{1|2}(\boldsymbol{x}_2), \nu + d_2),$$

where $\boldsymbol{\mu}_{1|2}(\boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and $\mathbf{\Gamma}_{1|2}(\boldsymbol{x}_2) = \frac{\nu-2+(\boldsymbol{x}_2-\boldsymbol{\mu}_2)'\mathbf{\Gamma}_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)}{\nu-2+d_2}(\mathbf{\Gamma}_{11} - \mathbf{\Gamma}_{12}\mathbf{\Gamma}_{22}^{-1}\mathbf{\Gamma}_{12}')$. Furthermore, $t_d(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Gamma}, \nu) = t_{d_1}(\boldsymbol{x}_1; \boldsymbol{\mu}_{1|2}(\boldsymbol{x}_2), \mathbf{\Gamma}_{1|2}(\boldsymbol{x}_2), \nu + d_2)\; t_{d_2}(\boldsymbol{x}_2; \boldsymbol{\mu}_2, \mathbf{\Gamma}_{22}, \nu)$.

Now consider a special case: a $(p+1)$–dimensional random vector $\mathbf{X} \sim t_{p+1}(\mu\mathbf{1}_{p+1}, \mathbf{\Gamma}_{p+1}, \nu)$, where $\mu \in \mathbb{R}$ and $\mathbf{\Gamma}_{p+1}$ is a symmetric positive definite Toeplitz matrix. Note that the mean vector $\mu\mathbf{1}_{p+1}$ and the covariance matrix $\mathbf{\Gamma}_{p+1}$ have structures similar to those of the mean and covariance matrix of a $(p+1)$–dimensional realization of a second order stationary process. More specifically, assume that $\mathbf{\Gamma}_{p+1}$ is the covariance matrix of a second order stationary AR(p) process.

Partition $\mathbf{X}$ as $\mathbf{X} = (X_1, \mathbf{X}_2) = (\mathbf{X}_1, X_{p+1})$ with X_1 and X_{p+1} real valued and $\mathbf{X}_1$ and $\mathbf{X}_2$ both $p \times 1$ vectors. The marginal distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ are $\mathbf{X}_1 \sim t_p(\mu\mathbf{1}_p, \mathbf{\Gamma}_p, \nu)$ and $\mathbf{X}_2 \sim t_p(\mu\mathbf{1}_p, \mathbf{\Gamma}_p, \nu)$, where the (symmetric positive definite Toeplitz) matrix $\mathbf{\Gamma}_p = Cov[\mathbf{X}_1] = Cov[\mathbf{X}_2]$ is obtained from $\mathbf{\Gamma}_{p+1}$ by deleting the first row and first column or, equivalently, the last row and last column (here the specific structures of $\mu\mathbf{1}_{p+1}$ and $\mathbf{\Gamma}_{p+1}$ are used). The conditional distribution of X_1 given $\mathbf{X}_2 = \boldsymbol{x}_2$ is

$$X_1 \mid (\mathbf{X}_2 = \boldsymbol{x}_2) \sim t_1(\mu(\boldsymbol{x}_2), \sigma^2(\boldsymbol{x}_2), \nu + p),$$

where expressions for $\mu(\boldsymbol{x}_2)$ and $\sigma^2(\boldsymbol{x}_2)$ can be obtained from above as follows. Partition $\mathbf{\Gamma}_{p+1}$ as

$$\mathbf{\Gamma}_{p+1} = \begin{bmatrix} \gamma_0 & \boldsymbol{\gamma}_p' \\ \boldsymbol{\gamma}_p & \mathbf{\Gamma}_p \end{bmatrix},$$

and denote $\boldsymbol{\varphi} = \mathbf{\Gamma}_p^{-1}\boldsymbol{\gamma}_p$ and $\sigma^2 = \gamma_0 - \boldsymbol{\gamma}_p'\mathbf{\Gamma}_p^{-1}\boldsymbol{\gamma}_p$ ($\sigma^2 > 0$ as $\mathbf{\Gamma}_{p+1}$ is positive definite). From above,

$$\begin{aligned} \mu(\boldsymbol{x}_2) &= \boldsymbol{\mu}_{1|2}(\boldsymbol{x}_2) = \mu + \boldsymbol{\gamma}_p'\mathbf{\Gamma}_p^{-1}(\boldsymbol{x}_2 - \mu\mathbf{1}_p) = \mu(1 - \boldsymbol{\gamma}_p'\mathbf{\Gamma}_p^{-1}\mathbf{1}_p) + \boldsymbol{\varphi}'\boldsymbol{x}_2, \\ \sigma^2(\boldsymbol{x}_2) &= \mathbf{\Gamma}_{1|2}(\boldsymbol{x}_2) = \frac{\nu - 2 + (\boldsymbol{x}_2 - \mu\mathbf{1}_p)'\mathbf{\Gamma}_p^{-1}(\boldsymbol{x}_2 - \mu\mathbf{1}_p)}{\nu - 2 + p}\sigma^2. \end{aligned}$$

References

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* **76**, 1481–1536.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *J. Finan. Economet.* **7**, 174–196.

Ding, P. (2016). On the conditional distribution of the multivariate t distribution. *Am. Statistician* **70**, 293–295.

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

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Glasbey, C. A. (2001). Non-linear autoregressive time series with multivariate Gaussian mixtures as marginal distributions. *J. R. Statist. Soc. C* **50**, 143–154.

Heber, G., Lunde, A., Shephard, N. & Sheppard, K. (2009). Oxford-man institute's realized library v0.2. Oxford-Man Institute, University of Oxford.

Heracleous, M. S. & Spanos, A. (2006). The Student's t dynamic linear regression: re-examining volatility modeling. In *Econometric Analysis of Financial and Economic Time Series (Advaces in Econometrics, Vol 20 Part 1)*, D. Terrell & T. B. Fomby, eds. Emerald Group Publishing Limited, pp. 289–319.

Kalliovirta, L., Meitz, M. & Saikkonen, P. (2015). A Gaussian mixture autoregressive model for univariate time series. *J. Time Ser. Anal.* **36**, 247–266.

Kalliovirta, L., Meitz, M. & Saikkonen, P. (2016). Gaussian mixture vector autoregression. *J. Economet.* **192**, 485–498.

Kotz, S. & Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge: Cambridge University Press.

Lanne, M. & Saikkonen, P. (2003). Modeling the US short-term interest rate by mixture autoregressive processes. *J. Finan. Economet.* **1**, 96–125.

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

Li, G., Guan, B., Li, W. K. & Yu, P. L. (2015). Hysteretic autoregressive time series models. *Biometrika* **102**, 717–723.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. Wiley.

Meitz, M. & Saikkonen, P. (2017). Testing for observation-dependent regime switching in mixture autoregressive models. HECER Discussion Paper No. 420, University of Helsinki, arXiv:1711.03959.

Pitt, M. K. & Walker, S. G. (2006). Extended constructions of stationary autoregressive processes. *Stat. Probabil. Lett.* **76**, 1219–1224.

Spanos, A. (1994). On modeling heteroskedasticity: the Student's t and elliptical linear regression models. *Economet. Theory* **10**, 286–315.

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

Wong, C. S., Chan, W. S. & Kam, P. L. (2009). A student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika* **96**, 751–760.

Wong, C. S. & Li, W. K. (2000). On a mixture autoregressive model. *J. R. Statist. Soc. B* **62**, 95–115.

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

Wong, C. S. & Li, W. K. (2001b). On a mixture autoregressive conditional heteroscedastic model. *J. Am. Statist. Assoc.* **96**, 982–995.

Supplementary material for "A mixture autoregressive model based on Student's t–distribution" by Meitz, Preve, and Saikkonen (not meant for publication)

This Supplementary Material includes proofs of Theorems 1–3, information on the numerical optimization methods employed for maximum likelihood estimation, simulation experiments, and further details of the empirical example.

1 Proofs

Proof of Theorem 1. Corresponding to $\varphi_0 \in \mathbb{R}$, $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p) \in \mathbb{S}^p$, $\sigma > 0$, and $\nu > 2$, define the notation $\boldsymbol{\Gamma}_p$, γ_0, $\boldsymbol{\gamma}_p$, μ, and $\boldsymbol{\Gamma}_{p+1}$ as in (4), and note that $\boldsymbol{\Gamma}_p$ and $\boldsymbol{\Gamma}_{p+1}$ are, by construction and due to assumption $\boldsymbol{\varphi} \in \mathbb{S}^p$, symmetric positive definite Toeplitz matrices. To prove (i), we will construct a p–dimensional Markov process $\boldsymbol{z}_t = (z_t, \ldots, z_{t-p+1})$ $(t = 1, 2, \ldots)$ with the desired properties. We need to specify an appropriate transition probability measure and an initial distribution. For the former, assume that the transition probability measure of $\boldsymbol{z}_t$ is determined by the density function $t_1(z_t; \mu(\boldsymbol{z}_{t-1}), \sigma^2(\boldsymbol{z}_{t-1}), \nu + p)$, where $\mu(\boldsymbol{z}_{t-1})$ and $\sigma^2(\boldsymbol{z}_{t-1})$ are obtained from the last two displayed equations in the Appendix by substituting $\boldsymbol{z}_{t-1}$ for $\boldsymbol{x}_2$. This shows that $\boldsymbol{z}_t$ can be treated as a Markov chain (see Meyn and Tweedie (2009, Ch. 3)). Concerning the initial value $\boldsymbol{z}_0$, suppose it follows the t–distribution $\boldsymbol{z}_0 \sim t_p(\mu \mathbf{1}_p, \boldsymbol{\Gamma}_p, \nu)$. Furthermore, if $\boldsymbol{z}_t^+ = (z_t, \boldsymbol{z}_{t-1}) = (\boldsymbol{z}_t, z_{t-p})$, we find from the Appendix that the density function of $\boldsymbol{z}_1^+$ is given by

$$t_{p+1}(\boldsymbol{z}_1^+, \mu \mathbf{1}_{p+1}, \boldsymbol{\Gamma}_{p+1}, \nu) = t_1(z_1; \mu(\boldsymbol{z}_0), \sigma^2(\boldsymbol{z}_0), \nu + p)\ t_p(\boldsymbol{z}_0; \mu \mathbf{1}_p, \boldsymbol{\Gamma}_p, \nu). \tag{A1}$$

Thus, $\boldsymbol{z}_1^+ \sim t_{p+1}(\mu \mathbf{1}_{p+1}, \boldsymbol{\Gamma}_{p+1}, \nu)$ and, as in the Appendix, it follows that the marginal distribution of $\boldsymbol{z}_1$ is the same as that of $\boldsymbol{z}_0$, that is, $\boldsymbol{z}_1 \sim t_p(\mu \mathbf{1}_p, \boldsymbol{\Gamma}_p, \nu)$ (the specific structure of $\boldsymbol{\Gamma}_{p+1}$ is used here). Hence, as $\boldsymbol{z}_t$ is a Markov chain, we can conclude that it has a stationary distribution characterized by the density function $t_p(\boldsymbol{z}, \mu \mathbf{1}_p, \boldsymbol{\Gamma}_p, \nu)$ (see Meyn and Tweedie (2009, pp. 230–231)). This completes the proof of (i).

To prove (ii), note that, due to the Markov property, $z_t \mid \mathcal{F}_{t-1}^z \sim t_1(\mu(\boldsymbol{z}_{t-1}), \sigma^2(\boldsymbol{z}_{t-1}), \nu + p)$ where $\mathcal{F}_{t-1}^z$ signifies the sigma-algebra generated by $\{z_s, s < t\}$. Thus we can write the conditional expectation and conditional variance of z_t given $\mathcal{F}_{t-1}^z$ as

$$\begin{aligned} E[z_t \mid \mathcal{F}_{t-1}^z] &= E[z_t \mid \boldsymbol{z}_{t-1}] = \mu + \boldsymbol{\gamma}_p' \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{z}_{t-1} - \mu \mathbf{1}_p) = \varphi_0 + \boldsymbol{\varphi}' \boldsymbol{z}_{t-1}, \\ Var[z_t \mid \mathcal{F}_{t-1}^z] &= Var[z_t \mid \boldsymbol{z}_{t-1}] = \frac{\nu - 2 + (\boldsymbol{z}_{t-1} - \mu \mathbf{1}_p)' \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{z}_{t-1} - \mu \mathbf{1}_p)}{\nu - 2 + p} \sigma^2. \end{aligned}$$

Denote this conditional variance by $\sigma_t^2 = \sigma^2(\boldsymbol{z}_{t-1})$ (and note that $\sigma_t^2 > 0$ a.s. due to the assumed conditions $\sigma^2 > 0$, $\boldsymbol{\Gamma}_p > 0$, and $\nu > 2$). Now the random variables ε_t defined by

$$\varepsilon_t \overset{def}{=} (z_t - \varphi_0 - \boldsymbol{\varphi}' \boldsymbol{z}_{t-1})/\sigma_t$$

follow, conditional on $\mathcal{F}_{t-1}^z$, the $t_1(0, 1, \nu + p)$ distribution. Hence, we obtain the 'AR(p)–ARCH(p)' representation (7). Because the conditional distribution $\varepsilon_t \mid \mathcal{F}_{t-1}^z \sim t_1(0, 1, \nu + p)$ does not depend

on $\mathcal{F}_{t-1}^z$ (or, more specifically, on the random variables $\{z_s,\ s<t\}$), the same holds true also unconditionally, $\varepsilon_t \sim t_1(0,1,\nu+p)$, implying that the random variables ε_t are independent of $\mathcal{F}_{t-1}^z$ (or of $\{z_s,\ s<t\}$). Moreover, from the definition of the ε_t's it follows that $\{\varepsilon_s,\ s<t\}$ is a function of $\{z_s,\ s<t\}$, and hence ε_t is also independent of $\{\varepsilon_s,\ s<t\}$. Consequently, the random variables ε_t are IID $t_1(0,1,\nu+p)$, completing the proof of (ii). ■

Proof of Theorem 2. First note that $\boldsymbol{y}_t$ is a Markov chain on $\mathbb{R}^p$. Now, let $\boldsymbol{y}_0 = (y_0,\ldots,y_{-p+1})$ be a random vector whose distribution has the density $f(\boldsymbol{y}_0;\boldsymbol{\theta}) = \sum_{m=1}^M \alpha_m t_p(\boldsymbol{y}_0;\mu_m\mathbf{1}_p,\boldsymbol{\Gamma}_{m,p},\nu_m)$. According to (8), (9), (11), and (A1), the conditional density of y_1 given $\boldsymbol{y}_0$ is

$$\begin{aligned} f(y_1 \mid \boldsymbol{y}_0;\boldsymbol{\theta}) &= \sum_{m=1}^M \frac{\alpha_m t_p(\boldsymbol{y}_0;\mu_m\mathbf{1}_p,\boldsymbol{\Gamma}_{m,p},\nu_m)}{\sum_{n=1}^M \alpha_n t_p(\boldsymbol{y}_0;\mu_n\mathbf{1}_p,\boldsymbol{\Gamma}_{n,p},\nu_n)} t_1(y_1;\mu(\boldsymbol{y}_0),\sigma^2(\boldsymbol{y}_0),\nu_m+p) \\ &= \sum_{m=1}^M \frac{\alpha_m}{\sum_{n=1}^M \alpha_n t_p(\boldsymbol{y}_0;\mu_n\mathbf{1}_p,\boldsymbol{\Gamma}_{n,p},\nu_n)} t_{p+1}((y_1,\boldsymbol{y}_0);\mu_m\mathbf{1}_{p+1},\boldsymbol{\Gamma}_{m,p+1},\nu_m). \end{aligned}$$

It follows that the density of $(y_1,\boldsymbol{y}_0)$ is $f((y_1,\boldsymbol{y}_0);\boldsymbol{\theta}) = \sum_{m=1}^M \alpha_m t_{p+1}((y_1,\boldsymbol{y}_0);\mu_m\mathbf{1}_{p+1},\boldsymbol{\Gamma}_{m,p+1},\nu_m)$. Integrating y_{-p+1} out (and using the properties of marginal distributions of a multivariate t–distribution in the Appendix) shows that the density of $\boldsymbol{y}_1$ is $f(\boldsymbol{y}_1;\boldsymbol{\theta}) = \sum_{m=1}^M \alpha_m t_p(\boldsymbol{y}_1;\mu_m\mathbf{1}_p,\boldsymbol{\Gamma}_{m,p},\nu_m)$. Therefore, $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are identically distributed. As $\{\boldsymbol{y}_t\}_{t=1}^\infty$ is a (time homogeneous) Markov chain, it follows that $\{\boldsymbol{y}_t\}_{t=1}^\infty$ has a stationary distribution $\pi_{\boldsymbol{y}}(\cdot)$, say, characterized by the density $f(\cdot;\boldsymbol{\theta}) = \sum_{m=1}^M \alpha_m t_p(\cdot;\mu_m\mathbf{1}_p,\boldsymbol{\Gamma}_{m,p},\nu_m)$ (cf. Meyn and Tweedie (2009, pp. 230–231)).

For ergodicity, let $P_{\boldsymbol{y}}^p(\boldsymbol{y},\cdot) = \Pr(\boldsymbol{y}_p \mid \boldsymbol{y}_0 = \boldsymbol{y})$ signify the p–step transition probability measure of $\boldsymbol{y}_t$. It is straightforward to check that $P_{\boldsymbol{y}}^p(\boldsymbol{y},\cdot)$ has a density given by

$$f(\boldsymbol{y}_p \mid \boldsymbol{y}_0;\boldsymbol{\theta}) = \prod_{t=1}^p f(y_t \mid \boldsymbol{y}_{t-1};\boldsymbol{\theta}) = \prod_{t=1}^p \sum_{m=1}^M \alpha_{m,t} t_1(y_t;\mu(\boldsymbol{y}_{t-1}),\sigma^2(\boldsymbol{y}_{t-1}),\nu_m+p).$$

The last expression makes clear that $f(\boldsymbol{y}_p \mid \boldsymbol{y}_0;\boldsymbol{\theta}) > 0$ for all $\boldsymbol{y}_p \in \mathbb{R}^p$ and all $\boldsymbol{y}_0 \in \mathbb{R}^p$. Now, one can complete the proof that $\boldsymbol{y}_t$ is ergodic in the sense of Meyn and Tweedie (2009, Ch. 13) by using arguments identical to those used in the proof of Theorem 1 in Kalliovirta et al. (2015). ■

Proof of Theorem 3. First note that Assumption 1 together with the continuity of $L_T^{(c)}(\boldsymbol{\theta})$ ensures the existence of a measurable maximizer $\hat{\boldsymbol{\theta}}_T$. For strong consistency, it suffices to show that a certain uniform convergence condition and a certain identification condition hold. Specifically, the former required condition is that the conditional log-likelihood function obeys a uniform strong law of large numbers, that is, $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |L_T^{(c)}(\boldsymbol{\theta}) - E[L_T^{(c)}(\boldsymbol{\theta})]| \to 0$ a.s. as $T\to\infty$. As the y_t's are stationary and ergodic and $E[L_T^{(c)}(\boldsymbol{\theta})] = E[l_t(\boldsymbol{\theta})]$, condition $E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |l_t(\boldsymbol{\theta})|\right] < \infty$ ensures that the uniform law of large numbers in Ranga Rao (1962) applies.

The validity of condition $E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |l_t(\boldsymbol{\theta})|\right] < \infty$ can be established by deriving suitable lower and upper bounds for $l_t(\boldsymbol{\theta})$. Recall from (10) and (15) that

$$l_t(\boldsymbol{\theta}) = \log\left(\sum_{m=1}^M \alpha_{m,t} t_1(y_t;\mu_{m,t},\sigma_{m,t}^2,\nu_m+p)\right),$$

where

$$t_1(y_t;\mu_{m,t},\sigma^2_{m,t},\nu_m+p)=C(\nu_m)\sigma^{-1}_{m,t}\left(1+(\nu_m+p-2)^{-1}\Big(\frac{y_t-\mu_{m,t}}{\sigma_{m,t}}\Big)^2\right)^{-\frac{1+\nu_m+p}{2}}$$

and $C(\nu)=\frac{\Gamma((1+\nu+p)/2)}{(\pi(\nu+p-2))^{1/2}\Gamma((\nu+p)/2)}$. The following arguments hold for some choice of finite positive constants $c_1,\ldots,c_{10}$, and all staments are understood to hold 'for all $m=1,\ldots,M$' whenever appropriate. The assumed compactness of the parameter space (Assumption 1) and the continuity of the gamma function on the positive real axis imply that

$$c_1\leq C(\nu_m)\leq c_2. \tag{A2}$$

Next, recall that $\sigma^2_{m,t}=\frac{\nu_m-2+(\boldsymbol{y}_{t-1}-\mu_m\mathbf{1}_p)'\boldsymbol{\Gamma}^{-1}_{m,p}(\boldsymbol{y}_{t-1}-\mu_m\mathbf{1}_p)}{\nu_m-2+p}\sigma^2_m$, where the matrix $\boldsymbol{\Gamma}_{m,p}$ is positive definite and $\sigma^2_m>0$. Thus, by the compactness of the parameter space, $\sigma^2_{m,t}\geq c_3$. On the other hand, as $\boldsymbol{\Gamma}_{m,p}$ is a continuous function of the autoregressive coefficients, the continuity of eigenvalues implies that the smallest eigenvalue of $\boldsymbol{\Gamma}_{m,p}$, $\lambda_{min}(\boldsymbol{\Gamma}_{m,p})$, is bounded away from zero by a constant. This, together with elementary inequalities, yields $(\boldsymbol{y}_{t-1}-\mu_m\mathbf{1}_p)'\boldsymbol{\Gamma}^{-1}_{m,p}(\boldsymbol{y}_{t-1}-\mu_m\mathbf{1}_p)\leq\lambda^{-1}_{min}(\boldsymbol{\Gamma}_{m,p})\|\boldsymbol{y}_{t-1}-\mu_m\mathbf{1}_p\|^2\leq c_4(1+y^2_{t-1}+\cdots+y^2_{t-p})$. Thus, by the compactness of the parameter space, we have $c_3\leq\sigma^2_{m,t}\leq c_5(1+y^2_{t-1}+\cdots+y^2_{t-p})$ so that also

$$c_5^{-1}(1+y^2_{t-1}+\cdots+y^2_{t-p})^{-1}\leq\sigma^{-2}_{m,t}\leq c_3^{-1}. \tag{A3}$$

Therefore

$$1\leq 1+(\nu_m+p-2)^{-1}\Big(\frac{y_t-\mu_{m,t}}{\sigma_{m,t}}\Big)^2\leq c_6(1+y^2_t+y^2_{t-1}+\cdots+y^2_{t-p}),$$

which, together with the compactness of the parameter space, implies that

$$c_7(1+y^2_t+y^2_{t-1}+\cdots+y^2_{t-p})^{-c_8}\leq\left(1+(\nu_m+p-2)^{-1}\Big(\frac{y_t-\mu_{m,t}}{\sigma_{m,t}}\Big)^2\right)^{-\frac{1+\nu_m+p}{2}}\leq 1. \tag{A4}$$

Using (A2)–(A4) it now follows that

$$c_9(1+y^2_{t-1}+\cdots+y^2_{t-p})^{-1/2}(1+y^2_t+y^2_{t-1}+\cdots+y^2_{t-p})^{-c_8}\leq t_1(y_t;\mu_{m,t},\sigma^2_{m,t},\nu_m+p)\leq c_{10}.$$

Using this and the fact that $\sum_{m=1}^M\alpha_{m,t}(\boldsymbol{\theta})=1$ we can now bound $l_t(\boldsymbol{\theta})$ from above by a constant, say $l_t(\boldsymbol{\theta})\leq\bar{C}<\infty$. Furthermore, for some $\underline{C}<\infty$,

$$-\underline{C}(1+\log(1+y^2_t+y^2_{t-1}+\cdots+y^2_{t-p}))\leq l_t(\boldsymbol{\theta}).$$

Hence, as the StMAR process has finite second moments, we can conclude that $E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|l_t(\boldsymbol{\theta})|\right]<\infty$.

As for the latter condition required for consistency, we need to establish that $E[l_t(\boldsymbol{\theta})]\leq E[l_t(\boldsymbol{\theta}_0)]$ and that $E[l_t(\boldsymbol{\theta})]=E[l_t(\boldsymbol{\theta}_0)]$ implies $\boldsymbol{\theta}=\boldsymbol{\theta}_0$. For notational clarity, let us make the dependence on parameter values explicit in the expressions in (5) and write $\mu(\cdot,\boldsymbol{\vartheta})$ and $\sigma^2(\cdot,\boldsymbol{\vartheta})$, and let $\alpha_m(\boldsymbol{y},\boldsymbol{\theta})$ stand for $\alpha_{m,t}$ (see (11)) but with $\boldsymbol{y}_{t-1}$ therein replaced by $\boldsymbol{y}$ and with the dependence on the parameter values made explicit ($m=1,\ldots,M$). Making use of the fact that the density of $(y_t,\boldsymbol{y}_{t-1})$ has the form $f((y_t,\boldsymbol{y}_{t-1});\boldsymbol{\theta})=\sum_{m=1}^M\alpha_m t_{p+1}((y_t,\boldsymbol{y}_{t-1});\mu_m\mathbf{1}_{p+1},\boldsymbol{\Gamma}_{m,p+1},\nu_m)$ (see proof of Theorem 2) and

reasoning based on the Kullback-Leibler divergence, we can now use arguments analogous to those in Kalliovirta et al. (2015, p. 265) to conclude that $E[l_t(\boldsymbol{\theta})] \leq E[l_t(\boldsymbol{\theta}_0)]$ with equality if and only if for almost all $(y, \boldsymbol{y})$,

$$\sum_{m=1}^{M} \alpha_m(\boldsymbol{y}, \boldsymbol{\theta}) t_1(y; \mu(\boldsymbol{y}, \boldsymbol{\vartheta}_m), \sigma^2(\boldsymbol{y}, \boldsymbol{\vartheta}_m), \nu_m + p) = \sum_{m=1}^{M} \alpha_m(\boldsymbol{y}, \boldsymbol{\theta}_0) t_1(y; \mu(\boldsymbol{y}, \boldsymbol{\vartheta}_{m,0}), \sigma^2(\boldsymbol{y}, \boldsymbol{\vartheta}_{m,0}), \nu_{m,0} + p). \tag{A5}$$

For each fixed $\boldsymbol{y}$ at a time, the mixing weights, conditional means, and conditional variances in (A5) are constants, and we may apply the results on identification of finite mixtures of Student's t–distributions in Holzmann et al. (2006, Example 1) (their parameterization of the t–distribution is slightly different than ours, but identification with their parameterization implies identification in our parameterization). Consequently, for each fixed $\boldsymbol{y}$ at a time, there exists a permutation $\{\tau(1), \ldots, \tau(M)\}$ of $\{1, \ldots, M\}$ (where this permutation may depend on $\boldsymbol{y}$) such that

$$\begin{aligned} \alpha_m(\boldsymbol{y}, \boldsymbol{\theta}) = \alpha_{\tau(m)}(\boldsymbol{y}, \boldsymbol{\theta}_0), \;\; \mu(\boldsymbol{y}, \boldsymbol{\vartheta}_m) = \mu(\boldsymbol{y}, \boldsymbol{\vartheta}_{\tau(m),0}), \;\; \sigma^2(\boldsymbol{y}, \boldsymbol{\vartheta}_m) = \sigma^2(\boldsymbol{y}, \boldsymbol{\vartheta}_{\tau(m),0}), \text{ and} \\ \nu_m = \nu_{\tau(m),0} \text{ for almost all } y \; (m = 1, \ldots, M). \end{aligned} \tag{A6}$$

The number of possible permutations being finite ($M!$), this induces a finite partition of $\mathbb{R}^p$ where the elements $\boldsymbol{y}$ of each partition correspond to the same permutation. At least one of these partitions, say $A \subset \mathbb{R}^p$, must have positive Lebesque measure. Thus, (A6) holds for all fixed $\boldsymbol{y} \in A$ with some specific permutation $\{\tau(1), \ldots, \tau(M)\}$ of $\{1, \ldots, M\}$. The fact that $\mu(\boldsymbol{y}, \boldsymbol{\vartheta}_m) = \mu(\boldsymbol{y}, \boldsymbol{\vartheta}_{\tau(m),0})$ for $m = 1, \ldots, M$, almost all y, and all $\boldsymbol{y} \in A$, can be used to deduce that $(\varphi_{m,0}, \boldsymbol{\varphi}_m) = (\varphi_{m,0,0}, \boldsymbol{\varphi}_{\tau(m),0})$ for $m = 1, \ldots, M$ (see (4), (5), and Kalliovirta et al. (2015, pp. 265–266)). Similarly, using condition $\sigma^2(\boldsymbol{y}, \boldsymbol{\vartheta}_m) = \sigma^2(\boldsymbol{y}, \boldsymbol{\vartheta}_{\tau(m),0})$ (and the knowledge that $(\varphi_{m,0}, \boldsymbol{\varphi}_m, \nu_m) = (\varphi_{m,0,0}, \boldsymbol{\varphi}_{\tau(m),0}, \nu_{m,0})$), it follows that $\sigma_m^2 = \sigma_{\tau(m),0}^2$ so that $\boldsymbol{\vartheta}_m = \boldsymbol{\vartheta}_{\tau(m),0}$ $(m = 1, \ldots, M)$. Now $\alpha_m = \alpha_{\tau(m),0}$ $(m = 1, \ldots, M)$ follows as in Kalliovirta et al. (2015, p. 266)). In light of (16), the preceding facts imply that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This completes the proof of consistency.

Given conditions (i)–(iii) of the theorem, asymptotic normality of the ML estimator can now be established using standard arguments. The required steps can be found, for instance, in Kalliovirta et al. (2016, proof of Theorem 3). We omit the details for brevity. ■

2 Estimation

2.1 Numerical optimization

Finding maximum likelihood estimates of the unknown parameters of an StMAR(p,M) model amounts to maximizing $L_T^{(c)}(\boldsymbol{\theta})$, a function in $M(p+4)-1$ variables, under several constraints. Our experience with both actual and simulated data indicates that this can be challenging, in part due to multiple local maxima, and that advanced numerical optimization methods are needed. We use a hybrid numerical optimization scheme combining randomized search methods and classical gradient based methods to efficiently search for a global maximum that satisfies the constraints. (Using the commonly employed EM algorithm is also a possibility; however, contrary to some previous mixture models, the mixing weights of the StMAR model depend on the autoregressive parameters implying that the optimization problem in the maximization step does not simplify much.)

We first employ a genetic algorithm using a variety of initial populations (collections of starting points; for discussions on the genetic algorithm, other popular algorithms, and their applications in econometrics, see Goffe et al., 1994, and Dorsey and Mayer, 1995). For each of the initial populations, the genetic algorithm is run for a small number of generations to reach the region near an optimum point relatively quickly. Corresponding to each initial population, the solution found by the genetic algorithm is then used as a starting point for MATLAB's optimization method `fmincon`, which is faster and more efficient for local search (for `fmincon` we further use a sequential quadratic programming method; see e.g. Nocedal and Wright, 2006). The final parameter estimate is the best solution found by `fmincon` for all the starting points considered. This hybrid optimization scheme combining multiple initial populations, the genetic algorithm, and `fmincon` allows us to efficiently search the parameter space and reduces the risk of ending up with a local, not global, maximum. We parallelize our code to consider multiple initial populations and starting points in parallel. This helps to speed up the optimization considerably. In view of the complexity of the estimation procedure, numerical gradients and Hessians are used for the optimization.

The StMAR code used in our S&P 500 realized kernel example, further described in our StMAR MATLAB Toolbox Documentation, is available for download through the second authors webpage at https://www.researchgate.net/profile/Daniel_Preve. R code by Savi Virolainen is available through the CRAN repository in the form of the 'uGMAR' package.

2.2 Simulation experiments

We carried out several Monte Carlo studies to evaluate the performance of the numerical optimization scheme described above. The results of two of these studies are reported in Tables 1 and 2. In these experiments, 500 independent simulated sample paths were generated from an StMAR(1,2), and also from an StMAR(4,2), process; the sample sizes and parameter values used are displayed in Tables 1 and 2.

Overall, the performance of the numerical optimization scheme is quite satisfactory. As is commonly known, the degrees of freedom parameter of a Student's t–distribution is relatively difficult to estimate, especially if its true value is large. This is also the case for our StMAR model, and our simulation results indicate that the ν_m parameters can be relatively difficult to estimate even in moderate or large samples. Similar difficulties were reported by Wong et al. (2009) when estimating their (constant mixing weights) version of a Student t-mixture autoregressive model using the EM algorithm (see their Table 3).

Table 1: Simulation results for a StMAR(1,2) with various sample sizes T and 500 replications. M, Md and SD denote the sample mean, median, and standard deviation, respectively.

		$T = 500$			$T = 1000$			$T = 2500$			$T = 5000$		
	Value	M	Md	SD	M	Md	SD	M	Md	SD	M	Md	SD
$\varphi_{1,0}$	−1.50	−2.31	−1.75	2.25	−2.02	−1.58	1.55	−1.64	−1.51	0.65	−1.52	−1.51	0.22
$\varphi_{1,1}$	0.85	0.73	0.83	0.20	0.78	0.84	0.16	0.83	0.85	0.08	0.85	0.85	0.03
σ_1^2	0.35	0.38	0.32	0.45	0.35	0.33	0.09	0.35	0.34	0.05	0.35	0.35	0.04
ν_1	4.00	14.01	5.29	44.43	6.45	4.59	11.47	4.47	4.16	2.22	4.12	4.03	0.47
$\varphi_{2,0}$	−5.50	−4.79	−5.33	2.92	−5.10	−5.43	1.65	−5.41	−5.48	0.84	−5.50	−5.49	0.39
$\varphi_{2,1}$	0.35	0.44	0.37	0.30	0.40	0.36	0.21	0.36	0.35	0.10	0.35	0.35	0.05
σ_2^2	0.30	0.96	0.30	14.03	0.31	0.30	0.06	0.30	0.30	0.03	0.30	0.30	0.02
ν_2	8.00	20.48	7.08	53.76	15.15	7.44	36.00	9.19	7.76	6.64	8.30	7.80	2.12
α_1	0.60	0.61	0.59	0.10	0.59	0.59	0.07	0.59	0.59	0.04	0.60	0.60	0.03

Table 2: Simulation results for a StMAR(4,2) with various sample sizes T and 500 replications. M, Md and SD denote the sample mean, median, and standard deviation, respectively.

		$T = 1000$			$T = 2500$			$T = 5000$			$T = 10000$		
	Value	M	Md	SD	M	Md	SD	M	Md	SD	M	Md	SD
$\varphi_{1,0}$	−1.00	−1.65	−1.16	1.03	−1.46	−1.08	0.83	−1.27	−1.03	0.67	−1.15	−1.02	0.51
$\varphi_{1,1}$	0.35	0.34	0.34	0.06	0.34	0.34	0.06	0.34	0.35	0.03	0.35	0.35	0.02
$\varphi_{1,2}$	0.20	0.16	0.17	0.07	0.18	0.19	0.05	0.19	0.20	0.04	0.19	0.20	0.03
$\varphi_{1,3}$	0.15	0.12	0.13	0.07	0.13	0.14	0.06	0.14	0.15	0.04	0.14	0.15	0.03
$\varphi_{1,4}$	0.10	0.08	0.08	0.06	0.09	0.09	0.05	0.09	0.09	0.03	0.10	0.10	0.02
σ_1^2	0.25	1.84	0.26	17.82	0.46	0.25	3.77	0.26	0.25	0.04	0.25	0.25	0.02
ν_1	9.00	9.06	7.02	16.71	8.14	8.32	3.64	8.50	8.87	2.62	8.70	8.88	1.91
$\varphi_{2,0}$	−3.00	−2.95	−2.92	2.86	−2.68	−2.97	0.92	−2.80	−3.00	0.72	−2.88	−2.98	0.53
$\varphi_{2,1}$	0.30	0.29	0.30	0.14	0.30	0.30	0.04	0.30	0.30	0.03	0.30	0.30	0.02
$\varphi_{2,2}$	0.10	0.11	0.12	0.13	0.12	0.11	0.06	0.11	0.10	0.04	0.11	0.10	0.03
$\varphi_{2,3}$	0.05	0.06	0.07	0.13	0.07	0.06	0.06	0.06	0.05	0.04	0.06	0.05	0.03
$\varphi_{2,4}$	0.05	0.04	0.04	0.14	0.05	0.05	0.05	0.05	0.05	0.03	0.05	0.05	0.02
σ_2^2	0.30	1.32	0.23	12.28	0.35	0.26	0.42	0.32	0.28	0.17	0.30	0.29	0.07
ν_2	3.00	23.26	4.63	69.40	4.96	3.38	3.68	3.98	3.11	2.47	3.48	3.04	1.66
α_1	0.55	0.62	0.59	0.10	0.57	0.56	0.05	0.56	0.56	0.04	0.55	0.55	0.03

3 Empirical example

3.1 In-sample results

We estimated 12 different StMAR models with $p = 1, 2, 3, 4$ and $M = 1, 2, 3$. Of these models, the BIC, HQC, and AIC information criteria chose the StMAR(4,1), StMAR(4,2), and StMAR(4,3) models, respectively. Estimation results for these three models are shown in Table 1 of the main paper. Higher-order models were also tried but their forecasting performance was inferior to the models with $p = 4$.

3.2 Out-of-sample results

3.2.1 Two-sided prediction intervals

Table 2 of the main paper reported the percentage shares of 1, 5, 10, and 22-day cumulative RK_t out-of-sample observations that belong to the 99%, 95%, and 90% one-sided upper prediction intervals based on the distribution of the simulated sample paths. The corresponding numbers for two-sided prediction intervals (for nominal levels 99%, 95%, 90%, 70%, and 50%) are presented in Table 3. Overall, it is seen that the empirical coverage rates of the StMAR based prediction intervals are closer to the nominal levels than the ones obtained with the reference models. The StMAR(4,1), and also the StMAR(4,2), does particularly well. By comparison, the accuracy of the prediction intervals obtained with the HAR quickly degrade as the forecast period increases.

Note that to generate prediction intervals for the reference AR and HAR models, we need to specify an error distribution in these models; we assume that the errors are Gaussian. The order of the AR model is chosen using AIC and BIC; both favour an AR(p) model with $p = 11$.

Table 3: The percentage shares of cumulative realized kernel observations that belong to the 99%, 95%, 90%, 70% and 50% two-sided prediction intervals based on the distribution of 500,000 simulated conditional sample paths.

	Daily					*Weekly*				
	99%	95%	90%	70%	50%	99%	95%	90%	70%	50%
AR(11)	98.39	94.35	89.92	65.52	45.97	95.12	88.01	79.47	61.18	42.48
HAR	98.59	95.16	89.72	66.13	44.96	94.31	86.79	79.67	60.37	41.87
StMAR(4,1)	98.79	94.35	88.91	64.92	45.56	96.34	91.26	84.76	64.63	45.53
StMAR(4,2)	98.79	94.56	89.31	66.53	48.39	96.34	89.63	83.13	63.82	45.53
StMAR(4,3)	98.59	93.95	89.11	66.13	47.38	96.75	89.02	81.91	60.98	45.12
	Biweekly					*Monthly*				
	99%	95%	90%	70%	50%	99%	95%	90%	70%	50%
AR(11)	93.22	83.98	77.82	60.99	43.94	93.47	84.00	78.11	60.00	41.26
HAR	92.81	83.78	77.41	58.32	41.07	90.11	80.84	76.42	56.63	38.32
StMAR(4,1)	96.92	89.94	84.39	65.71	48.46	99.16	90.53	86.74	66.11	45.05
StMAR(4,2)	96.71	88.50	81.52	62.01	44.15	97.47	86.53	82.32	65.26	43.16
StMAR(4,3)	96.10	86.04	79.06	59.55	42.71	95.79	84.00	80.21	63.16	41.89

3.2.2 Volatility point forecast evaluation criteria

Let RM denote a (cumulative) realized measure (volatility proxy), such as the realized variance or realized kernel, and $\widehat{RM}$ a forecast of RM. Although realized measures generally are consistent estimators of the underlying latent volatility, in practice they are noisy proxies. Because of this, care needs to be taken when choosing a loss function to evaluate and compare volatility forecasts. Following the literature on volatility forecast comparison (Patton and Sheppard, 2009; Patton, 2011), we consider the two most widely used loss functions, namely squared loss (MSE)

$$L_{MSE}(RM, \widehat{RM}) = (RM - \widehat{RM})^2$$

and QLIKE (quasi-likelihood) loss

$$L_{QLIKE}(RM, \widehat{RM}) = \frac{RM}{\widehat{RM}} - \log \frac{RM}{\widehat{RM}} - 1.$$

Moreover, as Patton and Sheppard (2009) recommend the use of QLIKE rather than MSE in volatility forecasting applications, we employ QLIKE loss as our primary loss function, and squared loss as our secondary loss function.

3.2.3 Point forecasts

Results for 1, 5, 10, and 22-day cumulative RK_t forecasts based on the sample median are presented in Figure 1. The left panel reports QLIKEs and the right panel MSEs. Forecast accuracy of the models is reported relative to the StMAR(4,2) model: The horizontal line (at 100) represents the StMAR(4,2) model, whereas the other lines represent the size of the forecast error measure made relative to this model (for instance, a value of 110 in the left panel is to be interpreted as a QLIKE 10% larger than for the StMAR(4,2) model). The overall performance of the StMAR(4,2) model is quite reasonable. The model does particularly well in terms of our primary loss function, QLIKE. Overall, the StMAR(4,3) performs somewhat more poorly in terms of MSE. Figure 1 also suggests that the more parsimonious StMAR(4,1) model may be preferred to the StMAR(4,2) model over longer (biweekly, monthly) forecast periods. The popular HAR model performs well under MSE, but considerably less so under QLIKE.

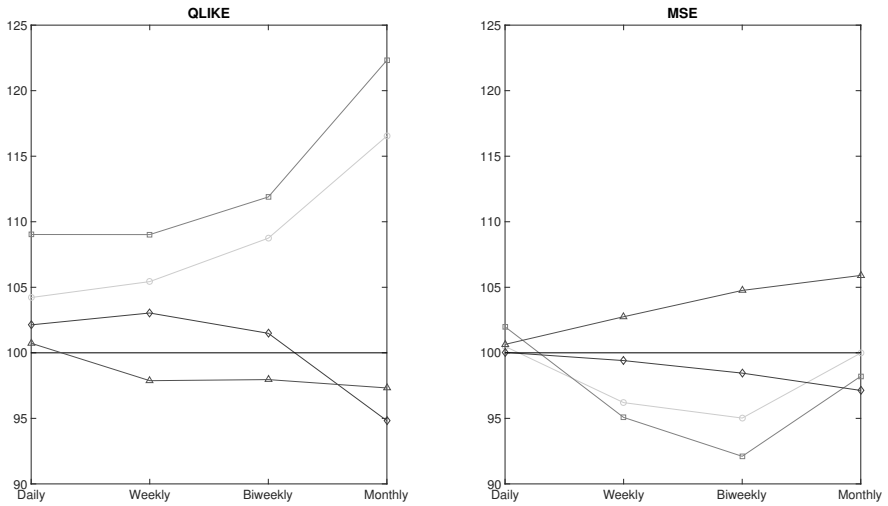

Figure 1: Relative forecast accuracies for the S&P 500 RK data in terms of QLIKE (left) and MSE (right). Results for the AR(11) (circle), HAR (square), StMAR(4,1) (diamond), StMAR(4,2) (solid), and StMAR(4,3) (triangle) models.

References

Dorsey, R.E., and W.J. Mayer (1995) Genetic algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features. *Journal of Business & Economic Statistics* **13**(1): 53–66.

Goffe, W.L., G.D. Ferrier, and J. Rogers (1994) Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* **60**(1–2): 65–99.

Holzmann, H., A. Munk, and T. Gneiting (2006) Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics* **33**: 753–763.

Kalliovirta, L., M. Meitz, and P. Saikkonen (2015) A Gaussian mixture autoregressive model for univariate time series. *Journal of Time Series Analysis* **36**: 247–266.

Kalliovirta, L., M. Meitz, and P. Saikkonen (2016) Gaussian mixture vector autoregression. *Journal of Econometrics* **192**: 485–498.

Meyn, S., and R.L. Tweedie (2009) *Markov Chains and Stochastic Stability (2nd ed.)*. Cambridge University Press, Cambridge.

Nocedal, J., and S.J. Wright (2006) *Numerical Optimization (2nd ed.)*. Springer, New York.

Patton, A.J. (2011) Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* **160**(1): 246–256.

Patton, A.J., and K. Sheppard (2009) Evaluating volatility and correlation forecasts, in T.G. Andersen, R.A. Davis, J.P. Kreiß and T. Mikosch (Eds.), *Handbook of Financial Time Series*. Springer, Berlin Heidelberg.

Ranga Rao, R. (1962) Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics* **33**:659–680.

Wong, C.S., W.S. Chan, and P.L. Kam (2009) A Student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika* **96**(3): 751–760.